9-12-2023

Combining Simulations and Single Molecule Spectroscopy To Understand SARS-CoV-2 Nucleocapsid Protein-RNA Interactions

Jhullian Jamille Alston
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Biochemistry, Biophysics and Structural Biology

Dissertation Examination Committee:
Andrea Soranno, Chair
Alex Holehouse, Co-Chair
Thomas Brett
Roberto Galletto
Nima Mossamaparast

Combining Simulations and Single Molecule Spectroscopy To Understand SARS-CoV-2 Nucleocapsid Protein-RNA Interactions

By

Jhullian J. Alston

A dissertation presented to
Washington University in St. Louis
in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2023
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

We offer special thanks to the Washington University School of Arts and Sciences and Washington University School of Medicine for the resources, support, and space to complete this dissertation. We also acknowledge Millipore Sigma and the National Institutes of Health, National Cancer Institute for their grants that funded part of this work.

Jhullian Alston

Washington University in St. Louis
August 2023

Dedicated to my Mother Iesha,

who inspired my curiosity to learn about how the world works,

and taught me to seek out the answers myself.

Abstract Of The Dissertation

Combining Simulations and Single Molecule Spectroscopy To Understand SARS-CoV-2 Nucleocapsid Protein-RNA Interactions

By

Jhullian Jamille Alston

Doctor of Philosophy in Biology and Biomedical Sciences

Biochemistry, Biophysics and Structural Biology

Washington University in St. Louis, 2023

Assistant Professor Andrea Soranno, Chair

Assistant Professor Alex S. Holehouse, Co-Chair

Disordered protein regions play crucial roles in various cellular functions, exhibiting high heterogeneity and sampling an ensemble of conformations distinct from folded domains. However, our understanding of their behavior and contributions to protein-protein and protein-nucleic acid interactions remains limited. This dissertation focuses on investigating the interactions between disordered regions and RNA, as well as folded regions of proteins, utilizing computational modeling and single-molecule fluorescence spectroscopy. The SARS-CoV-2 Nucleocapsid (N) protein serves as a model system to address broader questions concerning disordered protein behavior and N protein-mediated RNA genome packaging. I employed coarse-grained molecular dynamic simulations to characterize the cooperative binding of the first two domains of the N protein (NTD-RBD) to RNA. These simulations align with results from single-molecule experiments and offer insights into the sequence-specific contributions to binding. Notably, the simulations confirm that the disordered N-terminal domain enhances binding by approximately 50-fold through a transient

and highly heterogeneous bound state, rather than by acquiring a three-dimensional structure. Interestingly, while the folded RBDs of coronaviruses exhibit structural conservation, the disordered NTDs lack sequence conservation. By simulating six orthologous NTD-RBD constructs and conducting NTD-RBD chimeric swaps, this study suggests that complementary interactions between the NTD and RBD facilitate RNA binding, ensuring functional conservation despite variations in both RBD surface residues and NTD sequences. Furthermore, the research demonstrates that adjacent disordered regions to folded RNA binding domains can either enhance or suppress RNA binding depending on the specific sequence of the disordered region. Overall, this work provides valuable insights into the encoding of behavior within disordered regions and their impact on biological function.

Chapter 1: An Introduction to Disordered Proteins

1.1 The Free Energy Landscape of Folded Proteins is a Folding Funnel

When studying proteins, there are typically two main methods for comparing and contrasting them. The first approach involves using alignments to directly compare their linear sequences. In this analysis, each unit of the protein, called an amino acid, is examined for direct similarity or conserved physicochemical features. These features play a crucial role in driving interactions between amino acids and with the solvent environment in which the protein exists. Proteins are constructed from twenty amino acids, each possessing a distinct side chain that determines its specific physicochemical properties. These side chains can exhibit aromatic, nonpolar, positively or negatively charged, polar but uncharged, or hydrophobic characteristics. The interplay of interactions between amino acids themselves and between amino acids and the surrounding solvent dictates the behavior of the protein.

A significant focus of biological and biophysical research has been dedicated to understanding how these interactions mediate protein behavior, particularly in the process of protein folding[1,2]. Proteins strive to adopt the most thermodynamically favorable state. This involves leveraging the chemistries of their amino acids to create preferential interactions and minimize unfavorable ones. In an environment like the cytoplasm of a cell, proteins maximize interactions between hydrophilic residues and the solvent while minimizing unfavorable interactions between the solvent and hydrophobic amino acids. To sequester energetically unfavorable interactions, proteins fold upon themselves, concealing hydrophobic residues in an attempt to reach an energetically stable state. Conversely, proteins in a membrane environment may expose their hydrophobic residues to maximize preferential interactions with the hydrophobic tails of the membrane lipids[3–6]. Regardless

of the environment, the ultimate goal is for the protein to adopt conformations that enable it to reach its most energetically favorable state.

The potential conformations a protein can assume as it reaches its most thermodynamically stable state are often understood in terms of a free energy landscape depicted as a folding funnel that describes the potential conformations that a protein can take on and the energetic barriers it must cross to reach an energetically favorable final conformation[7–9]. At the top of the funnel are thermodynamically unfavorable conformations where the protein exists in an unfolded state with its unfavorable residues exposed to the solvent environment. The sides of the funnel are often portrayed with jagged edges and crevices, symbolizing local thermodynamically stable states. These local states can represent conformations where a portion of the protein forms a stable structure, or is stabilized by cooperative interactions between other domains in the protein or by chaperones.[10,11]

The descent from the top to the bottom of the funnel represents the process of protein folding, where the protein samples different but progressively more favorable conformations on the way to the bottom of the funnel. How proteins efficiently descend from unfolded to folded structure has remained an open question in structural biology and is articulated by Levinthals paradox. As a polymer with a large number of degrees of freedom, for even small proteins the number of potential configurational states is absurdly high, such that if a protein were to randomly sample all configurational states on its way to a final folded structure, it would never fold on timescales relevant to normal biological function[12,13]. However, proteins fold on micro- to millisecond timescales[14–16]. This paradox can be solved by the cooperative nature of protein folding. Each of the previously mentioned crevices in the folding funnel potentially represents the sequential stabilization

of native-like contacts that leads to cooperative folding of the final state[17]. These native-like contacts enable the folding of other regions of the protein and can speed up the folding process[18–20].

At the bottom of the funnel lies the most thermodynamically favorable state. For proteins that adopt three-dimensional structures, these funnels are typically depicted as very steep, with only one or a few thermodynamically most favorable states, separated by large barriers that prevent movement from the most thermodynamically stable state. Both the intermediates and final folded and thermodynamically favorable states are determined by the amino acids composing the protein sequence and the surrounding solvent environment. This sequence dependence underlies how diseases can be caused by small mutations to specific residues in a protein, because they perturb a proteins ability to adopt their correct conformation[21,22].

This leads us to the second method of comparing and contrasting proteins, which involves examining the three-dimensional structures they adopt. This arises from the classical structure-function paradigm, whereby the phrase "sequence dictates structure and structure dictates function" describes how the three-dimensional fold of a protein underlies its biological function, and similar structures should have similar functions[23–25]. Thus, by comparing the conformations of similar proteins to each other, we can look for similar biological functions.

Linear sequence and 3D structure are inherently linked. Proteins that are similar in sequence-space – as measured by sequence alignment – generally adopt the same 3D structures. However, the converse is not necessarily true, and structurally similar proteins can have highly divergent sequence[2626–3026]. Structural biology has made great strides in characterizing and quantitatively

understanding how sequence dictates protein structure, including the interplay between certain amino acid residues and specific structural motifs.

In recent years, significant advancements have been made in tools, particularly computational ones, that enable the direct determination of protein structure from sequence. While these tools do not fully solve the protein folding problem, which encompasses understanding the thermodynamic driving forces and mechanisms involved in protein folding, they offer valuable insights into how sequence influences structure. The current leading structure prediction tool, Alphafold, is based on the interplay of sequence alignment and structural similarity among proteins with similar sequences[31,32]. Alphafold has successfully determined approximately 200 million structures from 1 million species, covering almost every known protein. However, there are still challenges in predicting the certain types of protein regions by Alphafold, particularly those containing disordered regions.

1.2 The Free Energy Landscape of Disordered Proteins is Shallow

Unlike folded proteins that adopt well-defined thermodynamically favorable structures, disordered or unstructured proteins lack a definitive structure. Disordered regions are also typically less conserved at the residue level than their folded counterparts[33]. To better understand this difference, let's revisit the folding funnel model of protein folding. Folded proteins exist in a steep gradient with a favorable well at the bottom of the funnel where a 'single' conformation exists. However, disordered proteins exist in a shallow well, which can be imagined as virtually flat[34]. While the folded domain funnel has steep edges with a deep basin that locks it into a particular conformation, the

disordered protein has small dips that are easily overcome and a shallow and broad basin, enabling it to sample a conformational ensemble of relatively energetically equal states.

The structural description of disordered proteins is challenging due to the multitude of states they sample, which makes it incorrect to classify them using traditional structural biology classifications such as alpha-helices and beta-sheets. Instead, disordered proteins exhibit an ensemble of conformations that are sampled to varying degrees. To address this issue, concepts from polymer physics have been employed to describe disordered proteinss[31,35–39]. In polymer physics, homopolymers and simple heteropolymers are typically treated as ensembles of states, and statistical descriptions such as polymer sclaing laws are used to capture their behavior[40]. By applying quantitative measures of polymer behavior, such as the radius of gyration, to amino acid polymers, we can gain insights into the characteristics and global dimensions of disordered proteins. The radius of gyration for example, is able to describe the swelling and compaction of disordered proteins.

Disordered proteins and regions exhibit a spectrum of characteristics, ranging from fully disordered and heterogeneous ensembles to compact yet conformationally diverse molten globules, and even localized disorder within folded domains. This diversity in conformational heterogeneity arises from the relationship between the protein sequence and its ensemble of conformations. The chemical properties of the sequence directly impact the chain's behavior, leading to variations in sampled conformations and overall dimensions[41–43].

For instance, polyelectrolytes, which possess repulsive electrostatic interactions along the sequence, tend to be more expanded and adopt rigid, rod-like structures[44]. On the other hand, proteins rich in poly-aromatic residues tend to be more compact due to potential pi-stacking and cation-pi

interactions among the aromatic rings[45,46]. Disordered regions can also exhibit modularity, where specific subregions of the protein polymer contribute differently to the overall chain dimensions. In proteins with blocky sections of opposing charges, long-range attractive interactions can occur as the oppositely charged regions strive to maximize their attractive forces[47]. This continuum of disorder has a range of effects on the behavior of disordered proteins[48].

Traditionally, structural biology has focused on quantitatively understanding protein folding, but there is a separate problem of understanding the behavior of these disordered proteins that do not fold. These "floppy" polymers, acid blobs, and negative noodles have been recognized for decades, but it is only in the past decade that their importance has been realized by the larger biological community. Simultaneously we have realized our lack of quantitative understanding of this protein "unfolding" problem. To add fuel to the fire, disordered proteins are pervasive throughout biology, especially in eukaryotes, and are typically centers of interactions involving proteins and nucleic acids. The lack of understanding of how disordered regions contribute to biological function in normal context as well as in disease leaves much room for the betterment of human health.

1.3 Disordered Proteins and Their Modes of Binding to Proteins

While the focus of this thesis is on disordered proteins, it is of benefit to contrast their unique behaviors by comparing them with their counterparts, folded proteins. One of their more interesting features can be noted when comparing the way they can differ in binding ligands. Both folded proteins and disordered proteins can bind to other macromolecules. The classic concept of protein-ligand binding, proposed by Emil Fischer, describes the nature of protein-ligand binding as a puzzle

piece like complementarity where an enzyme and substrate must have a precise and specific matching of geometric shapes, similar to how a key fits into a lock to enable functionality.

However, this model was refined with the introduction of the induced fit model, which recognizes the conformational flexibility of proteins. It acknowledges that proteins can change shape in response to a specific substrate, allowing them to accommodate and bind the ligand effectively. This involves moderate rearrangement of residues to create a suitable binding environment. Nevertheless, the lock and key and induced fit models do not fully capture all the potential ways binding can occur for biological macromolecules. It is now understood that the binding process can be highly dynamic and "fuzzy." This refers to the numerous conformational states that disordered proteins can adopt and maintain, even in the bound state.

Disordered regions play a crucial role in facilitating these dynamic interactions. The modes of binding for disordered regions also exists along a continuum, ranging from rigid to highly dynamic interactions[49]. Some disordered proteins stabilize transient structures upon ligand binding[50]. While dynamic and disordered in the unbound state, specific conformations are selected and stabilized upon interaction with the ligand, resulting in a structured binding interaction. An example is the phosphorylated kinase inducible activation domain (pKID) of CREB binding to the KIX domain of the CREB binding protein. Initially dynamic and disordered, it forms a stabilized folded state composed of two alpha-helices upon binding[51]. Additionally, extensive studies have been done on P53 which, while composed of a large disordered region and DNA binding, adopts local structure when bound to numerous proteins[52–55].

On the other end of the spectrum, some disordered proteins remain highly dynamic even in the bound state. An “Extreme” example of this behavior is the "Disorder in an Ultrahigh Affinity Protein Complex" mediated by prothymosin-alpha and histone linker 1[56]. Both proteins are mostly intrinsically disordered and retain rapid dynamics both in the unbound state, consistent with unfolded proteins, and in the bound state. Surprisingly, they bind with picomolar affinity, while maintaining fast reconfiguration times as measured by nano-second fluorescence correlation spectroscopy, showing that even a complex with rapidly introconverting conformations can bind with high affinity. This work utilized single molecule fluorescence spectroscopy and simulations to capture sequence-specific effects and validate them with experimental measurements, a powerful combination for assessing disordered protein-protein and protein-nucleic acid interactions.

Between these two types of interactions, disordered regions can exhibit a variety of behaviors. For example, they may show context-specific folding, remain highly dynamic when interacting with some proteins, or stabilize a structural element upon binding to others. Some disordered regions can be both disordered and structured in the bound state, sampling a range of dynamic and stable interactions. The static pictures and models we have built over the years using structural techniques, which provide snapshots of stably bound states, have missed the transient and dynamic modes of interaction. Hence, continued efforts are needed to decipher the logic that enables such interaction.

1.4 Disordered Proteins Can Bind Nucleic Acids in a Variety of Ways

The range of binding modes exhibited by disordered proteins extends beyond protein-protein interactions and includes interactions with nucleic acids. Disordered proteins play crucial roles in various processes involving nucleic acids, such as transcription, RNA folding, spliceosome assembly,

ribosome assembly, and RNA packaging in viruses. Similar to their interactions with other proteins, disordered proteins exhibit diverse binding modes when interacting with nucleic acids. For example, disordered proteins can undergo folding upon binding to DNA, while maintaining a dynamic disordered state when bound to DNA[51,57–59].

Disordered regions also frequently act as linkers between folded RNA binding domains. This enables the folded domains to interact with the same RNA molecule, bind separate RNAs, fold upon dimerization with another disordered RNA binding protein, or directly bind to the RNA themselves[60]. Disordered regions can also enhance the ability of adjacent folded domains to interact with nucleic acids. However, our understanding of how disordered regions interact with nucleic acids and modulate the binding affinity of cognate folded domains is still limited, requiring further investigation.

1.5 The SARS-CoV-2 Nucleocapsid Protein as a Model System to Study IDPs

Over the past decade, the study of disordered proteins has gained significant attention due to their prevalence in biology and their involvement in normal cellular function, dysregulation, and diseases. The Coronavirus Disease 19 (Covid-19) pandemic caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) has highlighted the importance of studying disordered proteins. While SARS-CoV-2 has a highly ordered proteome, there are two non-structural proteins with considerable disorder, as well as the Nucleocapsid protein, which contains both folded and disordered regions.

The nucleocapsid protein, composed of five domains (three disordered and two folded), primarily functions in RNA genome packaging in Coronaviruses. These domains are referred to as the N-terminal Domain, the RNA Binding Domain, the Serine Arginine Linker, the Dimerization Domain, and the C-terminal Domain. Domains one, three, and five are disordered, while regions two and four are folded. The folded domains exhibit high sequence and structural conservation among coronavirus orthologs, whereas the disordered regions are less conserved.

The modular structure of the nucleocapsid protein offers a convenient system to investigate how disordered regions impact various important questions in the study of disordered proteins. Additionally, it provides an opportunity to gain a better understanding of a crucial protein involved in the ongoing pandemic, which has, at the time of this writing, caused the deaths of just under 7 million people worldwide.

Some of the questions that can be addressed include: 1) How do adjacent folded domains affect the conformational ensemble of disordered regions? 2) How does the interaction of disordered proteins, like the Nucleocapsid protein, with various proteins and nucleic acids concurrently influence macromolecular assembly? 3) How do specific and nonspecific interactions in disordered regions impact their macromolecular assembly? 4) How does the sequence composition of a disordered region flanking a folded domain affect nucleic acid binding?

1.6 Conclusions

Extensive studies on the nucleocapsid protein have been conducted in other coronaviruses, such as SARS-CoV-1, Murine Hepatitis Virus (MHV-1), Middle Eastern Respiratory Syndrome Virus

(MERS), and other orthologs. These studies, focusing on coronavirus nucleic acid interactions, specific interactions with motifs within their genomes, and the characterization of folded domains, have provided a substantial knowledge base to build upon. However, our approach to studying the nucleocapsid protein incorporates new methods that allow us to investigate and characterize its behavior using single molecule techniques, all-atom and coarse-grained simulations, providing unprecedented resolution and information regarding its conformational ensemble in both free and nucleic acid-bound states.

At a fundamental level, this work aims to provide a detailed characterization of the single molecule behavior of the disordered and folded domains of the SARS-CoV-2 nucleocapsid protein when interacting with single and double-stranded RNA, employing computational and experimental biophysics. On a broader scale, it offers methodologies to characterize the behavior of disordered proteins and addresses questions related to how the sequence composition of a disordered domain can influence the behavior of neighboring folded domains. Ultimately, it highlights the strength of a complementary set of techniques including simulations, theory, and single molecule fluorescence spectroscopy, which provide unparalleled resolution and characterization of disordered proteins.

Chapter 2: Integrating Single-Molecule Spectroscopy And Simulations For The Study Of Intrinsically Disordered Proteins

This chapter was published in the journal Methods as:
Alston JJ, Soranno A, Holehouse AS. Integrating single-molecule spectroscopy and simulations for the study of intrinsically disordered proteins. *Methods*. 2021 Sep;193:116-135. doi: 10.1016/j.ymeth.2021.03.018. Epub 2021 Apr 6. PMID: 33831596; PMCID: PMC8713295.

Author Contributions: J.J.A and A.S.H. conceived the manuscript. J.J.A., A.S. and A.S.H. wrote the manuscript.

2.1 Abstract

Over the last two decades, intrinsically disordered proteins and protein regions (IDRs) have emerged from a niche corner of biophysics to be recognized as essential drivers of cellular function. Various techniques have provided fundamental insight into the function and dysfunction of IDRs. Among these techniques, single-molecule fluorescence spectroscopy and molecular simulations have played a major role in shaping our modern understanding of the sequence-encoded conformational behavior of disordered proteins. While both techniques are frequently used in isolation, when combined they offer synergistic and complementary information that can help uncover complex molecular details. Here we offer an overview of single-molecule fluorescence spectroscopy and molecular simulations in the context of studying disordered proteins. We discuss the various means in which simulations and single-molecule spectroscopy can be integrated, and consider a number of studies in which this integration has uncovered biological and biophysical mechanisms.

2.2 Introduction

A structure-centric perspective has dominated our models of molecular function since the first folded proteins were visualized over 60 years ago [61–64]. Despite this, over a third of the eukaryotic proteome consists of regions or entire proteins that do not adopt a stable structure but instead sample a conformationally heterogeneous collection of structurally distinct states referred to as a conformational ensemble (Fig. 1) [48,65–67]. These intrinsically disordered proteins and protein regions (collectively referred to hereafter as IDRs) play a wide variety of roles that are critical for biological function [41,68,69]. As a result, the classical view that protein function is determined by folded proteins has expanded to recognize that function is driven by the combination of structure, conformation, and dynamics. There exists a continuum of structural heterogeneity, with well-folded hyper-stable proteins at one end and heterogeneous disordered regions at the other (Fig 1) [70]. While well-folded proteins lend themselves to various functions, including mechanical strength or enzymatic activity, disordered proteins are ideally suited for molecular recognition or biological self-assembly [41,71]. It is this repertoire of conformational plasticity that provides cells with a complex molecular toolkit, through which adaptive and responsive function can be encoded.

The three-dimensional structure of a folded domain is encoded by its primary sequence, an observation that has generally been referred to as the sequence-to-structure relationship [28,72,73]. Although IDRs do not adopt a set three-dimensional structure, they are far from "featureless random noodles." As such, an analogous sequence-to-ensemble relationship exists for IDRs in which the amino acid sequence of an IDR determines the conformational biases associated with its ensemble [68,74,75]. Just as the last four decades have focused immense attention on understanding the physical principles that map sequence to structure, the same types of questions are now being asked

of disordered regions. Beyond merely an exercise in understanding physical chemistry, the conformational biases in IDRs are a central determinant of their biological function [76–80]. As such, our ability to interpret IDR function rests at least partially on how well we understand their sequence-encoded conformational biases and transient structure.

A major challenge in studying conformational behavior in IDRs is posed by the structural heterogeneity and rapid dynamics associated with their ensembles. Due to the absence of a standard 'reference' structure, techniques such as X-ray crystallography are inherently limited in their ability to provide molecular information in the context of IDRs. Similar limitations can be extended to cryogenic electron microscopy (cryoEM), where class averaging across multiple particles is often limited to a few conformational subsets. While various techniques have been instrumental in elucidating the conformational behavior of IDRs, single-molecule fluorescence spectroscopy and all-atom simulations have played essential roles in contributing to our understanding of IDR conformational behavior and IDR dynamics. In this review, we focus on how combining single-molecule fluorescence spectroscopy and computational methods can provide quantitative and complementary insights into the solution state behavior of IDRs.

Single-molecule fluorescence spectroscopy offers a high-resolution readout of molecular behavior, making it ideal for investigating the complexities and heterogeneity of disordered proteins [36,81,82]. Specifically, single-molecule fluorescence spectroscopy enables measurements of intra- and inter-chain distances and protein dynamics with high temporal and spatial resolution. Paired with an understanding of the physics that underlie protein interactions, single-molecule approaches can be used to dissect the molecular mechanisms that drive protein behavior, dynamics, and binding. As an

example, fluorescence correlation spectroscopy (FCS) allows the diffusion coefficient of an IDR to be measured, from which the overall hydrodynamic radius of the protein can be estimated. Single-molecule Förster Resonance Energy Transfer (smFRET) provides a molecular ruler to quantify intramolecular distances within the protein [36,83,84], which can be directly tied to fundamental descriptors of polymer physics.

Finally, many single-molecule fluorescence approaches provide access to protein dynamics, over a broad range of times, from nanoseconds to hundreds of seconds, depending on the method of choice. Readouts of protein dynamics are often essential for adequately interpreting measured transfer efficiencies in smFRET experiments, particularly when discriminating whether a population reflects a static or dynamic conformation. More generally, experimentally-derived molecular dynamics offer an additional lens through which perturbations to an IDR (mutation, binding partners, solution changes) can be examined.

Molecular simulations include a robust set of tools that provide structural insight at an effectively infinite spatial resolution [85–88]. By generating large conformational ensembles, protein conformation and dynamics can be directly assessed (in the case of molecular dynamics), or ensemble-averaged properties can be computed (in the case of molecular dynamics and Monte Carlo simulations). Essentially any property that can be derived from the collection of conformations can be calculated, offering a window into a wide array of molecular information. Of particular relevance in the context of disordered proteins, all-atom simulations are especially well-poised to enable a structural interpretation of experimental data as a function of some perturbations, such as mutations, post-

translational modification, and changes in solution properties such as temperature, ion concentrations, or pH. [76,89–95].

Single-molecule fluorescence spectroscopy and all-atom simulations are highly complementary. Both techniques can, in principle, provide information at the resolution of a single molecule and do so at high temporal resolution. As such, the types of information available from single-molecule fluorescence spectroscopy and all-atom simulations are simultaneously overlapping, yet the assumptions and limitations are inherently orthogonal. As such, results from simulations can help interpret measurements made by single-molecule fluorescence spectroscopy, and *vice versa*.

The remainder of this review is laid out as follows. We provide a brief description of all-atom simulations and single-molecule fluorescence spectroscopy approaches used in the context of disordered proteins. We discuss theoretical approaches through which results from simulations and experiments can be formally integrated. We then consider specific examples in which simulation and experiment have been integrated to provide complementary insight. Finally, we conclude by summarising the outstanding questions and challenges.

2.3 Overview of all-atom simulations

Molecular simulations represent a large class of methods in which one or more molecules are explicitly described in terms of their spatial coordinates and their associated chemical and physical properties. All physics-based molecular simulations require two different components: a *representation scheme* and an *update scheme*.

The *representation scheme* reflects how a biomolecule is described within a simulation framework. This is typically achieved using a force field – a collection of equations, reference values, and rules that converts each three-dimensional conformation of a protein into a potential energy value [96,97]. The granularity of a representation scheme reports on the degrees of freedom that are explicitly encoded within that scheme (Fig. 2). We broadly categorize all-atom simulations here as those in which every biomolecule in the system is represented with atomistic detail, providing a one-to-one mapping between a simulated and real molecule. This would include representations that encode implicit and explicit hydrogens, given in both cases a clear mapping between a given biomolecule and atomic position are present. Commonly used modern forcefields that have shown good agreement in the context of disordered proteins include Amber ff03w, Amber ff03ws, Amber ff99SBws, a99SB-disp, *DES-Amber,* Amber ff99SBws-STQ, CHARMM36m, and the ABSINTH implicit solvent model [98–106].

In contrast to all-atom simulations, coarse-grained simulations sacrifice accuracy for a reduced number of degrees of freedom, facilitating larger, longer, and faster simulations. Disordered proteins have been well-described by a range of different coarse-grained models, including ultra-coarse-grained models, one-bead-per residue models, or mixed-resolution models [38,56,107–118]. While coarse-grained simulations have had remarkable success in capturing conformational behavior in disordered proteins, here we focus on all-atom simulations [119–124].

Not only must the protein of interest be represented, so too must the solution environment. The solvent can be represented using an explicit solvent model (in which water is described as individual molecules) or an implicit solvent (in which the solvation effects are 'felt' by the molecules through a

mean-field interaction) [125]. Explicit solvents are generally computationally expensive but benefit from directly capturing information related to the local solvent structure. While implicit solvents sacrifice molecular detail, the performance enhancement by reducing the number of atoms in the system by 90-99% is substantial. In the context of disordered proteins, the strength of attractive protein-water interactions has been the subject of substantial investigation, and may be one of the most important factors that determines forcefield accuracy in the context of disordered proteins [101,102,105,126,127].

The *update scheme* reflects how the molecules defined by the representation scheme evolve as the simulation proceeds. In Molecular Dynamics (MD) simulations, the update scheme converts changes in energy with respect to the atomistic position into force [96,128,129]. This force dictates the evolution of the system through a series of timesteps in which new forces are calculated and used to alter the velocity of each atom in a sequential manner. MD simulations can be used to obtain both ensemble-average values for observables of interest (e.g., end-to-end distance, the radius of gyration, local transient structure) as well as information on chain dynamics [94,99,130,131].

Monte Carlo (MC) simulations differ from MD simulations with respect to the update scheme. For MC simulations, changes to the protein conformation are made in a series of Monte Carlo steps [132]. During each step, i) a random perturbation (move) to the system is applied, leading to a temporary change in protein conformation ii) the potential energy associated with the new conformation is calculated, and iii) the new conformation is either accepted or rejected depending on the change in energy compared to an acceptance criterion. Typically Monte Carlo moves include rigid body moves (*e.g.,* translation or rotation of a molecule of interest), local moves that act on a single degree of freedom *(e.g.,* a single dihedral angle rotation or bond stretching), or more complex moves that

perturb several degrees of freedom simultaneously (e.g., in the context of concerted pivot moves or moves to perturb systems in specific ways [133–137].)

The acceptance criterion determines how moves are accepted or rejected. The most commonly used criterion here is Metropolis-Hastings, and when combined with an ergodic moveset that maintains detailed balance, this approach ensures that the collection of conformations generated sample the canonical (NVT) distribution [138,139]. Standard MC simulations cannot provide information on chain dynamics as there is no time component involved in the update scheme. However, if well-sampled ensembles are generated, equilibrium distributions of various ensemble properties such as global dimensions, average distances between residues, or transient structure can be obtained [80,90,140].

2.4 Limitations of all-atom simulations

There are several caveats associated with the interpretation of intrinsically disordered proteins with all-atom simulations. One area that has received considerable attention is that of force field accuracy [99,105,127,141]. Obtaining the correct balance of attractive and repulsive atomic interactions and dihedral angle distributions is an inherently challenging problem. For IDRs in particular, small inaccuracies can have a substantial impact on the final conformational ensemble due to the metastable nature of residual structure in IDRs. Many standard force fields lead to over-compaction of IDRs, influencing both final ensemble behavior and introducing local kinetic traps that can impair conformational sampling [101,102,141]. There has been a substantial effort over the last decade to address this challenge with IDRs in mind, with notable work from several key players including Best, Mittal, and Piana on this challenging problem [86,98–105,126,127,141–144].

A related but distinct challenge is that of conformational sampling. The heterogeneous conformational landscape of an IDR means that the total number of energetically accessible conformations is vast - much larger than there are for the same folded protein. Given conformational rearrangement takes time, there is a real and practical challenge in that for MD simulations, adequate sampling in unbiased simulations will typically require many microseconds of simulation time, even in the best-case scenario where there are no kinetic traps. Unfortunately, this requirement is often forgotten, with simulations run as a single replica for just a few hundred nanoseconds. These simulations can inherently only explore a small slice of conformational space and will inevitably lead to biased or noisy conclusions.

As mentioned above, simulations of IDRs also often experience "kinetic" traps - long-lived metastable states that impede conformational exploration. Both MD and MC simulations can suffer from these metastable states (Fig. 3). In the context of MC simulations, structurally-cooperative energetic minima raise a specific challenge, whereby the probability of the specific move(s) necessary for escape becomes vanishingly small. In the context of MD simulations, large energetic barriers between distinct states can yield slow conformational rearrangements that lead to locally trapped states domination ensembles. Even long MD or MC simulations may only sample a small region of phase space due to spending large fractions of simulations in a single state. In both cases, local conformational traps can lead to disparate levels of conformational sampling along a single polypeptide, with locally trapped structural 'nuggets' that may give the illusion of good sampling. All told, substantial care should be taken when assessing ensembles for goodness of sampling [145,146].

2.5 Single-molecule Förster Resonance Energy Transfer

Förster Resonance Energy Transfer (FRET) is a non-radiative energy transfer that can occur when the emission band of one fluorophore (the donor) overlaps in part with the absorption band of the other fluorophore (the acceptor), and the two fluorophores are in proximity to one another. FRET provides a spectroscopic ruler to measure distances between specific positions on a molecule of interest [147], such as a disordered protein (Fig. 4a). As derived by Förster, the rate of energy transfer, denoted here as k_{FRET}, is dependent on the sixth power of the distance r between the two fluorophores [148],

$$k_{FRET}(r) = k_D \left(\frac{R_0}{r}\right)^6 \quad \text{(Eq. 1)}$$

Here k_D is the inverse of the fluorescence lifetime of the intrinsic donor lifetime τ_D (i.e., in the absence of the acceptor) and R_0 is the Förster radius,

$$R_0^6 = \frac{9000(ln\ 10)\, k^2 Q_D J}{128\pi^5 n^4 N_A} \quad \text{(Eq. 2)}$$

where Q_D is the fluorescence quantum yield of the donor, n is the refractive index of the solution, J is the spectral overlap integral, N_A is Avogadro's constant, and $\varkappa$ is the dipole orientation factor which reports on the relative orientation of the dyes.

The efficiency of the energy transfer $E(r)$ can be computed by comparing the rate of the transfer k_{FRET} with the other radiative and non-radiative relaxation rates (in the absence of acceptor) from the excited state to the ground state of the donor, k_{rad} and k_{nrad},

$$E(r) = \frac{k_{FRET}(r)}{k_{FRET}(r) + k_{rad} + k_{nrad}} = \frac{R_0^6}{R_0^6 + r^{\ 6}} \quad \text{(Eq. 3)}$$

using Eq. 1 and the fact that $k_{rad} + k_{rnad} = k_D$.

In single-molecule experiments, the transfer efficiency can be measured by comparing the number of acceptor photons n_A over the total number of acceptor (n_A) and donor (n_D) photons,

$$E(r) = \frac{n_A}{n_A + n_D} \quad \text{(Eq. 4)}$$

or by measuring the change in the lifetime of the donor in the presence and absence of the acceptor,

$$E(r) = 1 - \frac{\tau_{DA}(r)}{\tau_D} \quad \text{(Eq. 5)}$$

where

$$\tau_{DA}(r) = (k_{FRET}(r) + k_{rad} + k_{rnad})^{-1} \quad \text{(Eq. 6)}$$

It is important to note that only a small number of photons are detected in a typical experiment. The low number of photons is determined by the relatively long interval between the detection of two consecutive photons (interphoton time), which is largely due to the fluorophores being trapped in long-lived dark states (*e.g.,* triplets state on the microsecond timescale) after excitation. Therefore, the measurement of transfer efficiencies is affected by shot-noise [149]. This means that, even when measuring a rigid distance across a folded domain where one single transfer efficiency is expected, a distribution of transfer efficiencies will be determined, and the mean and width of the distribution can usually be extracted.

The mean value of a shot-noise limited distribution reports on the configuration of the chain. For a rigid protein, this will coincide with a single distance as follows from Eq. 3. For IDRs, this mean value reports on the average value of transfer efficiency across the multiple conformations of the protein, and the factors that determine the average transfer efficiency are detailed below.

The width of a shot-noise limited distribution depends on the average total number of detected photons according to,

$$\sigma_{shot-noise} = \langle E\rangle(1-\langle E\rangle)\langle 1/N\rangle \leq \langle E\rangle(1-\langle E\rangle)/N_T \quad \text{(Eq. 7)}$$

where $\langle 1/N\rangle$ is the average of the inverse number of photons in a burst and N_T is the minimum number of photons in a burst (usually determined as acceptance threshold for burst identification) [150]. This implies that to determine whether a single population in the transfer efficiency distribution represents a static, rigid distance (as for folded domains) or a dynamic, flexible polymer (as for IDRs), an orthogonal measure is required. Particularly helpful in this context are measurements that report on chain dynamics, and many single-molecule fluorescence approaches provide access to protein dynamics, including the analysis of transfer efficiencies vs. fluorescence lifetimes, transfer efficiencies vs. time binning, burst variance [151], the use of Probability Distribution Analysis (PDA) [152–154], and analysis of photon trajectories of immobilized molecules [79,155–160].

Since the measured transfer efficiency is an average of a given interval of time, the measured dynamics will reflect the conformational changes occurring on the characteristic timescale of observation. The diffusion time of molecules in the confocal volume and the camera detection rate in TIRF microscopy set an intrinsic timescale of reference for the corresponding measurements, usually in the range of milliseconds. Another timescale is given by the time-data bin used to analyze the data. There are no special limitations in the range of binning times that can be applied besides the intrinsic limitations due to the detection rate, whether related to the acquisition rate of the instrument (*e.g.,* camera frame rate) or to the emission rate of the fluorophores (*e.g.,* only a limited number of photons are observed in freely diffusing molecules). However, the choice of bin width

dictates the averaging of FRET information over the selected time range. As a tangible example of what this assumption can mean, let us assume the case of two different conformational states with distinct conformations. When exchange dynamics are slower than the binning time, the two states will appear as separated peaks with distinct mean transfer efficiencies and widths. When dynamics are faster than the binning time, the transfer efficiencies associated with the two states will be averaged out together, giving rise to a single population. When using intermediate binning times, a partial averaging of the two populations occurs. Therefore, analysis of transfer efficiency histograms as a function of time binning can provide insights on conformational changes and dynamics [156,161,162]. When the distribution of transfer efficiencies is broader than shot-noise, Photon Distribution Analysis (PDA) can provide insights into the underlying populations as well as interconversion between different states [152–154]. The method appears to be more sensitive to interconversion occurring between 0.01 and 10 times of the burst duration [163]. Whereas PDA considers the differences in transfer efficiency among all the detected molecules, Burst Variance Analysis (BVA) quantifies how the transfer efficiency changes inside each molecule (burst) over time [151]. Consequently, BVA provides a measure of dynamics on timescale longer than the minimum binning of photons required to compute the transfer efficiency variance within the burst. Analysis of the photon trajectory with maximum likelihood methods do not require time binning and can provide access to fast dynamics (up to the microsecond timescale) by studying the statistics of detected photons [164].

Another intrinsic timescale in single-molecule measurements is the fluorescence lifetime of the fluorophore, which is typically in the nanosecond range. Therefore, contrasting the donor lifetime in the presence of the acceptor (Eq. 5) with the transfer efficiency determined from the number of

acceptor and donor photons detected in a burst (Eq. 4) provides a useful test for the occurrence of fast dynamics compared to the burst duration. Indeed, Eq. 5 provides information on the transfer efficiency adopted by the system on the lifetime timescale [165–168]. Instead, Eq. 4 computes transfer efficiencies from the number of donor and acceptor photons detected in a burst and, therefore, probes the transfer efficiency on the timescale associated with burst duration (or with the data binning time). The burst duration of freely diffusing species is commonly on the millisecond timescale. In the case of a rigid distance, we expect an identical transfer efficiency on the nanosecond and millisecond timescale probed by lifetime and bursts, respectively, as indicated by the linear relation between the two terms in Eq. 4 and 5. As a result, the measured static distribution should fall on the corresponding predicted linear trend. A deviation from this linear behavior is expected when the molecule of interest samples a broad conformational ensemble on a timescale longer than nanoseconds but shorter than milliseconds, as in the case of many IDRs [164,169,170]

$$\tau_{DA}/\tau_D = 1 - \langle E \rangle + \frac{\sigma^2}{1-\langle E \rangle} \qquad \text{(Eq. 8)}$$

where σ represents the variance of transfer efficiency due to fluctuations in the donor-acceptor distance.

A similar dependence can also be found when studying the characteristic delay acceptor emission [164]. If we denote $P(r)$ as the distribution of conformations adopted by the interdye distance and we assume the interdye dynamics are slower than the dye tumbling but significantly faster than the interphoton times, we can compute the average τ_{DA} from the dynamic distribution as defined by,

$$\tau_{DA} = \int_0^\infty \; tI(t)dt \; / \; \int_0^\infty \; I(t)dt \qquad \text{(Eq. 9)}$$

where $I(t)$ is the time-dependent fluorescence intensity and is given by [168],

$$I(t) = I_0 \int_0^\infty \; P(r)e^{-t/\tau_{DA}(r)}dr \qquad \text{(Eq. 10)}$$

By integrating over the distance *r*, Eq. 10 assumes that the lifetime decay occurs faster than the conformational change in *r* as sampled by the distribution of distances given by *P(r)*.

The corresponding mean transfer efficiency is computed as,

$$\langle E \rangle = \int_0^{l_c} E(r)P(r)dr \qquad \text{(Eq. 11)}$$

where l_c is the contour length between the dyes if the protein segment was fully extended.

Dye orientation is commonly described in terms of a parameter defined as $\varkappa$, with the typical result of "$\varkappa^2 = 2/3$" for isotropic orientation of the fluorophores[36,81]. This is commonly valid if the dye tumbling is faster than the protein dynamics. However, if the dynamics of the protein are instead much faster than the tumbling of the dyes, the relative orientation of the dyes becomes coupled to the transfer efficiency. Under this regime, the mean transfer efficiency is given by the combination of the distribution of distances sampled by the protein and of k sampled by the fluorophores with the transfer efficiency dependence of distance and k:

$$\langle E \rangle = \int_0^4 \int_a^{l_c} E(r,\kappa^2)P(r)\, p(\kappa^2)\, dr\, d\kappa^2 \qquad \text{(Eq. 12)}$$

where *a* is contact radius between the dyes, *P(r)* is the inter-dye probability distribution as described previously, $E(r, \varkappa^2)$ is the transfer efficiency dependence on $\varkappa$ is as given by,

$$E(r,\kappa^2) = \left(1 + \frac{2}{3\kappa^2}(r/R_0)^6\right) \qquad \text{(Eq.13)}$$

and the probability distribution $p(\varkappa^2)$ is given by:

$$p(\kappa^2, 0 \le \kappa^2 \le 1) = \frac{1}{2\sqrt{3\kappa^2}} ln(2+\sqrt{3}) \qquad \text{(Eq. 14)}$$

and,

$$p(\kappa^2, 1 \leq \kappa^2 \leq 4) = \frac{1}{2\sqrt{3\kappa^2}} ln(\frac{2+\sqrt{3}}{\sqrt{\kappa^2+\sqrt{\kappa^2-1}}}) \quad \text{(Eq. 15)}$$

Analogously, if the chain dynamics are faster or comparable to the fluorescence lifetime, the energy transfer rate will depend on the distribution of states sampled by labeled molecules,

$$\langle E \rangle = \int_a^{l_c} (R_0/r)^6 P(r)\, dr / (1 + \int_a^{l_c} (R_0/r)^6 P(r)\, dr)$$

(Eq. 16)

where, as before, l_c is the contour length of the chain and a the dye-dye contact radius.

Experimentally, time-resolved lifetime and anisotropy measurements can provide information on the tumbling rate of the fluorophores [171], and more extensive discussion of the influence of the different timescales at play on transfer efficiency histograms can be found in the fundamental works of Gopich and Szabo [149,150,165,172].

Finally, the functional form of the inter-dye probability distribution *P(r)* is typically approximated using simple polymer models or inferred from molecular simulations. While the mean transfer efficiency can be used to constrain the mean value of the distribution, the variance of transfer efficiency fluctuations σ can be used as a further constraint [173] for the distribution given that,

$$\sigma^2 = \int_0^{\infty} E(r)^2 P(r) dr - \langle E \rangle^2 \quad \text{(Eq. 17)}$$

Various closed-form analytical and numerical models have been applied to describe FRET data, including the freely jointed (or Gaussian) chain, worm-like chain, and the self-avoiding walk [36,40,174–176]. Popularity of these models is largely due to the fact that they rely on single fitting parameters, enabling association of the mean transfer efficiency with a mean square distance, persistence length,

or excluded volume term. While the worm-like chain and self-avoiding walk distributions provide descriptive parameters to capture excluded volume effects (and repulsive interaction in general), more advanced polymer models are required to capture the transition from good to poor solvent often observed by tuning solution conditions (*e.g.,* denaturant), temperature, or by altering the sequence [177–181]. Ziv *et al.* adapted the coil-to-globule theory of Sanchez, introducing a conversion factor between the mean radius of gyration and the corresponding distribution of end-to-end distances [174,182,183]. More recently, by comparing single-molecule FRET and small-angle X-ray scattering (SAXS) data with simulations, Zheng *et al.* have proposed an empirical adaptation of the self-avoiding walk distance distribution that depends on the solvent quality through the scaling exponent ν [173,184]. These polymer models have been employed extensively to study disordered and unfolded proteins in many different contexts, where they have shown remarkable success [35,89,90,185–187].

2.6 Fluorescence correlation spectroscopy

Fluorescence correlation spectroscopy (FCS) is a powerful complementary tool to smFRET that measures the correlations of fluorescence fluctuations caused by diffusion and dynamics of labeled molecules as well as other photophysical effects [188–192]. This correlation can be computed as,

$$G(\tau) = \frac{\langle I(t)I(t+\tau)\rangle_t}{\langle I(t)\rangle_t^2} \quad \text{(Eq. 18)}$$

where $\langle ... \rangle_t$ represents the average over all measured times, τ is the lag time at which the correlation is computed, and $I(t)$ and $I(t + \tau)$ is the fluorescence intensity at times t and $(t+\tau)$.

When applied to single-photon counting measurements, the expression in Eq. 18 can be interpreted as the joint probability of observing a photon at time t and $(t + \tau)$ compared to the joint probability of observing two photons at any time,

$$G(\tau) = \frac{p(photon\ at\ t\ and\ t+\tau)}{p(photon\ at\ any\ t)^2} \quad \text{(Eq. 19)}$$

Eq. 19 provides an intuitive way to understand how the correlation decays of FCS relates to molecular diffusion through the confocal volume or other physical properties. If the lag time is shorter than the average residence time of a molecule in the confocal volume, the joined probability of observing two photons that are separated by that given lag time will be high since they are emitted by the same molecule. When the lag time increases and approaches the average residence time of the molecule in the confocal, the decrease in the joined probability reflects the increased probability of the emitting molecule exiting the confocal volume without being immediately replaced by a new one. Ultimately, if the lag time is much longer than the average residence time of the molecule, the joined probability of observing two photons at times t and $(t + \tau)$ will be identical to the probability of observing two photons at any time. Therefore, for very long lag times, the correlation (as described by Eq. 18 and Eq. 19) tends to unity. The same reasoning can be applied to understand the correlation decay connected to photophysical effects that result in dark states (*e.g.*, quenching or triplet states).

To better understand the properties of the correlation function, we can express the intensity as,

$$I(t) = \sum_j \quad i_j(t) + I_{bg}(t) \quad \text{(Eq. 20)}$$

with i_j and I_{bg} being the intensity of a single fluorophore and the background intensity at time t, respectively. Under this description, the correlation function from Eq. 18 adopts the form,

$$G(\tau) = \frac{N\langle i(t)i(t+\tau)\rangle_t + N(N-1)\langle i(t)\rangle_t^2 + 2N\langle I_{bg}(t)\rangle_t + \langle I_{bg}(t)\rangle_t^2}{N^2\langle i(t)\rangle_t^2} \quad \text{(Eq. 21)}$$

Assuming that $\langle I_{bg}(t) \rangle_t$ has a negligible contribution compared to other quantities, Eq. 21 reduces to,

$$G(\tau) \simeq \frac{\langle i(t)i(t+\tau)\rangle_t}{N \langle i(t)\rangle_t^2} + 1 \quad \text{(Eq. 22)}$$

where the amplitude of the correlation clearly depends on the inverse of the average number of molecules N observed in the confocal volume. As implied by Eq. 22, FCS is not exclusively restricted to the single-molecule regime and is often applied in conditions under which multiple molecules diffuse through the confocal volume. Importantly, when measurements are performed at sufficiently low concentrations of N molecules, the background term may contribute to the correlation amplitude and, if not accounted for, can affect a proper determination of *N*. Importantly, the ability to function at extremely low concentrations makes FCS an ideal technique in the context of IDRs that are prone to undergo self-assembly [193,194].

Nanosecond FCS (nsFCS) extends conventional FCS to sub-microseconds timescales by distributing photons across multiple detectors. The application of multiple detectors circumvents the intrinsic limitations (deadtime after pulse) that affect the correlation on individual detectors. Access to the sub-microsecond timescales allows the assessment of the contribution of static quenching (e.g., caused by dye-residues and dye-dye stacking), protein dynamics, and other photophysical effects [185,195,196]. Of particular interest in the context of IDRs is the application of nsFCS to provide an estimate of chain conformational dynamics. ns-FRET-FCS provides a measure of the protein dynamics through the characteristic correlated relaxation in the donor-donor and acceptor-acceptor correlations and the anti-correlated relaxation of the donor-acceptor cross-correlation. The anti-correlated decay directly reflects the anticorrelated intrinsic nature of FRET, where an increase in acceptor emission corresponds to a decrease in the donor emission and vice versa. When performed at the single-molecule level in a subpopulation specific way, the amplitude of the dynamic component (in the absence of quenching) of the correlation is directly related to the variance of transfer efficiency fluctuations in the solution, according to [197,198],

$$C_{ij}(\tau) = A_{ij}(1 - c_{AB}e^{-\tau t \tau_{AB}})(1 - c_T e^{-\tau t \tau_T})\left(1 - c_b^{ij}\ e^{-\tau t \tau_{CD}}\right) \qquad \text{(Eq. 23)}$$

with $i, j\ =\ A, D$, $c_b^{DD} = \frac{\sigma^2}{\langle E \rangle^2}$, $c_b^{AA} = \frac{\sigma^2}{1-\langle E \rangle^2}$, $c_b^{AD} = \frac{\sigma^2}{\langle E \rangle^2\ \ (1-\langle E \rangle^2)}$

Here, A_{ij} is an amplitude component related to the number of fluorescent molecules in the confocal volume, c_{AB} is the antibunching amplitude, τ is the lag time between the two detected photons, τ_{AB} is the correlation time of the antibunching component, c_T is the amplitude of the triplet component, τ_T is the correlation time of the triplet component, and τ_b is the correlation time associated with chain dynamics.

It is important to stress that the relaxation time τ_b of these three correlations represents a FRET-filtered value of the real reconfiguration time of the protein. Gopich *et al.* determined a simple correction factor that enables the extraction of the reconfiguration time of the protein [197]. This reconfiguration time can be directly linked to polymer quantities such as the characteristic times derived in Rouse and Zimm models [199–203]. Finally, since this approach provides access to the variance in the transfer efficiency fluctuations, it can be combined with single-molecule FRET and lifetime measurements to infer properties of the distribution of transfer efficiencies.

Single-molecule contact formation dynamics can also be probed using photon-electron transfer (PET) between a single fluorophore and an aromatic residue (or other quencher attached to the protein) [185,194,196,204,205]. In PET-FCS experiments, the fluorophore forms transient static complexes with the quencher. Therefore, the amplitude c_q and characteristic time τ_q associated with the static quenching in the correlation contains information on both the on- and off-rate of contact formation: $\tau_q = 1/(k_{on} + k_{off})$ and $c_q = k_{on}/k_{off}$. Importantly, static quenching is not diffusion limited, such that the on-rate must be calibrated with a known diffusion-limited quenching process

to extract the real on-rate of contact formation. For comparison, the dynamic quenching between dyes and aromatic residues, as measured by changes in the fluorescence lifetime, has been found very close to the diffusion-limited regime and offers a convenient strategy for extracting correction factors for reaction-limited quenching.

Furthermore, the on-rate of contact formation as measured in PET-FCS experiments can be related to the reconfiguration time measured by ns-FRET-FCS when computing the first passage time of the corresponding polymer model [185,206,207]. In the scenario where internal friction dominates the protein dynamics, Cheng *et al.* proposed a convenient equation where the contact time τ_c^{IF} is computed by using the Szabo-Schulten-Schulten theory[208] in terms of 1D diffusion in a potential of mean force for the Rouse and Zimm model for internal friction [201]. This leads to the remarkably simple expression,

$$\tau_c^{IF} = \left(\frac{\pi}{6}\right)^{0.5} \frac{R}{R_c} \tau_i \qquad \text{(Eq. 24)}$$

Where τ_i is the internal friction characteristic time, R is the root-mean-square separation between the dye and the quencher and R_c is the contact radius for quenching.

2.7 Challenges and practical considerations in single-molecule fluorescence spectroscopy

Recent cross-lab verification demonstrated that smFRET can provide highly reproducible results across different laboratories when the instruments are properly calibrated [209]. Calibration of experimental setups can be obtained by measuring reference samples that provide an estimate of the excitation and detection efficiency of the detectors and correct for the different quantum yields of the fluorophores [36,81]. An elegant solution has recently been proposed [209–212] and relies on the use of alternating-laser excitation (ALEX) [213,214] or pulsed interleaved excitation (PIE) [215,216]. In brief,

fluorescence detection of donor-only and acceptor-only molecules provides insights on the direct excitation of acceptor and cross talk, while a comparison of the stoichiometry ratio of donor-acceptor labeled molecules as a function of transfer efficiency (*e.g.,* polyproline or other systems of interest) enables estimates of the relative corrections for detection efficiency and quantum yield across the donor and acceptor channels. Investigating the dependence of the stoichiometry ratio vs. transfer efficiency requires either multiple samples with different mean transfer efficiency or altering transfer efficiency by changing the solution conditions of the same sample, although it should be noted that altered solution conditions may alter the quantum yield of the fluorophores or introduce quenching, which may further complicate this analysis.

An important decision in designing smFRET experiments is the choice of the experimental strategy, *e.g.,* whether one is investigating freely diffusing or immobilized molecules. Which approach to take is determined by several factors, including the accessible experimental setup and the biophysical or biochemical question being addressed. A common solution for the investigation of immobilized molecules is the use of Total Internal Reflection Fluorescence (TIRF). TIRF microscopy relies on evanescent illumination of samples tethered to the surface [217,218], reducing background fluorescence from labeled molecules in solution. TIRF microscopy often uses camera-based detection, enabling the simultaneous observation of multiple molecules and the study of out of equilibrium kinetics. Confocal single-molecule fluorescence microscopy enables measurements of both freely diffusing and immobilized molecules.

The use of single-photon counting avalanche photodiodes and Time Correlated Single Photon Counting (TCSPC) electronics provide access to fast dynamics, kinetics, and photophysical

properties of the systems such as triplet and fluorescence lifetimes. Owing to the high temporal resolution, confocal single-molecule fluorescence experiments have captured even rare events such as the transition path time from a folded to unfolded state or from a bound to unbound state [79,219]. Several approaches have been developed to enable the investigation of higher concentrations regimes and out of equilibria phenomena in confocal setups. For example, zero-mode waveguides have been used to extend the concentration boundaries of single-molecule confocal detection up to micromolar concentrations [220,221]. Similarly, microfluidic devices with fast mixing allows following the kinetics of the system of interest, at the single-molecule level, from hundreds of microseconds up to tens of seconds. [222–228]. Recurrence analysis of single particles (RASP) also captures the kinetics of freely diffusing molecules by identifying those molecules that after passing through the confocal volume re-enter in the confocal volume. By studying how the conformations of these molecules changes at different lag times, information on kinetics can be reconstructed [229,230]

Once the experimental setup and strategy have been chosen, the next step is the selection of appropriate labeling positions. The average Förster radius across the fluorophores suitable for single-molecule FRET lies between 5 and 7 nm, limiting the sensitivity of the method to distances approximately larger than 2-3 nm and smaller than 10 nm (see Fig. 4c). While knowledge of the protein structure allows for the tailoring of dye placement in folded proteins, more difficult is the choice of label position when studying IDRs, since the sequence properties of the chain can significantly alter the root-mean-square interdye distance. A distance of approximately 50 - 60 amino acids provides an appropriate dynamic range for sequences with a broad range of charge compositions, ranging from expanded polyelectrolytes to collapsed polyampholytes [231]. It is important to note that proline-rich sequences can adopt very extended configurations [93,232]. Sampling

different interdye positions within the same IDR can further improve the ability to to quantify the dependence of the related interdye distance with the sequence length of the measured segment, providing access to the associated scaling exponent [35,56,76,233]. As mentioned, an estimate of the expected distance between two pairs of residues can be derived using appropriate polymer models or simulations [119,120,162,231,232,234–238].

The amino acid sequence raises additional constraints with respect to the optimal strategy for labeling. Both FCS and FRET measurements rely on covalently labeling proteins of interest with one or more fluorescent dyes. The labeling strategies typically take advantage of endogenous cysteine residues or introduce novel cysteines via mutation. These cysteine residues can be covalently modified with fluorescent dyes via maleimide chemistry [239]. Given the general scarcity of cysteine residues in most protein sequences, it is not uncommon for an IDR of interest to contain one or even zero endogenous cysteines. In this scenario, mutations that convert small polar amino acids (*e.g.,* serine or glutamine) to cysteine (or vice versa when removing unwanted endogenous cysteines) are generally expected to have minimal impact on the conformational behavior of a disordered protein owing to the approximate chemical equivalence of the residues. Nevertheless, scenarios in which altering the number of cysteine residues in the protein can arise, in which case alternative labeling strategies are required.

The introduction of non-natural amino acids and enzymatic reactions for site-specific labeling presents a set of approaches that move beyond the intrinsic limitations of cysteine-based labeling methods. For example, the use of the enzyme sortase A has enabled site specific labeling of proteins that contain substantial cysteines and would be otherwise unamenable to site specific labeling by

maleimide chemistry [240,241]. Sortase A catalyzes the ligation of an "LPETG" motif with a "GGG" motif [242,243]. In this way, a linker containing a fragment of a protein that harbors a single cysteine can be utilized to enable maleimide chemistry on the sole cysteine residue [244]. The rest of the protein that contains multiple cysteines can be ligated to the singular-cysteine containing protein fragment. Conversely, the use of split-inteins can enable maleimide labeling of multiple cysteine residues across a protein that has been separated into fragments that contain one cysteine each, followed by ligation of the fragments with native chemical ligation [245,246]. Non-natural amino acids, alone and in conjunction with Click chemistry, can enable site specific labeling which can be critical in the context of three- or four-color smFRET experiments, although the incorporation of non-natural amino acids can lead to complications in protein expression yields [186,247–253]. Additionally, in sequences where mutating endogenous cysteine residues is likely to disrupt protein conformation, non-maleimide chemistry methods offer an alternative labeling strategy. For example, short peptide sequences (A4/Q-Tags) have been used to site-specifically label several proteins [254–258]. Q tags utilize a transglutaminase catalyzed reaction to ligate cadaverine functionalized fluorophores to the glutamine residue present in Q-tag motifs (PNPQLPF, PKPQQFM, GQQQLG) [254]. Unlike sortase or maleimide chemistry, the A4 tag utilizes a phosphopantetheinyl transferase reaction to conjugate CoA conjugated substrates to the serine present in the A4 motif (DSLDMLEM) [256,259–261]

The subsequent key step rests on the choice of dye. With the advent of superresolution microscopy, a broad range of fluorophores and donor-acceptor combinations have become available, each with different photophysical and chemical properties. It is worth mentioning that, when targeting the cellular environment in single-molecule experiments, a choice of red-shifted fluorophores (compared to the often used 480-520 nm range of excitation) has proven to reduce the fluorescence

contribution originated by the cellular background [230]. Each fluorophore differs not only in excitation and emission wavelengths, but also in terms of geometry, hydrophobicity, net charge and linker flexibility and length. As a result, different dyes have different possible impacts on protein conformation, depending on the sequence-encoded physical chemistry of the given protein.

Although several studies have implicated dyes as a source of non-native interactions that can alter conformational behavior [262–265], with careful dye selection and validation, these issues can be minimized, and a number of studies have found that dyes can have a minimal impact on ensemble behavior [56,90,266,267]. However, some relevant examples do require attention. Due to the high hydrophobicity, the popular ATTO 647N dye has been reported to cause a substantial collapse of an IDR, at variance with many other dyes [89]. This result suggests that caution must be taken when using this fluorophore on IDRs. Focusing on the role of dye charges, many of the commonly used fluorophores, such as Alexa 488 and 594, carry a -2 negative charge each. This net charge may become particularly relevant when investigating polyampholytic sequences with local regions of net positive charge, or with proteins that possess a net positive charge, such that these electrostatic effects must be accounted for when modeling or interpreting the experimental data [231]. Finally, the choice of the dye may also depend on the specific environment in which the protein is located: recent work has revealed preferential interaction of specific fluorophores with lipids [268].

The reality is that there is no "one-size fits all" solution for choosing dyes. For some proteins, certain dyes will likely have an impact on molecular details, while in others they will not. The determinants of dye effects reflect the physicochemical properties of dyes and the sequence-encoded physical chemistry present in a given protein. As such, due diligence is required when considering if

and how dyes may be impacting conformational behavior. This may include testing different combinations of dye pairs to ascertain if different dyes reveal different results. Ideally, orthogonal verification with other techniques (either computational and/or experimental) offers a convenient route to refute or confirm findings [266].

When approaching the data analysis of smFRET experiments, several assumptions undergo transforming the measured transfer efficiency into a distance distribution. The most commonly cited assumption is the isotropic orientation of the fluorophores described by the κ^2 parameter in the definition of the Förster radius (R_0). Although simulations may achieve a quantitative estimate of κ^2, a measurement of the steady-state and/or time resolved anisotropy of the two fluorophores provides quantitative insight into possible conformational restrictions of fluorophores orientation [171,269].

Less discussed but equally important when comparing results from single-molecule fluorescence experiments with simulations is an appropriate estimate of the characteristic timescales at play. It is crucial to consider the timescales' impact on the interpretation of the mean transfer efficiency, with particular attention needed with respect to the rate of fluorophore tumbling and fluorescence lifetime, as well as chain dynamics. As mentioned in section 2.3, Eq. 6 assumes that the dynamics of the chain are faster compared to the interphoton time, but slower than both dye tumbling and the fluorescence lifetime (Fig. 5). This behavior is a precondition for invoking the approximation that fluorophores are experiencing isotropic orientation.

Once all these aspects are considered, a root mean square distance is extracted based on the mean transfer efficiency $\langle E \rangle$. The measured distances provide a readout on the separation of the

fluorophore, as opposed to a direct readout on the residue-residue distance between labeling positions and fluorophores linker needs to be accounted for [171]. For the dye pair Alexa 488 and 594, the dye linkers' contribution to the root-mean-square interdye distance for an unstructured protein corresponds to an increase in the protein sequence length of about nine amino acids [235,236,270].

2.8 Approaches for the integration of single-molecule fluorescence spectroscopy and simulations

Various theoretical frameworks appropriate for the integrations of single-molecule fluorescence spectroscopy with results from atomistic simulations have emerged over the last decade, with many of these being directly applicable to the study of disordered proteins. Rather than providing an exhaustive technical description of these methods, we will briefly overview the conceptual approaches and practical methodologies available.

The most straight-forward approach involves performing unbiased molecular simulations of a protein of interest without dyes, computing relevant observables from the simulations, and then comparing those observables with the analogous values obtained from experiment [76,271,272]. In the context of smFRET experiments, this would involve computing distributions of inter-residue distances and then comparing those distances with the analogous distribution obtained from experiments [212]. For FCS, this would involve computing a hydrodynamic radius (R_h) from simulations and comparing that value with the apparent R_h obtained from the diffusion constant [273,274]. For nsFCS, this would involve computing molecular reconfiguration times and comparing those times with timescales measured by experiment [185,275,276]. This approach makes two key assumptions. Firstly, it assumes that the dyes do not significantly contribute to the conformational ensemble obtained from simulations, such that the ensemble generated in the presence/absence of

dyes is equivalent. Secondly, it assumes that analytical models (i.e., *P(r)*, for determining inter-dye distance, see section 2.3) offer an appropriate route to back-calculate molecular properties that can also be obtained from simulations. Both these assumptions are reasonable, well established for many systems, and often taken to be true irrespective of if a comparison between simulation and experiment is to be performed. This naive comparison offers a convenient first approach to demonstrate agreement between simulation and experiment. Moreover, if the agreement is poor, it provides a starting point to diagnose the origin of discrepancies.

While simulations lacking fluorophores are – by definition – reporting on the naturally occurring state of the protein, for a quantitative comparison with single-molecule spectroscopy, this approach has some shortcomings. For one, the absence of explicit dyes ignores their conformationally heterogeneous nature, and as such, these simulations are unable to interpret/assess dye-protein interactions, should they occur. Furthermore, a simulation that lacks explicit dyes does not generally take dye photophysics into account. Consequently, an alternative approach involves the explicit inclusion of dyes in the simulations [198,212,266,277]. Here, simulations of biomolecules with fluorescent dyes are run, and then relevant observables (e.g., FRET transfer efficiencies) are calculated from ensembles using dye orientation directly. The resulting computationally-derived FRET results can then be directly compared with mean transfer efficiencies obtained from smFRET experiments. While conceptually appealing, the inclusion of fully parameterized dyes in all-atom simulations is somewhat less common than one might expect. This reflects several challenges that dyes introduce in the context of all-atom simulations.

One challenge in the inclusion of explicit dyes is the appropriate forcefield parameters. As mentioned, even for protein-only systems, correctly parameterized force fields that accurately describe IDR configurational rearrangement and dynamics are challenging. This is despite the wealth of data surrounding protein physical chemistry and structure. In contrast, large heterocyclic aromatic dyes are comparatively less well-studied. Consequently, the validity of dye parameters is less clear. Furthermore, there is good reason to expect that fixed-charge force fields may struggle to correctly capture the physical chemistry of large heterocyclic dyes due to the complex delocalized electron systems that are distributed across them. Finally, interpreting transfer efficiencies directly from dyes requires consideration of the dye photophysics, including the orientational dependence of the dipole-induced energy transfer that gives rise to FRET [212,277]. In principle, the explicit inclusion of fluorophores allows the impact of dye-protein interactions and the associated photophysics to be directly taken into account when computing transfer efficiencies, which, on the surface, appears ideal. However, in practice, it also introduces many potentially poorly-defined parameters that may bias or confound the calculation of FRET transfer efficiencies if done incorrectly. Moreover, given fluorescent lifetimes are inherently stochastic, this necessitates sufficient sampling to capture both IDR conformational rearrangement and dye-rearrangement. Taken together, the inclusion of explicit dyes is certainly the appropriate long-term strategy. However, with the exception of a small number of groups who have pioneered the aforementioned technical and theoretical issues, in the absence of well-characterized parametrization of dye and protein force fields, it remains unclear if the additional challenges introduced by including explicit dyes is more of a help or hindrance.

A final approach is one in which simulations are performed initially without dyes, but in a *post hoc* processing step the resulting ensemble has dyes (or clouds of dyes per protein conformation) re-built

[90,140,171,278–281]. Using this approach, transfer efficiencies (or dye-dye distances) can be back-calculated. This offers a convenient middle-ground in that dye geometry and size are explicitly taken into account, yet the challenges associated with dye parameters are avoided. It does, however, operate under the assumption that the presence of dyes has no impact on the conformational ensemble explored in the simulations. Depending on the implementation details, this approach also runs the risk of over-representing conformations in which dye-attachment residues are more exposed, given only conformers where dyes can be added are included in the calculations of transfer efficiencies. Finally, this type of reconstruction requires further assumptions regarding the timescales associated with the fluorophores tumbling. The reconstruction of dye ensembles can be achieved in a number of ways and is facilitated by specific software tools [238,279,282].

The three approaches described above far operate under the assumption that simulation and experiment will agree "out of the box". In reality simulations and experiments frequently do not show quantitative (and sometimes even qualitative) agreement. This is generally taken (fairly or unfairly) to reflect weaknesses on the side of the simulations, specifically due to force field errors and/or limited sampling. One solution to this challenge is development and improvements in both force fields (as mentioned) and the development of more powerful supercomputers [283–285]. In parallel, a number of approaches for ensemble re-weighting (also known as ensemble refinement) have emerged. These reweighting strategies alter the probability of each conformation in the ensemble to shift the expected values to better match the experiment. While a mismatch between simulations and experiments is generally taken to mean the simulation is at fault, this need not necessarily be the case, and scrutiny with respect to possible experimental artifacts (*e.g.,* fluorophore quenching altering transfer efficiencies) should be taken [195].

To summarize briefly, reweighting involves the process of re-defining the probability of each conformation in an ensemble. For clarity, we define conformation here in terms of a frame or snapshot of the simulation - *i.e.,* in the case of a non-reweighted, correctly sampled set of *n* conformations taken from an NVT ensemble it is assumed that any conformation *i* selected at random from the ensemble is present with probability,

$$p_i = \frac{1}{n} \quad \text{(Eq. 25)}$$

Where, as for any discretized probability distribution,

$$\sum_{i=1}^{n} \quad p_i = 1 \quad \text{(Eq. 26)}$$

As such, the ensemble-average value for any given observable with an instantaneous value (*e.g.,* end-to-end distance, $\langle R_e \rangle$) can be computed as,

$$\langle R_e \rangle = \sum_{i=1}^{n} \quad p_i R_e^i \quad \text{(Eq. 27)}$$

where R_e^i reflects the end-to-end distance of conformation *i*. There is nothing complex about Eq. 27 - in fact, this is simply a reformatted version of the arithmetic mean. When we calculate the mean we inherently assume every element in that calculation is equally important, such that every element appears with the same probability of $1/n$. Reweighting reflects a change in this assumption where we instead re-define the probabilities such that not every element is equally likely, under the constraint that the probabilities must sum to 1.

Several key factors must be considered for ensemble refinement. Firstly, when re-weighting a large ensemble of states, we typically wish to apply systematic changes that simultaneously alter our observable to match some experiment while doing so in a manner that minimizes the loss of

entropy. As such, maximum entropy-based methods have emerged as a key component of most reweighting schemes [286–288].

During maximum entropy reweighting, the collection of conformation-specific probabilities are altered such that the resulting probability distribution of an observable matches a prescribed distribution, or the reweighted ensemble average matches some experimentally observed value.

This requirement is reached under a constraint in which probabilities must sum to 1 and the entropy *S(p)*, defined as,

$$S(p) = -\sum_{i=1}^{n} p_i ln(p_i) \qquad \text{(Eq. 28)}$$

is maximized.

Entropy-maximization does not explicitly take uncertainty into account. This uncertainty can lie on the side of the experiment in terms of precision or accuracy but can also reflect uncertainty in the simulation. This uncertainty is often considered through some kind of Bayesian approach that allows fine-tuning of uncertainty in a variety of ways [289–294]. Specifically, Bayesian inference provides a general framework through which a posterior model can be generated based on a prior model and the inclusion of newly observed data [286]. Several modern frameworks have emerged to facilitate simulation reweighting with large ensembles of disordered proteins in mind. These include Bayesian Inference of Ensembles (BioEn), Convex OPtimization for Ensembl Reweighting (COPER), Bayesian/Maximum Entropy (BME), and Extended Experimental Inferential Structure Determination (X-EISD) [289,291,292,294]. Although these tools have been recently developed and applied to disordered proteins, a large number of additional tools have been developed over the years (as

reviewed by Bonomi *et al.* [287]). An in-depth discussion of the theoretical and practical differences between these methods goes beyond the scope of this review. However, each approach offers distinct advantages and disadvantages, and in principle are compatible with the integration of multiple different types of experimental data with distinct uncertainties.

An important caveat with respect to reweighting strategies reflects the fact that these approaches are ultimately limited by the quality of the starting ensemble [295,296]. Put another way - you cannot reweight what is never observed in the original simulations. Consequently, when starting ensembles are sufficiently large and sufficiently close to reality, reweighting can be a powerful approach to fine-tune simulation results to improve the signal-to-noise. However, if a starting ensemble is sufficiently incorrect, no amount of reweighting can rescue it. The gold standard here is to include two orthogonal methods and show that re-weighting simulation results with respect to one experimental dataset improves agreement with the other [186,296]. In this context, small-angle X-ray scattering is a good complementary technique to verify reweighted ensembles generated when ensembles are reweighted based on results from single-molecule spectroscopy.

As a final note, rather than reweighting unbiased simulations to match experimentally measured distributions, an alternative set of methodologies involve applying restraints or bias terms directly to the simulation. In this approach, a cost function that penalizes conformational behavior that deviates from experimentally compatible results is applied [119,297–300]. The nature of the cost function, how it is applied over long-timescale simulations, or how experimental uncertainty is dealt with vary depending on the implementation. While this approach has been used extensively in the context of structure determination, it has been used less frequently in the context of integrating single-molecule

spectroscopy with all-atom simulations. For a comparison between restraints and reweighting in molecular simulations see work by Rangan *et al.* [301]

The preceding section introduced maximum entropy and Bayesian inference as theoretical frameworks through which reweighting or restraining can be achieved. It is worth noting that the alternative and complementary approaches including maximum parsimony, maximum likelihood, and maximum caliber provide alternative theoretical frameworks for ensemble selection and reweighting. These approaches can be applied either to bias simulations or as a *post-facto* reweighting strategy, as reviewed by Bonomi, Gaalswyk, and Ghosh, respectively [287,297,302].

2.9 Single-molecule spectroscopy and solvent quality

The impact of solvent quality on denatured proteins was evident already in early studies of protein denaturation with single-molecule FRET [162,303] as a shift in the transfer efficiency population associated with the unfolded state. The work of Sherman & Haran directly implied a coil to globule transition in the conformations of the unfolded state [183]. In this context, important early work that integrated single-molecule spectroscopy and simulations was performed by Best, Gopich, Eaton, and Schuler [236,304]. Using both all-atom MD simulations and simple coarse-grained Langevin simulations, Merchant *et al.* showed a continuous transition in global dimensions of Protein L and cold shock protein Csp*Tm* observed by smFRET is reproduced as a function of solvent quality by simulations [304]. The integration of simulation and experiment here played a crucial role in helping to interpret smFRET data by demonstrating that the inferred radius of gyration (R_g) obtained from smFRET matched the R_g values obtained from simulations. This study represents one of the earliest examples in which all-atom simulations and smFRET were combined, and in many ways, defined

the template for this class of study. Subsequent work using coarse-grained models has arrived at similar conclusions and shows good agreement with extant smFRET data [121,305].

The importance of solvent quality for disordered and unfolded proteins was again the topic of further study by Best and Schuler. In a series of papers, a comprehensive investigation of chain dimensions in response to denaturant concentration combined several different disordered proteins and a collection of methods including all-atom simulations, FCS, smFRET [89,131,271]. In work by Zheng *et al.*, unbiased all-atom simulations without explicit dyes were performed as a function of denaturant concentrations [271]. Using these ensembles, intermolecular distances were then back-calculated, revealing a modest but continuous expansion in IDR global dimensions as a function of denaturant concentration. These computational results compared favorably with analogous measurements made by smFRET and SAXS. In a separate study by Borgia & Zheng *et al.*, smFRET and SAXS data were used to reweight ensembles generated from all-atom simulations using a Bayesian approach. The resulting ensembles were compared against changes in global dimension obtained by FCS and dynamic light scattering (DLS) [89]. This study also identified a modest but meaningful chain 'contraction' as denaturant concentration is decreased (Fig. 6). In parallel, analogous integrative biophysical studies made on several other systems came to similar conclusions, supporting a model in which the solvent quality tunes the dimensions of unfolded protein ensembles, but that these ensembles do remain relatively expanded [186,306,307]. This is in reasonable agreement with measurements made by SAXS that inferred that if any chain-compaction occurred at all, it would be modest [308,309]. Taken together, these results have helped establish that as unfolded polypeptides transition from high concentrations of denaturant into native conditions, there is a sequence-dependent contraction in global and local chain dimensions. The magnitude of this

contraction depends on the chemical nature of the denaturant and protein sequence. In many foldable proteins, this contraction appears to be in the range of 10-25% in global dimensions prior to *bona fide* folding [307]. For disordered proteins the extent of compaction (or lack thereof) this contraction can range from a few percent to over 50%, depending on the amino acid sequence and denaturant [35,80,89,93,271,310,311].

Despite this substantial effort, a quantitative and absolute agreement between SAXS and FRET-derived measurements remains contentious for at least some systems [262,263,307]. Despite the valid and important concerns regarding the impact of dyes, a general consensus that disordered/unfolded proteins are sensitive to changes in their solution environment seems undeniable [307]. These conclusions need not be at odds with the observation that foldable proteins undergo a sharp folding transition when solution condition conditions permit [312].

As a final point, the magnitude, modality, and physical origin of solution-dependent changes in IDR conformational behavior will depend on the amino acid sequence and the chemical nature of the co-solute [306,313–316]. This sequence-encoded sensitivity has been proposed to offer IDRs a mechanism to act as biological actuators and sensors of cellular state [91,311].

2.10 Reconciling length-scale dependent conformational heterogeneity with smFRET and simulations

The apparent discrepancy between SAXS and smFRET has an additional possible origin: residual structure leading to deviations from homopolymer models used to infer smFRET-derived distances [90,122,123,317]. Analytical homopolymer models are remarkably good at quantitatively describing the conformational behavior of IDRs [35,76,187]. However, for IDRs with a substantial amount of residual

structure or peculiar sequence patterning, there is an expectation that homopolymer models will become progressively less reliable [120,317].

The possible impact of structural heterogeneity was examined simultaneously and independently in two studies. Song *et al.* applied simulations and theory to analyze extant smFRET data for unfolded proteins to argue that anisotropic biases in the underlying conformational ensemble could explain apparent discrepancies between SAXS and smFRET data [122–124]. Using coarse-grained simulations to construct transfer efficiency distributions, the authors show that even relatively small but persistent conformational biases can have a substantial impact on distances derived from transfer efficiencies.

In independent but complementary work, Fuertes & Ruff *et al.* performed an integrative study that combined all-atom simulations, smFRET, and SAXS of both labeled and unlabelled IDRs under native and strongly denaturing conditions. In this work, a dye-reconstruction approach was applied in which clouds of dyes were rebuilt around simulations run in the absence of dyes. A key result from this study reflects the fact that homopolymeric models are better equipped to describe conformational behavior under denaturing conditions. This result reflects the fact that in the limit of high denaturant concentration, the chain has - in effect - become a *bona fide* homopolymer. In contrast, under native conditions, sequence-dependent residual structure can lead to deviations from true homopolymeric behavior, limiting the accuracy when pairwise intra-chain distances are used to inform on global dimensions.

An analogous study by Gomes *et al.* integrated smFRET with nuclear magnetic resonance (NMR) spectroscopy, SAXS, and simulations and came to similar conclusions [267]. Here, coarse-grained

simulations in which explicit dyes with modeled photophysics were used to construct realistic transfer efficiency histograms. In agreement with Fuertes & Ruff, the authors found that integrative modeling is necessary to fully reconcile seemingly discordant observations due to local conformational biases. The need for several distinct methods that provide unfolded-state behavior across distinct length-scales has also emerged in other systems [80,186,318].

Taken together, the application of homopolymer models remains a critical tool for the analysis and interpretation of IDRs. As it turns out, the specific choice of polymer models often introduces only small systematic variations on the extracted root-means-square distances from single-molecule data [36,173]. However, underlying assumptions baked into polymer models may not hold true across various interdye distances of the protein due to long-range anisotropic interactions or local residual structure [78,90,319]. It is therefore important to test whether the assumptions associated with a given model are robust across multiple interdye distances. Polymer models can be assessed by comparing the persistence length for a wormlike chain model or the Kuhn segment for a Gaussian Chain. The origins of any observed deviations must then be examined. At the same time, heteropolymer theories often describe the local contribution of compositional heterogeneity over a specific inter-residue distance in terms of an effective bond segment that rescales the second moment of the ideal chain distribution. Different segments of the chain will adopt different effective bond lengths, such that no single effective bond length is expected to fit an entire chain. The expected heterogeneity in the effective bond lengths along a heteropolymeric protein provides a possible explanation for the empirical success of using freely jointed chain (or similar) homopolymer models on systems that are clearly far from theta-solvent conditions. As such, one should carefully consider the physical meaning of the extracted distance in the context of appropriate theories and models. In this respect,

the application of homopolymer models to the interpretation of heteropolymeric IDRs should be used under the guise of "*What is the homopolymer that best describes my data?*" as opposed to "*Does my heteropolymer behave as a homopolymer?*". [318].

2.11 Conformational dynamics as assessed by single-molecule spectroscopy and simulations

The ability of single-molecule spectroscopy to provide direct insight into the molecular dynamics of a given IDR has opened up additional avenues of experimental characterization and comparison between simulation and experiment.

Soranno *et al.* combined simulations, single-molecule spectroscopy, and theory to build a complete molecular dissection of the determinants of internal friction in unfolded proteins [168,185]. By combining smFRET and ns-FCS, the authors were able to probe how fast chain dynamics depends on the interdye sequence length and solvent viscosity, demonstrating that under native condition protein dynamics are often not dictated only by solvent conditions, but more significantly by internal friction effects, where internal refers to intrinsic properties of the protein, such as transient intramolecular interactions and dihedral angle constraints. These results were in remarkable agreement with extant simulation data performed by Piana *et al.* [126,185]. Moreover, the conclusions drawn in this study were further confirmed via integrative analysis of alpha-synuclein dynamics using smFRET, NMR, and MD simulations [275]. The integration of simulation and experiments provided a comprehensive molecular readout that implicates non-local intramolecular interactions and a second contribution from the retardation of dihedral rotation, although these two effects may be inherently coupled.

Integrating smFRET with simulations allowed Metskas & Rhoades to reconcile apparent discrepancies between published structures of the intrinsically disordered C terminal domain of troponin-1 [320]. Multiple high-resolution structures lacked agreement with each other and with NMR based measurements, highlighting the conformational heterogeneity that exists in the system. MD simulations performed with discordant published structures as starting points allowed them to gain an understanding of the conformational landscape the protein adopted. Interestingly, although good agreement between smFRET measurements and MD simulations was obtained when comparing folded subregions, substantial disagreement was arrived at when smFRET measurements of the intrinsically disordered C terminal domain were compared with MD simulations. Hypothesizing that this discrepancy reflected a difference in the timescales of the techniques, the authors applied MC simulations to construct a large ensemble of conformations for the disordered region. This ensemble showed good agreement between smFRET, MD, and MC simulations, and the most populated conformations present in the MC simulations matched the three published structures that were 'incongruent.' This study elegantly demonstrates that if distinct timescales are probed, it is possible to obtain apparently contradictory yet entirely valid results.

Zosel *et al.* integrated extensive single-molecule fluorescence data and all-atom simulations to assess complex binding kinetics between the disordered protein ACTR and its conformationally heterogeneous folded partner NCBD [79]. Single-molecule experiments revealed that an evolutionarily conserved proline in NCBD undergoes slow cis-trans isomerization. The binding affinity of NCBD for ACTR depends heavily on the isomerization state of this slow-switching proline. MD simulations provided a cogent molecular explanation for the proline-dependent affinities and demonstrated that the molecular structure of the bound complexes differs depending on the proline

isomerization state. The ability to reconcile complex and counterintuitive kinetic behavior was entirely dependent on the ability to observe conformational rearrangement on a range of timescales and length scales. Similarly, the ability to offer a cogent structural explanation for this behavior rests on the application of molecular simulations to the binding event. Taken together, this study offers an example in which simulations and experiments offer complementary insights into the structure and dynamics of a complex molecular mechanism.

Medina *et al.* utilized MD simulations paired with smFRET and hydrogen-deuterium exchange mass spectrometry to probe the conformational heterogeneity and dynamics present within intermediate binding steps of FOXP1 [321]. This approach allowed the authors to probe low-population conformations that would be hidden if ensemble experiments were used exclusively. By applying single-molecule spectroscopy and simulations, this complex structural landscape was disentangled, enabling the development of a model in which domain switching involves intermediate states populated by a heterogeneous population of conformations.

Finally, while not strictly an IDR, Chung *et al.* utilized a combination of long-timescale simulations and single-molecule spectroscopy to determine the physical basis for slow protein folding in a small triple-helix designed protein [219]. By first analyzing photon trajectories from FRET histograms using a maximum likelihood method [322] to obtain relaxation rates, the authors reveal a sharp pH dependence on the folding rates, where folding is dramatically faster at low pH. A similar pH dependence on folding is also observed in all-atom molecular dynamics simulations. By strengthening or weakening the non-bonded interactions associated with salt bridges by altering the underlying forcefield, the authors are able to perform a computational experiment to decouple the observed rate effects on

salt-bridge strength vs. net charge of the molecule. This ingenious analysis revealed that salt-bridge strength is the key determinant of the transition time, providing a clear example in which the types of theoretical experiments that simulations afford offers direct insight into a physical process that would otherwise be impossible to measure.

2.12 Multi-molecular assemblies as measured by single-molecule spectroscopy and simulations.

The integration of single-molecule spectroscopy and simulations has more recently played key roles in providing a high-resolution window into dynamic protein:protein and protein:RNA complexes [56,281,323]. Ensemble methods typically hide the heterogeneous nature of IDPs, masking dynamic interactions that may underlie biological function. In a series of papers exploring polyelectrolytic complexes, the integration of smFRET, nsFCS, and MD simulations has been essential to deconvolve complex heterogeneous systems.

In a landmark study, Borgia, Borgia, & Bugge *et al.* demonstrated that a binary complex formed between the negative polyelectrolyte prothymosin alpha (ProTα) and the positive polyelectrolyte linker histone H1.0 (H1) formed a high-affinity complex in which both proteins remain fully disordered [56]. Using a bespoke coarse-grained model that is directly compared against 28 distinct intra- and inter-molecular distances measured by smFRET, the authors demonstrate remarkably good agreement and provide a comprehensive molecular picture of the resulting high-affinity complex. Importantly, on the experimental side, the authors compare results with two different sets of dye pairs, and on the computational side, simulations are run both with and without explicit dyes. In addition to smFRET and simulations, extensive NMR data corroborate the disordered nature of the complex and provide additional key insights.

In two subsequent studies, Holmstrom and Heiðarsson & Mercadante probed the dynamic nature of intrinsically disordered proteins in the context of protein:protein, protein:RNA, and protein:DNA interaction [323,324]. In both of these studies, single-molecule spectroscopy was combined with coarse-grained MD simulations were able to capture the dynamic nature of the association of an IDP with another protein or nucleic acid. In the bound state, the IDP in question remains both disordered and dynamic upon association with its ligand, where this dynamic association underlies the biological function. For Holmstrom *et al.* this dynamic association enhanced the folding of an RNA hairpin, providing an electrostatic screening effect analogous to high concentrations of monovalent salts. For Heiðarsson & Mercadante *et al.* a ternary electrostatic competition mechanism driven through a dynamic protein assembly enabled the dissociation of Histone H1 from the nucleosome.

In addition to providing insight into individual molecules or small complexes in a dilute solution, single-molecule fluorescence spectroscopy can be used to peer into the interior of biomolecular condensates formed through liquid-liquid phase separation [80,325,326]. Martin, Holehouse, & Peran *et al.* combined turbidity, FCS, and coarse-grained simulations to calculate full phase diagrams of the low-complexity domain of the RNA binding protein hnRNPA1 [80]. More broadly, both smFRET and FCS offer a means to examine the conformational behavior of IDRs inside and around phase-separated droplets [325,326].

2.13 The application of simulations and single-molecule spectroscopy to offer molecular insight into biophysical mechanism

The true power of integrating molecular simulations with single-molecule spectroscopies lies in the ability to uncover novel biophysical mechanisms. In our final results section, we consider a

collection of studies in which specific molecular details have been unraveled through the combination of single-molecule fluorescence spectroscopy and simulations.

A long-standing question in cell biology pertains to the molecular basis of recognition and translocation of nuclear transport receptors by the phenylalanine and glycine-rich (FG) disordered regions that line the interior of the nuclear pore complex [327–330]. An integrative study by Milles & Mercadante *et al.* combined all-atom simulations with smFRET, NMR, and SAXS to offer a direct molecular picture of the nature of FG interactions with their associated cargo proteins [331]. This work revealed a degenerate network of transient molecular contacts between a nuclear pore protein and its corresponding nuclear transport receptors. These interactions were encoded by distributed adhesive phenylalanine residues in FG motifs where they interact in a multivalent fashion across the surface of the cognate transportin proteins. Despite the lack of specific binding sites and the microscopically weak binding affinities of individual motifs, the resulting macroscopic binding affinity is remarkably high. As such, nuclear transport receptors are tightly bound, yet relatively free to diffuse. This work provides a molecular explanation for the selective partitioning and rapid translocation of transportin-bound cargo proteins across the nuclear pore complex.

The physical basis for temperature-induced collapse of disordered and unfolded proteins has been examined via smFRET interpreted via all-atom replica exchange molecular dynamics simulations, pointing to the role of sidechain solvation in driving compaction [234]. This observation was confirmed in subsequent work where temperature-dependent free energies of solvation were used with all-atom implicit-solvent Monte Carlo (MC) simulations to explain corresponding smFRET experiments for a number of different IDRs [332]. In both cases, unbiased simulations without explicit dyes were

performed and the radius of gyration (R_g) from simulations compared with the apparent R_g calculated from smFRET-derived inter-dye distances.

Beyond these classic examples, there are many cases in which single-molecule spectroscopy and simulations have been combined to address specific mechanistic questions. In the context of protein folding, all-atom MD simulations have been used to identify transient non-native salt bridges that are the dominant determinant of transition-path times along the folding barrier [219]. All-atom simulations have been used in conjunction with smFRET of aggregation-prone polyglutamine (polyQ) to demonstrate that – contrary to naive expectation – the biophysical behavior of polyglutamine tracts do not show a discontinuous transition as polyQ length extends between physiological and disease-associated lengths [140,193]. In a similar vein, residual structure in the monomeric state of the aggregation-prone amyloid-beta peptide was examined through an in-depth study that combined smFRET will all-atom MD simulations where explicit dyes were included [198]. By combining FCS and simulations, Mao *et al.* demonstrated that the sequence net charge plays a crucial role in determining the global dimensions of disordered regions [272]. Similarly, FCS, smFRET, and simulations help demonstrate that sufficiently long polyglutamine and polyglycine repeats undergo chain collapse to form compact yet heterogeneous ensembles [140,193,306]. By combining MD simulations and an extensive set of smFRET experiments, Vancraenenbroeck *et al.* showed that IDR-binding affinity can be directly modulated by solution-dependent changes to conformational behavior, hinting at a complex, environmental-dependent protein:protein interaction network inside cells, a conclusions supported by more recent work that combines simulations and ensemble FRET [76,311].

IDRs are frequently involved in molecular recognition, and single-molecule spectroscopy and molecular simulations are well-poised to provide molecular detail on those interactions. A crucial aspect of microtubule function in axons is their ability to undergo dynamic instability, where they experience periods of elongation and depolymerization, a process that is highly regulated by a family of intrinsically disordered Tau proteins[333,334]. To better understand the first step of microtubule assembly, where tau protein binds soluble tubulin heterodimers, Melo *et al.* completed an extensive mapping of free and tubulin-bound tau conformations using smFRET [335]. Subsequently, they generated an ensemble of possible tau conformations using Monte Carlo simulations constrained by distances generated from their smFRET measurements. When modeled in proximity of coarse grained tubulin dimers it was possible to visualize how tau binding to multiple dimers could be accomplished. Importantly, this gave insight into the dynamic nature of the interaction. Instead of adopting a fixed structure upon tubulin binding, a "fuzzy complex" is observed, where the disordered nature of Tau allows for the binding of multiple tubulin dimers and highlighted the significance of conformational flexibility upon binding, a phenomena later seen with other IDP binding interactions as well [56,323,324].

Finally, in an integrative study that combined MD and MC simulations with single-molecule spectroscopy, Cubuk *et al.* performed a comprehensive dissection of the three disordered regions in the SARS-CoV-2 nucleocapsid protein [187]. This work revealed distinct structural features that provide a molecular explanation for several previously described binding interactions.

In short, the ability to ascribe atomistic-level insight from simulations with analogous observations for a specific subset of intramolecular distances affords high-resolution physical descriptions of complex phenomena in a way that most other techniques do not.

2.14 Discussion and Conclusions

The integration of single-molecule spectroscopy and simulations has emerged as a fruitful approach to provide molecular insight into the complex and heterogeneous behavior of disordered proteins. A recurrent theme in many of the studies described above is the need to consider a range of length-scales and time-scales to construct a holistic understanding of IDR conformational behavior. While smFRET provides high spatial accuracy and precision with respect to specific pairs of distances, it is largely blind to conformational behavior that occurs distally to the labeling positions. In contrast, while simulations provide high-precision insight into both global and local conformational behavior, they are limited by possible force field or sampling inaccuracies. As such, the most comprehensive – and arguably informative – studies are those in which smFRET empowers confidence in the simulations (either by confirming simulated results or providing a means to refine them), which in turn allows simulations to report on features that are not directly captured by smFRET [56,89,90,185,219,304,323]. When smFRET and simulations can be combined to make predictions that can be tested via orthogonal methods such as FCS, SAXS, NMR, DLS, or any additional method, the accuracy of inferences made through integrative studies can be directly assessed [56,126,267,291,331].

Despite substantial successes, several open challenges remain for the effective integration of single-molecule spectroscopy and simulations. A significant challenge is the need for better methods to describe dyes and their photophysics. A number of groups have pioneered work in this arena, yet

despite notable successes, the inclusion of dyes in all-atom or coarse-grained simulation simulations is by no means standard practice [56,173,212,277]. As mentioned in the introduction, large heterocyclic dyes are inherently challenging for fixed-charge force fields due to their aromatic nature. The emergence of polarizable force fields offers a potential solution to this challenge [336–339]. While in fixed-charge all-atom force fields, each atom has a fixed partial charge, in polarizable force fields the local charge density is responsive and variable, depending on the local chemical environment. As a result, polarizable dyes models may offer a more realistic route to describe their physicochemical effects and, potentially, help identify scenarios in which protein:dye interactions are likely. Beyond facilitating better interpretation of smFRET data, an accurate and transferable description of fluorescent dyes would allow experimental groups to computationally screen distinct pairs of dyes to help identify those which are least likely to interact with a given protein. While polarizable models (such as AMOEBA) have historically been viewed as substantially slower than fixed-charge models, recent major efforts to improve performance have yielded simulation times on the order 10-30 ns/day in AMOEBA [340,341]. As a result timescales relevant for comparison with single-molecule spectroscopy are firmly within reach, suggesting further application of polarizable forcefields is a promising future avenue.

A more general challenge for simulations of disordered proteins represents robust methods for the quantification and assessment of conformation sampling. While limitations in standard molecular force fields persist with respect to disordered proteins, even if a perfect forcefield existed, it would not guarantee that accurate estimates of chain conformations and dynamics could be reached. Recent work from Lincoff *et al.* has argued that while over compaction of standard force fields when describing IDRs is a known problem if better conformational sampling was available, some of the

force field limitations may be less severe than they appear [342,343]. This is not to suggest that forcefield limitations are overblown, but simply to urge a critical assessment of local and global conformational heterogeneity when performing molecular simulations of disordered proteins. Simulations of a few hundred nanoseconds are rarely sufficient for even modestly sized disordered proteins. General best-practices for assessing conformational sampling in IDRs are lacking but would help to guide researchers to understand if poor agreement between simulation and experiment is due to forcefield weaknesses, insufficient conformational sampling, or a combination of the two.

The integration of single-molecule fluorescence spectroscopy and all-atom simulations has been instrumental in our modern understanding of sequence-encoded conformational behavior in disordered proteins. As more advanced methods for multi-dimensional data integration emerge, integrative studies in which multiple experimental techniques are used to better understand a specific system will likely become more commonplace and more effective. The ability to obtain insight over multiple length-scales and timescales is an essential feature that integrative studies provide. For disordered proteins especially, the need to consider a range of length scales and timescales reflects the inherently heterogeneous and stochastic nature of the conformational transition. Given the fact that molecular simulations and single-molecule fluorescence spectroscopy offer a comparative spatial and temporal resolution, they are an inherently complementary and powerful combination.

Figures

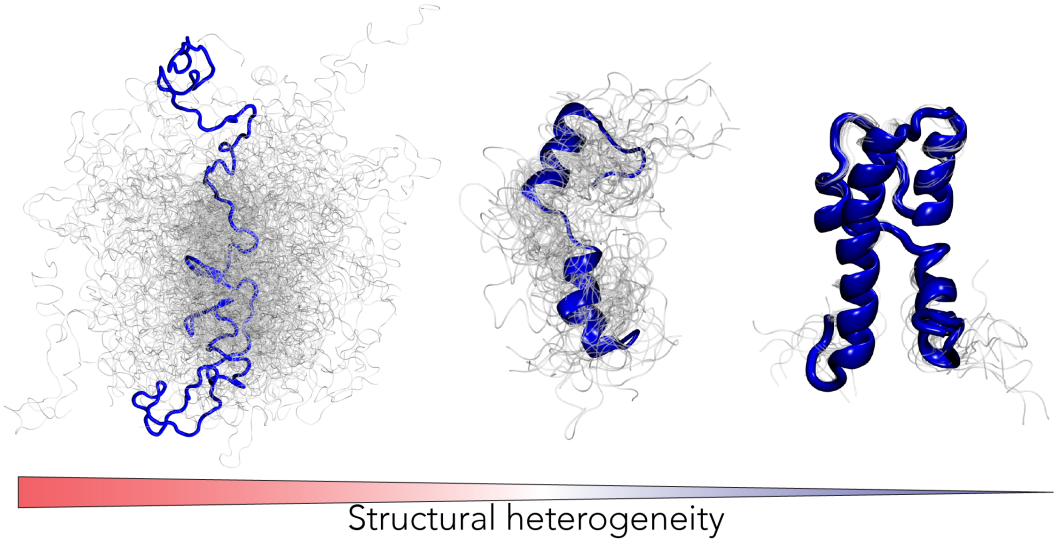


Figure 1. Proteins exist along a continuum of structural heterogeneity.

While some proteins adopt well-defined tertiary structures (far right), intrinsically disordered protein regions (IDRs) lack a defined reference state (far left). Importantly, all proteins are defined by an ensemble, where function is ultimately determined by the combination of chain dynamics and preferential conformations [69,70,344]. IDRs are not fundamentally different from folded proteins but are distinguished by conformational fluctuations so large that a single native-state reference frame is no longer applicable nor useful.

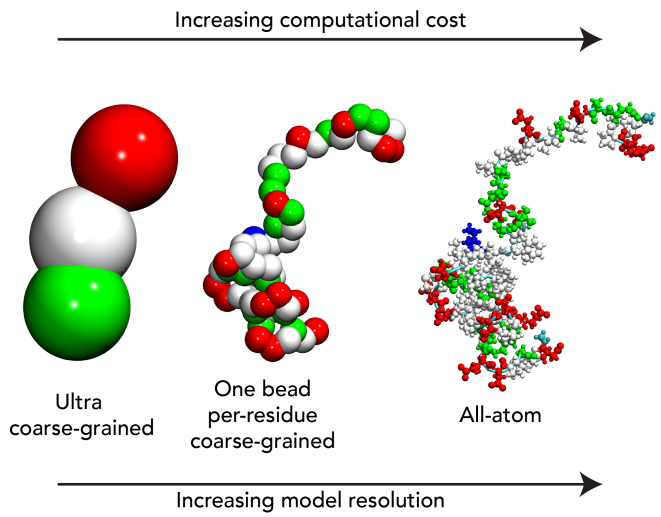


Figure 2. Examples of distinct levels of granularity of the representation schemes.

As the number of degrees of freedom increases (from left to right), as does the computational cost. In principle, more degrees of freedom should yield a higher accuracy model, although this depends on the actual fidelity of the model. A model with many degrees of freedom is only more accurate if those degrees of freedom and described correctly.

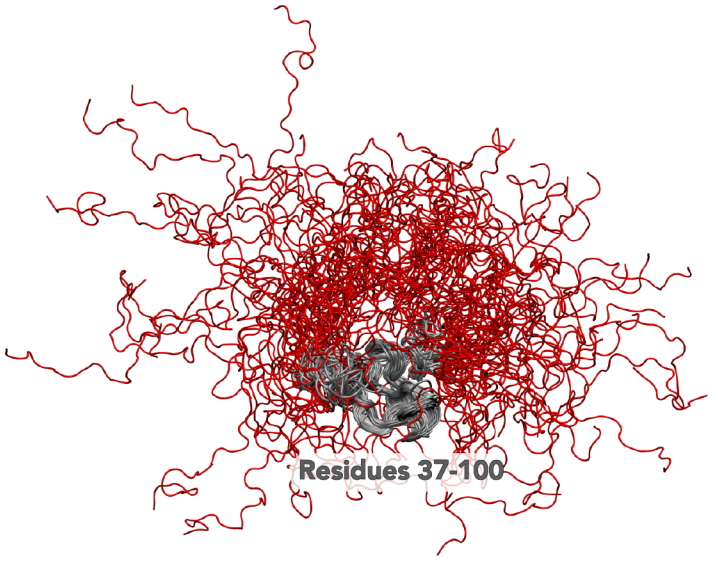


Figure 3. Schematic of Disordered Protein

Snapshots taken from a simulation trajectory of α-synuclein reveal a scenario in which a subregion of the protein is kinetically trapped while the N and C-termini explore a diverse collection of conformational states.

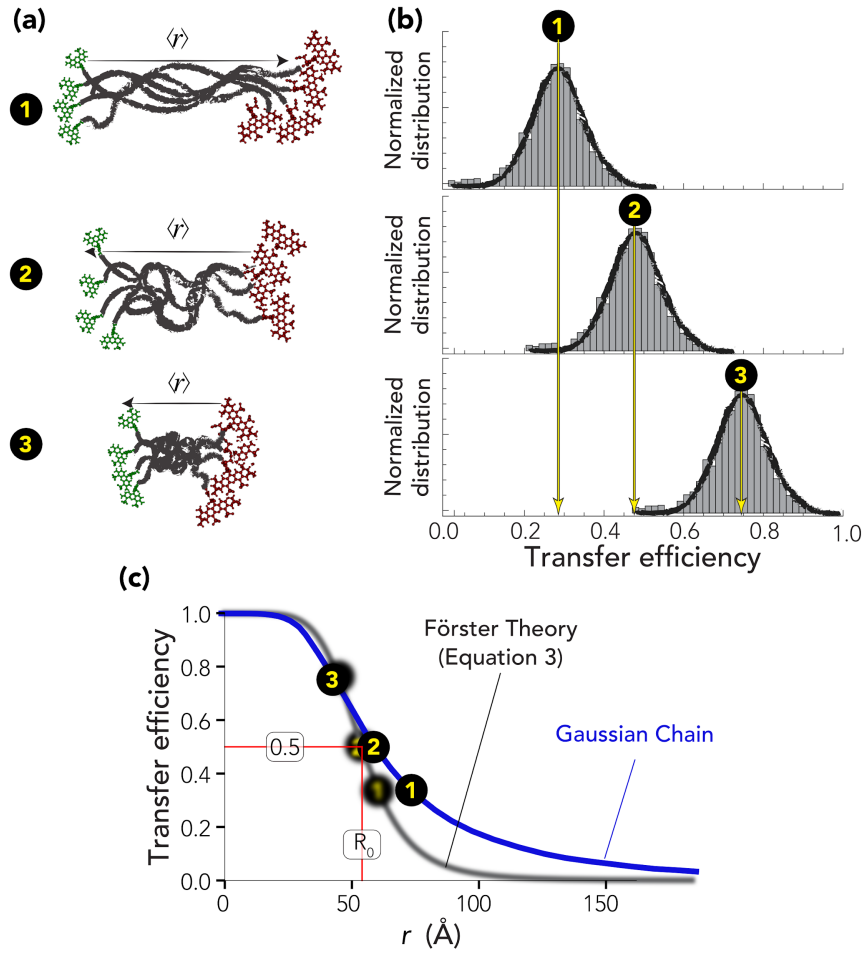


Figure 4. Overview of single-molecule FRET experiment and data

(a) Schematic representation of disordered proteins with different mean end-to-end distances. **(b)** Histograms of photon bursts for the hypothetical ensembles in corresponding panels in (a). **(c)** The black curve represents the dependence of the mean transfer efficiency on the inter-dye distance as predicted by Förster's theory (eq. 3), shown with conformations annotated. The blue curve depicts the transfer efficiency of a fluctuating Gaussian chain as a function of the average root mean square inter-dye distance.

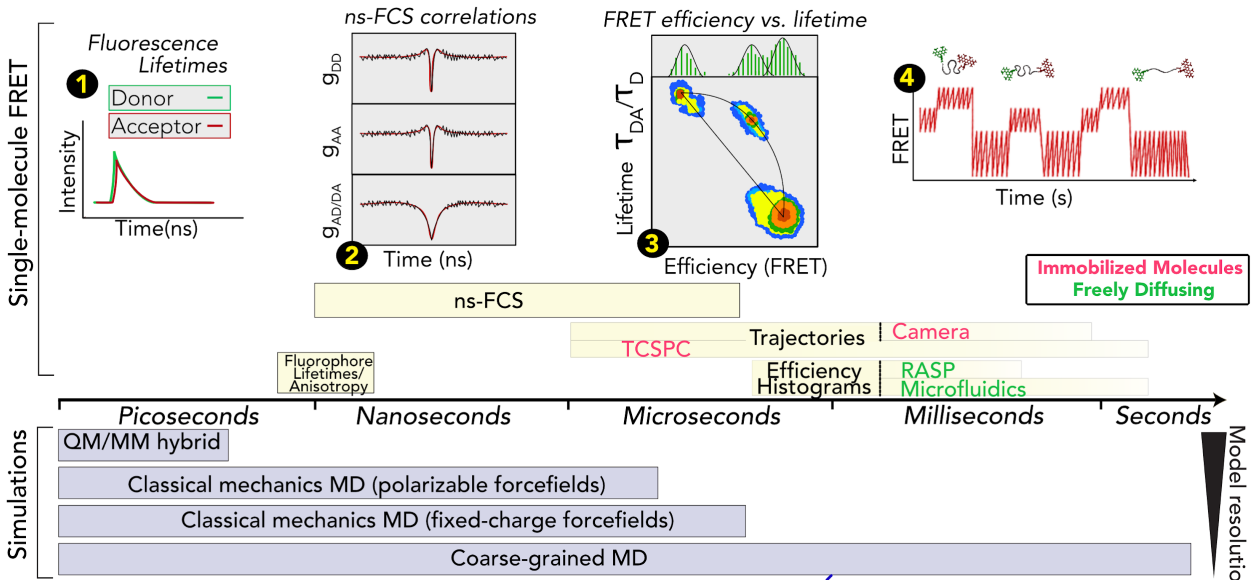


Figure 5. Experiments and simulations inform over a broad range of timescales.

Schematic detailing the different timescales accessible to single-molecule fluorescence spectroscopy and simulations. **(1)** Time-resolved fluorescence provides access to donor and acceptor lifetimes (which are influenced by the FRET process) and to anisotropies (which reports about tumbling of the dyes and of the overall molecule. **(2)** The correlate decay in the donor (DD) and acceptor (AA) autocorrelations as well as the anticorrelated rise in the donor-acceptor (AD/DA) cross-correlation reports about protein dynamics. **(3)** 2D-histogram of donor lifetime in the presence of the acceptor (normalized by the donor lifetime in the absence of the acceptor) vs. transfer efficiency. The diagonal line represents the result for a static configuration of the protein and the curved line represents dynamics exchange in the protein conformational ensemble. **(4)** Transfer efficiency trajectory of immobilized molecules can reveal slow conformational changes of the protein up to minutes.

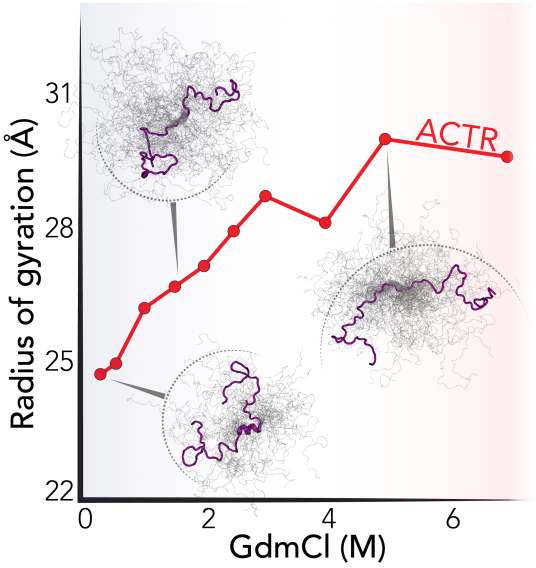


Figure 6. The conformational ensemble of the 71-residue ACTR

The conformational ensemble of the 71-residue ACTR as a function of denaturant, as obtained from smFRET and all-atom simulation by Borgia & Zheng *et al.* [89].

Chapter 3: Condensation goes viral: a polymer physics perspective

This chapter was published in the Journal of Molecular Biology as:
Alston JJ, Soranno A. Condensation Goes Viral: A Polymer Physics Perspective. *J Mol Biol.* 2023 Jan 26:167988. doi: 10.1016/j.jmb.2023.167988. Epub ahead of print. PMID: 36709795.

Author Contributions. J.J.A and A.S. conceived and wrote the manuscript. J.J.A. performed and analyzed simulations used as exemplifications of the models.

3.1 Abstract

The past decade has seen a revolution in our understanding of how the cellular environment is organized, where an incredible body of work has provided new insights into the role played by membraneless organelles. These rapid advancements have been made possible by an increasing awareness of the peculiar physical properties that give rise to such bodies and the complex biology that enables their function. Viral infections are not extraneous to this. Indeed, in host cells, viruses can harness existing membraneless compartments or, even, induce the formation of new ones. By hijacking the cellular machinery, these intracellular bodies can assist in the replication, assembly, and packaging of the viral genome as well as in the escape of the cellular immune response. Here, we provide a perspective on the fundamental polymer physics concepts that may help connect and interpret the different observed phenomena, ranging from the condensation of viral genomes to the phase separation of multicomponent solutions. We complement the discussion of the physical basis with a description of biophysical methods that can provide quantitative insights for testing and developing theoretical and computational models.

3.2 Introduction

Viruses exploit various cellular membrane-bound organelles in their life cycle as a means of viral entry, genome replication, assembly, and egress, as well as as a means of evading the host immune system[345–350]. Often, these compartments are hijacked and undergo dynamic reorganization to aid viral proliferation[348,349]. Efforts to understand the ability of viruses to utilize host organelles have led to the design of drugs that deliberately target these processes[351]. During the past decade, it has been recognized that some viruses can also harness host-membraneless organelles and form new biomolecular condensates as a part of their life cycle, utilizing these compartments to mediate activities such as transcription, genome replication, and host immune system evasion. In normal contexts, biomolecular condensates often act as a molecular reservoir for specific proteins and RNA, while in aberrant contexts they can sequester key proteins needed for a specific cellular response[352]. Sequestration could play a key role for viruses in the evasion of the cellular immune response by partitioning host cell immune-response factors into condensates that impede their function[353–355]. Furthermore, understanding the permeability of viral condensates to immune response proteins such as pattern recognition receptors can offer novel ways of targeting viruses[356]. The large constellation of phenomena emerging on the cellular level requires, on one hand, the adoption of a new set of concepts for describing and rationalizing experimental observations and, on the other hand, the development of new technologies for investigating mechanisms and testing existing models. Here, we provide a polymer physics perspective on the mechanisms controlling the condensation of proteins and nucleic acids, which can help to sort out and interpret some of the different phenotypes in biomolecular condensates related to viruses. We complement the theoretical description with a brief discussion of current enabling methodologies that provide access to

fundamental observables to quantify and connect molecular interactions and mesoscopic observables, from single-molecule to demixed solutions.

3.3 What is a biomolecular condensate?

According to the empirical definition introduced by Banani *et al.* [357] and recently reprised by Mittag and Pappu [358], condensates are membraneless entities that concentrate a given set of components (e.g., proteins, nucleic acids, metabolites…) in non-stoichiometric complexes[359]. While the term biomolecular condensate is often automatically linked to phase separation and liquid-liquid demixing, the observations of an increased concentration of components, *in vitro* or *within the cell*, does not necessarily imply a single mechanism of action or specific material properties [358]. As such, the term "condensate" captures different emerging phenomena on the mesoscale (e.g, phase separation and percolation), which all give rise to condensation. As we will discuss below, part of the ambiguity is inherent to the fact that many of the underlying mechanisms are or may be interrelated. Identification of the specific mechanism at play requires a precise interrogation of the molecular properties of the object, with particular respect to the concentration and temperature dependence of its components, which are a common result of the thermodynamics of the system. Note that measurements of transport properties (as often quantified by recovery after photobleaching experiments) only highlight the kinetics within the object, which are indirectly linked to the equilibrium properties of the system and, as such, cannot prove a specific mechanism[358,360,361].

A common question surrounding the formation of biomolecular condensates pertains to their ability to select for specific components and filter out others, controlling their overall internal composition and buffering the concentration of species outside. The role of viral components in biomolecular

condensates brings this concept to the extreme, since the virus introduces extraneous viral components into the infected cell that may invade existing condensates or form new ones. Understanding the rules that control the mechanism of selection (or exclusion) of these components from the cellular environment can play an essential part in decoding how the virus succeeds in hijacking the cellular machinery.

3.4 Viruses and Biomolecular condensates.

Lacking their own complete replication machinery, viruses require the infection of a host cell to assist them in the replication of their genomic material. With few exceptions, the nature of the genome dictates the localization of replication, which for DNA-based viruses happens in the nucleus whereas for RNA-based viruses occurs in the cytosol. Viruses with a double-stranded DNA or RNA genome rely on host polymerase proteins or their own RNA-dependent RNA-polymerase (RdRp) to transcribe the viral genome. The same occurs for single-stranded negative-sense RNA genomes, which also requires the assistance of RdRps. In contrast, single-stranded positive-sense RNA viruses can directly harness the host cell ribosomes to translate viral protein components. Another essential step in the life cycle of the virus is the correct packaging of the viral genome after replication. Interestingly, this requires organization and often compaction of large nucleic acids into relatively small volumes.[362] This can be further complicated, since some viral genomes consist of multiple independent nucleic acid molecules, which must be packaged within the same virion ("segmented") or within distinct virions ("multipartite") [362,363].

In the following sections, we summarize three notable cases that exemplify how viruses can harness biomolecular condensates for: i) viral genome replication, ii) hijacking the stress machinery of the cell, iii) packaging the viral genome.

3.5 Viral factories for genome replication.

The Mononegavirales family are non-segmented single-stranded negative-sense RNA viruses, including rabies virus, Ebola virus, measles virus, mumps virus, and human respiratory syncytial virus[364]. Infection of host cells with this family of viruses leads to the accumulation of cytoplasmic inclusion bodies within the cell. While some of these bodies have been known for a long time and even used as diagnostic markers (such as the Negri Bodies for rabies virus[365]), only later was it noted that these compartments contained both viral RNAs and viral and host proteins[366,367]. The presence of essential components for viral replication supports the idea that these inclusion bodies are *bona fide* "viral factories". Among the components, there are the RdRp necessary for transcribing the viral genome and Nucleoproteins. Nucleoprotein coats and protects the viral genome and is among one of the highest expressed viral components after infection. The current hypothesis is that at low concentrations, the Nucleoproteins protect the genomic RNA and allows RdRp to transcribe subgenomic RNA; at sufficiently high concentrations, the Nucleoprotein promotes phase separation and, by stabilizing the polymerase, inhibits termination and favors synthesis of genome-long RNA.[368] In other words, the phase separation of Nucleoproteins (also called anti-terminator proteins) may play the key-role of enriching the protein concentration and enabling the switch between subgenomic and genome-long transcription of mRNA.

Viral factories have since been found in the host cell post infection for a plethora of viruses, including other single-stranded negative-sense RNA viruses (such as vesicular stomatitis virus) and double-stranded RNA viruses (such as rotaviruses)[369–373]. Indeed, as mentioned above, double-stranded RNA viruses face similar challenges as single-stranded negative RNA viruses, since they cannot directly interact with host cell ribosomes and require the intervention of a viral polymerase. Therefore, it is plausible that these biomolecular condensates may serve an analogous function.

Overall, the round shape of viral factories[372–374], their ability to fuse[369,371–373,375] to deform against physical barriers[375], their response to osmotic stress[375], and their ability to exchange material with the surrounding milieu [369,371–373,375] supports that these micron sized objects[365,366,376] are likely the results of an intracellular phase separation. Interestingly, these inclusion bodies do not fuse with other stress granules found in the cell[375], suggesting that the viral components encode for precise compartmentalization of these "factories" from other condensates.

3.6 Hijacking stress granules.

Viruses can also exploit existing intracellular condensates and their components, as in the case of stress granules. As implied by the name, stress granules regulate the translational machinery of the cell in response to various stress factors, harboring several ribonucleoprotein complexes (including arrested pre-initiation complexes), mRNAs, and related translational initiation factors[377,378]. Therefore, it is not surprising to observe that their assembly and function is hijacked upon viral infection. For example, upon infection by mammalian orthoreovirus (double-stranded RNA virus), the host's protein expression is shut-down and cellular mRNAs accumulate in stress granules, where they are maintained transcriptionally inactive[379]. Interestingly, stress granules arise after viral

disassembly (uncoating) inside the host cell, but independently from viral transcription[379]. The accumulation of typical stress granule markers like G3BP1 and TIA-1 and the dose response to phosphorylation of eIF2α [379] suggests that viral components trigger the assembly of these stress granules, repurposing their function. Indeed, when virus-induced stress granules are disassembled because they are no longer necessary to the virus, further formation of stress granules in the cell is inhibited, even under exposure of infected cells to stress factors such as arsenite[380].

Similar observations are found also for single-stranded positive-sense RNA viruses such as the Semliki Forest virus[381], Polio virus[382–384], and Hepatitis C virus[385], though each exhibit different phenotypes regarding viral transcription and recruitment or exclusion of specific components.

3.7 Viral genome packaging.

Packaging of the viral genome poses a key challenge for viruses, requiring the condensation of large nucleic acids or the combination of multiple segments. Various strategies have evolved to achieve this result and some intersect with the formation of and partitioning into biomolecular condensates. For example, the Influenza A Virus (single-stranded negative-sense RNA virus) has a segmented genome containing eight distinct elements that are unusually replicated in the host cell nucleus and then transported to the cytosol in the form of viral ribonucleoprotein complexes, each comprising a single segment of genome[386]. These ribonucleoprotein complexes colocalize with intracellular foci [387–389] that form near the endoplasmic reticulum exit sites and exhibit spherical shape, fusion and fission events, fast exchange of components with the surrounding cytosol and rapid response to environmental stimuli [388]. While these biomolecular condensates do not appear to contribute to

evasion of cellular innate antiviral response, they have been proposed to facilitate the organization of the different genome segments for proper assembly and export at the plasma membrane[388].

Many proteins involved in viral packaging have been found to easily partition into membraneless organelles. After infection with the Measles virus (single-stranded negative-sense RNA virus), the nucleoprotein and phosphoprotein, which are responsible for packaging the viral genome, colocalize within puncta identified as transcription factories. *In vitro* experiments support that both the nucleoproteins and phosphoproteins can co-phase-separate in solution and, upon addition of RNA, facilitate the assembly of nucleocapsid-like particles at a significantly higher rate than the one observed in absence of phase separation[390]. These observations pose an interesting question on which type of interactions facilitate the formation of nucleocapsid-like particles within the complex environment of a condensate.

Recent experiments on the SARS-CoV-2 nucleocapsid protein, which is responsible for coating and condensing the viral genome[362], also reported colocalization in cytosolic inclusions[391,392]. While there is evidence that genomic transcription happens in membrane-bound compartments, the strong propensity for phase separation of nucleocapsid protein poses a question on how its recruitment to the correct nucleic acid is controlled and how phase separation of the viral genome in the cytoplasm is avoided. This does not necessarily imply that a phase separation mechanism should be at play; conversely it can be seen as the necessity of mechanisms in place to instead avoid or control the phase separation propensity of certain components[393].

How viruses impart specificity to packaging their own genomic nucleic acids, while excluding host cell nucleic acids, and their own subgenomic nucleic acids has been a longstanding question. We will discuss some simple models below.

Further reading. While these are some notable cases, additional examples and further details on the functional role of biomolecular condensates upon viral infection can be found in the following recent reviews: [368,394–397].

3.8 A polymer theory framework.

Understanding how different components can lead to the assembly of biomolecular condensates or alter the function of existing ones requires accounting for the mechanisms that can lead to solution demixing. Proteins and nucleic acids are biological polymers each composed of a specific set of monomers, their fundamental units (residues and nucleotides, respectively). Given their polymeric nature, the language of polymer physics offers a simple framework to explain the driving forces controlling solution demixing concepts. The framework we present here is through the lens of Flory-Huggins solution theory[398].

Let's start by considering the case of a homopolymer, such as a nucleic acid composed of one single type of nucleotide. While this is an obvious oversimplification of what happens in realistic protein and nucleic acid sequences, it is the simplest case scenario where we can define the key quantities of interest. Note that often homopolymer theories can be applied to capture important features of a heteropolymer sequence (like a "real" protein or nucleic acid) because the properties encoded in the heteropolymer averages out on sufficiently long distances.

What controls the dimensions of such a biomolecule in a solvent (for example, an aqueous buffer solution)? What causes a group of molecules to aggregate or segregate from others? It is reasonable to assume that the properties of a biomolecule in that solvent (dimensions, aggregation propensity, demixing, etc…) will be controlled by the set of interactions between the molecule and the solvent or between the molecule and another molecule. More precisely, the overall set of interactions controlling a polymeric molecule is encoded in each monomer. Therefore, in the case of a homopolymer, we can define specific contributions for the monomer-solvent interaction (ε_{ms}) and monomer-monomer interaction (ε_{mm}). For comparison, it useful also to consider a parameter that accounts for solvent-solvent interactions (ε_{ss}). The behavior of the homopolymer in the solution will be dictated by the balance across these interactions. Stronger monomer-solvent interactions than monomer-monomer interactions will favor the interaction of the homopolymer with the solvent; in the opposite case, with stronger monomer-monomer interactions than monomer-solvent, this will favor intrachain interactions. In polymer physics, there is a key-parameter that accounts for this balance and this is commonly referred as the χ interaction parameter:

$$\chi = \left(\frac{2\varepsilon_{ms}-\varepsilon_{mm}-\varepsilon_{ss}}{k_B T}\right) \quad \text{(Eq. 1)}$$

where k_B is the Boltzman constant and T is the temperature.

This is the only equation within this review and up to this point, we have focused only on the contribution of interactions. However, one additional key factor is the homopolymer concentration.

3.9 Biopolymers in dilute conditions.

When the concentration of homopolymers is so low that interactions between different molecules can be neglected, we can consider the polymer to be under dilute conditions. The only interactions at play are intramolecular contacts and χ defines whether the polymer chain undergoes compaction or expansion. When favorable solvent-monomer interactions dominate over monomer-monomer interactions, the polymer maximizes its interaction with the solvent by adopting expanded configurations. Because the solvent interactions are favorable, the solvent is referred to as "a good solvent". One practical example of biopolymers in good solvent is given by the conformations of a single-stranded nucleic acid, which, in the absence of ligands, adopts expanded conformations. The negative net charge of the chain (unfavorable monomer-monomer interactions) and the favorable interactions between nucleotides and aqueous buffer solution lead to expanded configurations. A similar behavior is seen for highly charged disordered proteins (the net charge of the chain disfavors monomer-monomer interactions) and for proteins in high concentrations of denaturants (denaturing solvents favors the interactions between residues and solvent) [35,37,231,399,400].

Viceversa, when monomer-monomer interactions dominate, the polymer minimizes its exposure to the solvent by adopting compact configurations. In this case, the solvent is defined as a "poor solvent". Examples of biopolymers in poor solvent are disordered proteins dominated by hydrophobic interactions and folded domains.

It is important to note that according to this definition, the same solvent (e.g. an aqueous buffer) can be both a good and a poor solvent depending on the polymer that one is studying.

There is a third case to consider, which is the case where monomer-solvent interactions exactly counterbalance the contribution of solvent-solvent and monomer-monomer interactions. In this case, the χ parameter defined in **Eq. 1** is equal to zero. This condition defines the transition limit between the good and poor solvent conditions. In this scenario, the solvent is considered a "theta solvent" or "ideal solvent". Interestingly, many disordered proteins sit close to this interface, which makes them very sensitive to the surrounding environment[35]. Therefore, by tuning the solvent quality (e.g. by altering the strength of interactions via temperature), a dilute molecule can undergo a coil-to-globule transition from compact configurations in poor solvent to expanded configurations in good solvent[401].

In **Fig. 1** we exemplify the common result of Flory-Huggins model with a bead model of the polymer where we tune the strength of the interactions between each bead and the solvent. Notable examples of "tuning the strength of interactions" are ion screening of electrostatic interactions of nucleic acids and charged disordered proteins and the contribution of kosmotropic or chaotropic agents [231,399].

3.10 Biopolymers and phase separation.

What happens when the homopolymers are no longer in dilute conditions? With increasing concentration of molecules, intermolecular interactions become relevant and monomer-monomer interactions can occur between different polymer chains. In this scenario, the solution can either remain well-mixed and contain an increasingly dense phase of polymers or undergo demixing (see **Fig. 1**). Demixing occurs when the total free energy of the solution is better minimized by partitioning the polymer into a dilute and dense phase rather than maintaining a single well-mixed phase. In the case of a homopolymer, at a given constant temperature, the concentration of the

dilute and dense phase are fixed (determined by the free energy and chemical potential of the polymer-solvent mixture) and only the relative fraction of each phase changes with increasing concentration of the polymer. This is a key expectation of a single homopolymer system and it is of relevance when characterizing phase separation of single components *in vitro*. In other words, when increasing the concentration within the demixing range, the homopolymer will partition in two phases, one with a lower and one with a higher concentration of homopolymer. The concentration of these two phases remains constant independently of the total concentration of homopolymer. Indeed, the concentration in the lower concentration phase is equivalent to the saturation concentration (i.e., the minimum concentration at which phase separation can occur), whereas the one in the higher concentration phase is equivalent to the maximum concentration at which phase separation can occur. While the concentrations remain constant, the relative abundance of each phase is dictated by the total concentration of homopolymer. Therefore, it is reasonable to expect a small volume of the dense phase when the total concentration of homopolymer in solution is close to the saturation concentration and vice versa.

If we further increase the concentration of molecules beyond the dense phase boundary, we will re-enter into a well-mixed phase that is now characterized by a concentrated solution of the polymer. These three cases (dilute solution, demixed solution, concentrated solution) constitute the basic predictions of Flory-Huggins theory. It is interesting to note that all of these different conditions are the result of the same set of interactions and different phases emerge as a function of the total concentration of the polymer in the solution.

The theory also provides more insights on the conditions under which phase separation can occur. The Flory-Huggins theory identifies critical χ values at which phase separation occurs, which depends on the length of the polymer. However, there is no specific prescription regarding the exact strength of monomer-monomer, solvent-solvent, and monomer-solvent interactions besides the constraints imposed to satisfy at least this critical χ value. This is important since often weak interactions have been invoked in the context of biomolecular condensates as a prerequisite, but there is no such requirement for phase separation to happen; rather, the nature of the interaction (repulsive or attractive) and the balance between the polymer and solvent components favors or disfavors demixing.

One important element that emerges from the Flory-Huggins framework is that the same set of interactions controls the conformations of polymers under dilute conditions as well as their phase separation propensity. After determining conformational properties of a disordered region in isolation (*via* NMR and small-angle X-ray scattering) and corresponding phase separation boundaries (*via* absorbance and fluorescence correlation spectroscopy), Martin et al.[80] used the Flory-Huggins theoretical model and simulations to extract the molecular interactions and confirm the interconnection between these two phenomena, including the role of valence and patterning of specific residues. Another notable example are elastin-like peptide sequences, whose single chain conformations and phase separation have been extensively characterized and compared [402–404]. This intrinsic connection between the single chain and multi-chain behavior, which is encoded in the monomer interactions, has enabled the development of physics-driven models that capture the phase-separation propensity of disordered proteins and nucleic acids[80,113,405–408]. Overall, these

examples demonstrate that the implications of the Flory-Huggins model maintain validity when applied to biopolymers and can be deployed to rationalize the mechanism at action.

The Flory-Huggins model can be further complicated to account for the heterogeneous nature of the sequence, starting from the realization that specific monomers may be "stickier" than others. The theory of associative polymers proposed by Semenov and Rubinstein[409] has offered a robust scaffold to conceptualize the role of "stickers" (elements that encode for associative interactions) and "spacers" (elements that do not contribute or contribute minimally to association)[359,410] in controlling phase separation propensity. In the context of disordered proteins, distributed π-systems (e.g., π-π and cation-π interactions)[80,411–420], oppositely charged residues[181,326,421–427], and hydrophobic aromatic and aliphatic residues[423,428] have emerged as different typology of stickers. The nature of stickers encodes for context-dependent response of the phase behavior, enabling them to be modulated by environmental factors such as pH, ion screening, and post-translational modifications[410]. The number and patterning of stickers contribute a further layer of tuning for the phase separation behavior[80,429,430]. While spacers may not be directly involved in key associative interactions between chains that control phase-separation, they should not be regarded as intrinsically inert as they also encode for specific monomer-monomer and monomer-solvent interactions. Consequently, modulation of the spacer composition can alter the free energy of mixing and control phase behavior (concentrations, dynamics, and material properties)[94,429,431,432]. Importantly, addition of adhesive contacts between the monomers can give rise to percolation networks through the solution [409], where percolation represents a "networking transition" compared to the "density transition" observed in phase separation [358]: as such, the two phenomena can be disjointed (giving rise to distinct phases) or coupled (creating networks within condensates and

modulating viscoelastic properties of the material) [359,433–437]. While these concepts have been extensively explored in the context of disordered proteins, nothing precludes their application to entire protein domains, with folded domains acting as sticker elements and disordered regions flanking the folded domains acting as spacers[438].

3.11 Viral genome condensation: one vs many.

We have previously mentioned the problem of viral genome condensation and the phase separation propensity of specific components such as the viral nucleoproteins or nucleocapsids. The problem of nucleic acid condensation and the interplay between single chain condensation and multichain phase separation has been longly discussed in literature[439]. The Flory-Huggins polymer framework suggests a direct connection between the case of a single polymer condensation (a necessary step in the packaging of viral genome) and the phase separation of many polymers. In this respect, it is interesting to look back to the work of Post and Zimm published in 1982[440]. At that time, direct measurements of single nucleic acid conformations were not possible and the authors embraced expectations from Flory-Huggins theory of a polymer solution and proposed to quantify the association of multiple nucleic acid molecules (phase separation) as a way to measure intrachain interactions. As they stated, "*It is not, of course possible to measure the behavior of a single polymer molecule experimentally; therefore, one must consider the consequences of intermolecular associations on the free energy of the system, recognizing the possibility of an aggregated polymer state*." Indeed, in the framework of Flory-Huggins theory, the same interactions control the intermolecular association of multiple chains and the intramolecular interactions leading to a single chain expansion or collapse. Post and Zimm added an important hypothesis to the Flory-Huggins theory, incorporating in the χ parameter the interactions of other molecules that control the condensation of the large polymer. In this respect the phase

diagram in (**Fig. 1A)** acquires a new meaning, where changes in the χ-parameter are now modulated by ligand concentrations and specificity, favoring or disfavoring single chain compaction or multi-chain phase separation.

It is interesting to explore the implications of this model in a viral context. In the case of coronaviruses, compaction of the viral genome is driven by the interplay between a single-stranded nucleic acid and many copies of the Nucleocapsid protein. Our group and many others [392,441–443] have observed a strong propensity of the SARS-CoV-2 nucleocapsid protein in promoting phase separation when interacting with nucleic acids, both *in vitro* and *in living cells*. The extreme phase separation promiscuity of this protein poses an interesting conceptual problem: how is the protein rescued and recruited on the correct nucleic acid? To address this question, we used a simple polymer bead model (the same described above in the Flory-Huggins paragraph) and asked how binding of a multivalent ligand (the nucleocapsid protein) to large polymers (the nucleic acids) can lead to phase separation of the solution[441]. In absence of specific interactions, with increasing concentration of the multivalent ligand, we identified a set of conditions under which the solution undergoes phase separation (**Fig. 2**). In other words, we observe a partitioning of the solution in a dilute and dense phase, the first one depleted of polymers and ligands, the second one enriched in both. This is largely consistent with the picture proposed by Post and Zimm[440] and experimental observations for SARS-CoV-2 nucleocapsid protein in presence of non-specific nucleic acids [392,441–443]. However, the viral genome encodes for specific interactions along its sequence and it has been hypothesized that particular binding sites with high affinity help in packaging the viral genome. These sites are often referred to as "packaging signals" and have been previously proposed in the context of retroviruses, beta-coronaviruses, mammalian C-type viruses, and influenza A virus[444–448].

To understand the contribution of packaging signals, we tested the effect of introducing a high affinity site in the equivalent of the RNA sequence in our simulation[441]. Interestingly, we found that a single high affinity site is sufficient to alter the phase separation propensity of the system, suppressing phase separation and facilitating condensation of individual chains[441]. While the model is clearly simplistic, it suggests a possible mechanism of action by which nucleic acid with high affinity motifs can attract the nucleocapsid protein and avoid phase separation with nonspecific sequences (**Fig. 3A**). In additon, phase separation propensity can be modulated by interactions with specific protein components[449], which could alter the nucleic acid and protein-component multivalence and possibly disfavor the demixing of the solution [393]. Once the single genome is condensed, a further transition can occur, leading to a structured organization of the nucleic acid (helical, smectic, etc…) or inducing a more disordered arrangement similar to beads on a string, as in the case of SARS-CoV-2 genome[450] (**Fig. 3B**).

3.12 Multi-components solution demixing.

The picture emerging from the Post and Zimm theory operates a strong simplification by "hiding" the binding of the ligand within the interaction χ parameter. Indeed, the theory does not explicitly account for its dependence on the concentration of ligands. The explicit treatment of these effects requires expanding the Flory-Huggins theory to include multiple species[451,452]. For the typical case of a ternary solution, including two homopolymers and a solvent, the model predicts different scenarios depending on the set of interactions and concentrations of each component. For example, if each polymer only phase separates on its own (i.e. obligate homotypic interaction), this reduces the problem to the case described above for a single type of homopolymer. Alternatively, as observed for the majority of protein and nucleic acid interactions, if the two polymers have

favorable cross-interactions and demix together,the two-phase region of the phase diagram has a closed-loop topology (see **Fig. 4**). Tie-lines identify the co-existing concentrations of the dilute and dense phase and all the intermediate concentrations at which the solution will partition at those specific concentrations. In the two-dimensional space identified by the concentrations of the two polymers, the slope of the tie lines indicates whether there is a preferential partitioning of a polymer with or versus the other. Tie lines parallel to one of the axes imply that one of the two polymers has no preferential partitioning between the two phases; positive slopes indicate an enrichment of both polymers in the dense phase, whereas a negative slope reveals segregation of one component with respect to the other.

An obvious implication of the ternary (or higher components) Flory-Huggins model is that, differently from the case of a single homopolymer and solvent, the concentration of polymers in the light and dense phase depends on the total concentration of each polymer. Indeed, if we keep the concentration of one polymer component constant and we increase the concentration of the other polymer within the phase-boundaries of demixing, we will cross many different tie lines, which implies that different concentrations of each polymer occur in the light and dense phase. While this may partially limit the buffering capacity of condensates[453], it is possible that condensate-driving components have evolved to conserve specific buffering capacity within the range of concentrations accessible in the cell. Similarly, viral invading components may be optimized to specifically favor the solution demixing of existing components. Note that in this case preservation of a specific function does not imply a strict amino acid sequence conservation, as has been shown already for functional evolution of disordered proteins[454,455]. Therefore, a deeper understanding of the molecular grammar

dictating assembly and emerging properties of condensates will provide predictive tools for classifying proteins and molecules that may favor or disfavor condensation.

It is important to remark how the original Flory-Huggins theory that we discussed does incorporate only preferential interactions between the monomeric components and does not account for chemical binding that may affect the conformational properties of components or lead to persistent contacts. This is partially addressed by the theory of associate polymers that we discussed above, where explicit attractive contacts can be introduced[456,457]. The theory can easily be extended to account for distinct associative components. Recently, Nandi *et al.* [458] have included explicit treatments that account for polynomial binding in the Flory-Huggins framework by constructing independent lattices of defined interaction for each component, with the model qualitatively reproducing experimental observations.

From a biological point of view, comparing these different models teaches us that different types of interactions (e.g monomer solubility and preferential interactions, binding, ion screening or ion condensation) contribute differently to the mixing free energy of the system. This can be particularly important in the context of testing and designing drugs for targeting condensates. Modulating solubility of the protein may modulate partitioning of small molecules of interest, whereas altering the physico-chemical properties of condensates (e.g by increasing crosslinking) may affect their function. Indeed, the recent finding that drugs can cause hardening of viral intracellular bodies and block viral replication suggests these as viable routes to identify therapeutic strategy[459].

3.13 Accessing condensation: from single-molecules to phase separation.

Forty years after the Post and Zimm paper[440], there is now a broad array of single-molecule methodologies that enables the investigation of single nucleic acid condensation. While single-molecule approaches can provide important insights on the molecular interactions at play, ensemble methods such as Nuclear Magnetic Resonance and small angle x-ray scattering (SAXS) also enable access to essential complementary information on protein and nucleic acid conformations and interactions. At the same time, the incredible interest for phase separation has led to the development of new methods to characterize phase boundaries as well as transport and rheological properties of condensates. Rather than a comprehensive review of each of the techniques, which is outside of the scope of this review, we decided to highlight the principal biophysical methods that can be used to study the interactions occurring at the level of single-molecules (as for the condensation of single nucleic acids) and the phase separation of many. In particular, we exemplify their application to investigate viral nucleic acid and, when this is not possible, discuss their potential application based on current experiments on other systems to study viruses moving forward.

Turbidity measurements

Turbidity experiments can generate a quantitative description of condensate phase diagram boundaries[460–462]. This technique takes advantage of the light scattered by condensates when the system is above its saturation concentration, i.e. the concentration where the components separate into a dense and dilute phase. As such, turbidity is commonly measured at wavelengths where no absorption is expected from the sample and the decrease in light transmission reflects the opacity of the demixed solution. When paired with centrifugation of the sample, which enables separating the

dense and dilute phases, it allows for careful measurements of the concentration in each phase and construction of tie lines[463]. Phase separation of the SARS-CoV-2 nucleocapsid protein and corresponding phase boundaries have been studied using turbidity assays[392,441,443,464]. A limitation of the approach pertains to the impossibility of distinguishing phase separation from aggregation or percolation and inspection of the sample *via* imaging is required to confirm the underlying assembly process.

Imaging via light microscopy

Light microscopy imaging provides one of the most common methods for identifying the formation of biomolecular condensates, both *in vitro* and *in living cells*. Macroscopic demixing *in vitro* can often be visualized without any specific labeling. However, the identification of specific components participating in the condensate, particularly within the cells, is helped by fluorescent labels, which are either covalently attached to a protein and/or nucleic acid or genetically encoded. The condensate formation is then studied as a function of solution conditions: ion concentration, pH, crowding agents, temperature, and obviously molecular concentration[465]. To access qualitative phase boundaries, images are taken as conditions are perturbed and the presence (or absence) of condensates is visualized to quantify entrance into and exit from the two-phase regime[466–468]. Microscopy can also provide insights into the partitioning of protein and nucleic acid components within the condensates, as in the case of the partitioning of SARS-CoV-2 nucleocapsid protein into stress granules[469].

Fluorescence Recovery After Photobleaching

A common approach that is combined with fluorescence microscopy is the assessment of molecular mobilities within biomolecular condensates and with the surrounding milieu via fluorescence recovery after photobleaching (FRAP)[191,361,470]. In typical FRAP experiments, fluorescently labeled molecules are photobleached using a high-power laser. The mobility of molecules within the bleached spot is quantified by monitoring the change in fluorescence signal before and after photobleaching and by following its recovery, which can be even just partial. Though it provides one of the easiest approaches to assess transport properties within the condensate, a careful interpretation of FRAP results is required to identify the mechanisms at play on the molecular scale and disentangle diffusive effects from kinetically-trapped states[358,361,471].

Fluorescence correlation spectroscopy (FCS)

FCS provides a method to determine both the concentration and the diffusion of molecules by measuring the fluctuations of the fluorescence signal within the confocal spot. As such, FCS enables quantifying both concentration and stoichiometry of species in the dilute and dense phase of a demixing solution as well as the mobility of molecules within and outside the condensates[80,361,472,473]. The resolution of FCS and the concentration regime in which it can be applied makes FCS a perfect approach to connect properties of the phase separated solution to the properties of single nucleic acid compaction [474]. Compaction of nucleic acids can be visualized by quantifying the change in hydrodynamic radius of fluorescently labeled molecules as a function of solution conditions, proteins, and crowders[475–477]. Importantly, FCS can be easily performed both *in vitro* and *in cell* [478]. In the context of nucleic acid condensation, Sabanayagam *et al.* used FCS to measure Bacteriophage T4 DNA packaging, utilizing changes in diffusion to assess translocation of bacteriophage DNA from

the bulk solution to the interior of the bacteriophage prohead[479]. Gopal *et al.* showed that viral RNAs are more compact on average than non-viral RNA, alluding to a potential evolutionary reduction in dimensions that could aid in packaging genomic RNA into relatively small virions[480]. By treating RNA molecules as branched polymers, FCS measurements of the hydrodynamic radii of long RNAs have also been used to help predict RNA dimensions from secondary structure predictions[481].

Förster resonance energy transfer (FRET)

FRET provides a 'spectroscopic ruler' to measure distance changes within the molecule of interest[232,482], taking advantage of the distance-dependent non-radiative energy transfer between the donor and acceptor fluorophores. Ensemble and single-molecule FRET methods have been largely used to study nucleic acid condensation and conformational changes of proteins upon binding to the nucleic acid[217]. Like FCS, FRET can be performed both *in vitro* and *in cell* [483,484], including its single-molecule applications[485,486]. FRET measurements can also be applied to proteins within biomolecular condensates[427,487,488]. FRET has been applied to study DNA and RNA bending and folding in viruses: oftentimes, viral nucleocapsid proteins are highly positively charged and act as macromolecular counterions, facilitating charge screening and enabling proper folding of the nucleic acid[324,489]. Taken further, these charged proteins could also be utilized to shift nucleic acid protein-solvent interactions and facilitate condensation and viral genome packaging.

Single-molecule force spectroscopy

Another class of techniques that offers robust characterization of nucleic acid compaction are single-molecule micromanipulation techniques which include optical and magnetic tweezers as well as atomic force microscopy. These three techniques enable not only quantitative observables of nucleic acid compaction but also report on the forces that enable compaction[490–492]. These single-molecule micromanipulation techniques have been used to measure nucleic acid compaction as a function of pH and ionic strength[493–497], protein induced condensation[498–500], and crowding agent effects[501]. In particular, the combination of fluorescence and force spectroscopy enabled quantifying the forces generated in protein:nucleic acid co-condensates and determining how those forces can act on non-condensate localized nucleic acids to mediate DNA condensation[499,502]. In regards to viral nucleic acid compaction, Gien *et al.* were able to understand how the HIV-1 nucleocapsid protein mediates viral genomic DNA compaction using a combination of optical tweezers and atomic force microscopy[500]. Optical trapping techniques can also be used to study the microrheology of *in-vitro* reconstituted condensates, which can provide quantitative insights into the condensate material properties[503]. Indeed, optical traps enable the physical manipulation of individual droplets and the study of droplet fusion, providing insights on surface tension properties of different biomolecular condensates and kinetics of fusion [325,467,503–506]. In addition, as demonstrated by Jawert *et al.* [505], optical traps can be harnessed to measure the viscoelastic moduli within a single droplet. This is realized by trapping two polystyrene beads within the condensate and by monitoring the perturbation experienced by one bead when displacing the second bead at different frequencies [505].

Nucleic acid curtains

Curtains enable high throughput single-molecule measurements of nucleic acid compaction[507–510]. This approach requires the attachment of long fluorescently labeled nucleic acids to a surface (e.g. by using supported lipid bilayers) within a microfluidics flow cell. The flow stretches the nucleic acid, while other components (protein or cations) can complex with the nucleic acid. Labeling of the components enables direct visualization of the effects of solution conditions or protein concentrations on nucleic acid compaction. Calcines-Cruz *et al.* utilized DNA curtains to develop viro-mimetic scaffolds and programmable bioinspired nanomaterials, which can be used for studying how sequence specificity enables packaging of viral nucleic acids [511].

Nuclear Magnetic Resonance (NMR) Spectroscopy

NMR spectroscopy enables the interrogation of intermolecular interactions with atomistic resolution. By probing the interactions of atomic nuclei with a magnetic field, information on the chemical environment being experienced by each nucleus can be determined[512]. Due to its ability to capture the structures of conformationally dynamic and heterogeneous molecules, in both physiological and non-physiological conditions, NMR has been used extensively to study the structure of viral nucleic acid elements[513–520] viral proteins[521,522], and the role of disordered regions[523,524]. The amino acid residues involved in nucleic acid binding can be determined by chemical shift perturbation[525–527], giving insight into residues that may be important for phase separation or nucleic acid condensation[392]. Furthermore, NMR spectra of biomolecules within condensates[94,528–530] offer the opportunity to understand how condensate environments affect proteins and nucleic acids.

Small-angle X-ray scattering (SAXS)

SAXS reports on the size, shape, conformation, and molecular weight of molecules in solution[531,532]. Its ability to measure compaction and expansion of molecules makes it a useful technique for studying nucleic acid condensation, and has been applied particularly for studying the effects of RNA folding[533–536], DNA compaction [537–541], and RNA compaction[542]. It can also be used to study the conformations of proteins undergoing phase separation[543].

Computational Methods

When it comes to interrogating the molecular driving forces for nucleic acid condensation and phase separation, simulations can provide insights that may be challenging to extract from experimental measurements. Simulations rely on a representation scheme of the biomolecules of interest and of its interactions [97,544]. These are encoded by a force field that either describes each component at atomic resolution (all-atom) or as a coarse-grained model that approximates the scale of observation. For simulating macroscopic behaviors, such as peptide- and protein-induced nucleic acid condensation or phase separation, coarse grained models are typically favored over all atom representations because of the decreased computational effort[110], though extensive all-atom simulations may provide details on parameters that are difficult to coarse-grain, such as the contribution of ion screening and counterion adsorption[545–550]. Recent years have seen the development of coarse-grained force-fields aimed to describing biomolecular condensates. At the ultra coarse-grained level, there are lattice-based models such as LASSI[408,551] and PIMMS[552], where multiple residues or even entire domains are described as single beads: both have been successfully used to describe the phase separation of mixtures of nucleic acids and proteins. Less coarse grained models utilize one bead per nucleotide or residue representations of biomolecules. These include the

hydropathy scale (HPS) model[110,113,325,326,553–558], the Kim/Hummer (KH) model[110], and the multi-scale π-π (Mpipi) model[405]. The HPS, KH, and Mpipi models each contain bonded, electrostatic, and short range pairwise-interaction terms. For the HPS model amino acid specific pairwise interactions are encoded by a hydrophobicity scale[553,559] and parametrized against SAXS and FRET measurements of intrinsically disordered regions. The KH model is based on amino acid pairwise interactions, which are derived from experimental data (second virial coefficient of lysozyme and binding affinity of the ubiquitin-CUE complex)[560]. Mpipi differs in that its sequence specific pairwise potentials are based on potential of mean force calculations of all-atom simulations of amino acid and amino acid:nucleic acid pairs in explicit solvent with ions. It also accounts for π-π interactions based on the frequency of π-π interactions that occur in a set of PDB structures[561]. Mpipi quantitatively predicts the radius of gyration of disordered proteins and was able to recapitulate experimental observables such as multiphasic droplets of polyarginine, polylysine, and polyU RNA[405,562].

The ability to recapitulate experimental findings with simulations provides the ability to systematically interrogate sequence features of disordered proteins and nucleic acids that may drive phase separation and/or nucleic acid compaction. Development of robust models tested and supported by experimental observations may enable better understanding of the sequence to function relationship between condensate components.

3.14 Recent advances on protein-nucleic acid coacervation

An increasing number of studies has turned to address the role of protein-nucleic acid condensates in regulating the cell machinery, in particular exploring nuclear organization418,460,502,555–557.

While many of these studies do not concern viral components, they well exemplify the relevant interactions that will drive phase separation of viral proteins. Here, we briefly summarize some of the biophysical concepts that have emerged from the recent study of protein-RNA condensates, focusing on general concepts that can be readily applied to viral components.

Because of the negatively charged nature of nucleic acids, one of the obvious driving forces of phase separation is the interaction with positively charged proteins. The phase separation of oppositely charged polymers in polymer physics is commonly referred as "complex coacervation"[563–565] and concerns the electrostatic attractions between oppositely charged groups as much as the role of ions condensing on the highly charged polymers. The specific base of the nucleic acid has a direct impact on the morphology of the condensate, with poly-rG sequences leading to amorphous condensates compared to other liquid-like condensates obtained for other homo-oligonucleotides[566]. Transient interactions due to base stacking also modulate dynamics of exchange within the coacervate [566]. When mediating the interaction with protein, the length dependence of the interaction (short range if involving cation-pi interactions, long range if purely electrostatic) further contribute in modulating the phase boundaries [567]. The balance of interactions in phase transition of RNA−protein complexes can lead even to ordered hollow condensates [325,421]. The stoichiometry and length of the nucleic acid further modulate the surface tension of the condensate [568] and its internal rearrangements [393]. Analogously, the pattern of charged residues in the sequence alters the viscoelastic properties of coacervates [569].

3.15 Conclusions

While the past decade has seen growing evidence that viral infections harness intracellular condensates, the challenge that lies ahead is in understanding the rules controlling the participation of viral components to specific condensates and uncovering the mechanisms related to the emerging phenotypes. The role of viral proteins and nucleic acids in modulating cell condensates provides an intriguing perspective on how "external" components can modulate phase separation propensity, shifting phase boundaries of protein and nucleic acids in the cytoplasm or nucleus. It may also provide important insights into how evolution has encoded for robustness in cellular compartmentalization and whether viral genomes that are highly mutated conserve specific properties related to phase separation. At the same time, the large promiscuity identified for specific components such as the nucleocapsid protein of SARS-CoV-2 poses the complementary challenge of understanding which cellular or biochemical factors may suppress or limit the phase separation propensity of viral proteins.

Bridging the molecular and mesoscopic scales and decoding the phase separation grammar encoded in proteins and nucleic acids will require both top-bottom and bottom-up approaches. Top-bottom approaches where condensates are studied within the cellular complexity are essential to ensure that findings are pertinent to the function of the compartments, but are limited in granting access to molecular details. Rigorous investigation of intracellular puncta are necessary to understand the nature of the observed objects and the role of each component. Bottom-up approaches may be limited in testing biomolecular function, but are essential for precise measurements of the interactions at play. Direct measurement of interactions in both well-mixed and dense phases are

necessary for precise models that account for different modes of binding and stoichiometries, in particular if disordered regions are involved.

Undoubtedly, the complexity of the cellular environment highly complicates the simplified models presented here, with multiple components coexisting and demixing at the same time. However, a better understanding of the interactions connecting intra- and inter-molecular chain interactions will help validate theories and computational models of phase separation. This is particularly important since a direct test of all possible condensate components may be impracticable, but theoretical predictions and computational models may offer a strategy to explore buffering capacity and response to extraneous components.

Condensates are more than the sum of the parts and determination of the collective properties of viral condensates (e.g., surface tension, viscoelasticity, etc…) may provide important insights on how these components alter the normal function of these intracellular bodies. At the same time, investigation of viral condensates will enable development of new therapeutics aimed to target these specific components during the virus life cycle.

3.16 Acknowledgements.

The contribution of J.J.A and A.S. to this work is supported by the NIH National Institute on Allergic and Infectious Diseases (Soranno, R01AI163142) and the National Cancer Institute (Alston, F99CA264413). We thank Alex Holehouse and Melissa D Stuchell Brereton for insightful discussions.

Conflict of interests. The authors declare no conflict of interests.

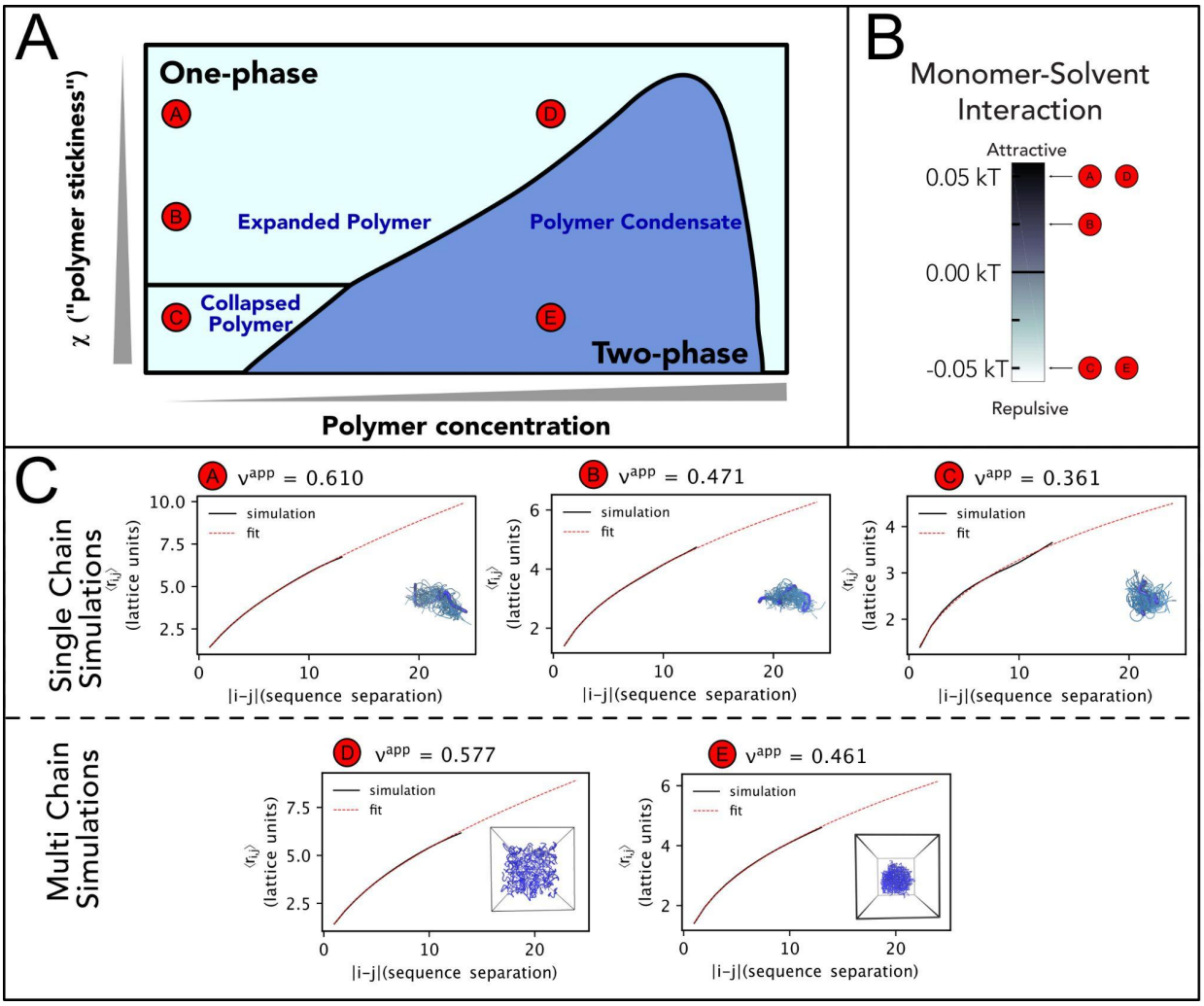


Figure 1. Flory Huggins Theory expectations

A. Representation of the phase diagram proposed by Post and Zimm in their "Theory of DNA Condensation"[440]. Here we compare it with the behavior of a 15-bead homopolymer in lattice-based PIMMS simulations[407] and study how solution conditions affect single polymer conformations and multi-chain interactions. We report polymer concentration on the x-axis and polymer stickiness (as represented by χ, **Eq. 1**) on the y-axis. At low concentrations, the polymer can adopt different degrees of collapse or expansions depending on the χ parameter. With increasing concentrations,

depending on the polymer stickiness, the solution can undergo demixing and separate in two phases. **B.** In coarse grained simulations, χ can be modulated by altering polymer-solvent interactions, whereas polymer-polymer and solvent-solvent (implicit) interactions remain fixed throughout all simulations. The graph depicts the corresponding pairwise interaction energy between each polymer bead (the monomer unit) and the solvent for a given solution condition. **C.** We compare results obtained for single-chain and multi-chain. *Single-polymer simulations* show a change in the scaling exponent ν as a function of internal distance $|i-j|$ and χ. Under solution conditions that favor monomer-solvent interactions, we observe a scaling exponent ν close to ⅗, as expected for a polymer in good solvent. However, by making the monomer-solvent interactions more and more unfavorable, we observe a decrease in the scaling exponent till reaching the configuration of a collapsed globule, with a scaling exponent ν equal to $\sim 1/3$, consistent with poor solvent conditions. It is interesting to note that the theta-solvent condition (ν equal to ~½) is not achieved when the monomer-solvent interactions are equal to zero, but when they are favorably attractive. This reflects the inherent contribution of monomer-monomer and solvent-solvent interactions to the χ parameter. In *multi-chain simulations*, we observe two different scenarios. If solvent-monomer conditions are favorable, but monomer-monomer interactions are unfavorable, the homopolymer exists as a homogenous solution of overlapping polymers where each individual polymer has a certain degree of expansion. In the specific case simulated above, the scaling exponent ν is equal to .577. However, altering the delicate balance of monomer-monomer and solvent-monomer interactions can lead to solution demixing. Of note, the scaling factor has shifted from 0.361 to 0.461. As the polymer forms a condensate, it 'solvates' itself with other polymers. This result mimics expectations for the scaling exponent of a polymer in a melt, where the scaling exponent is ½.

Interestingly, a similar result is expected also for the expanded chains at sufficiently high concentration[570,571].

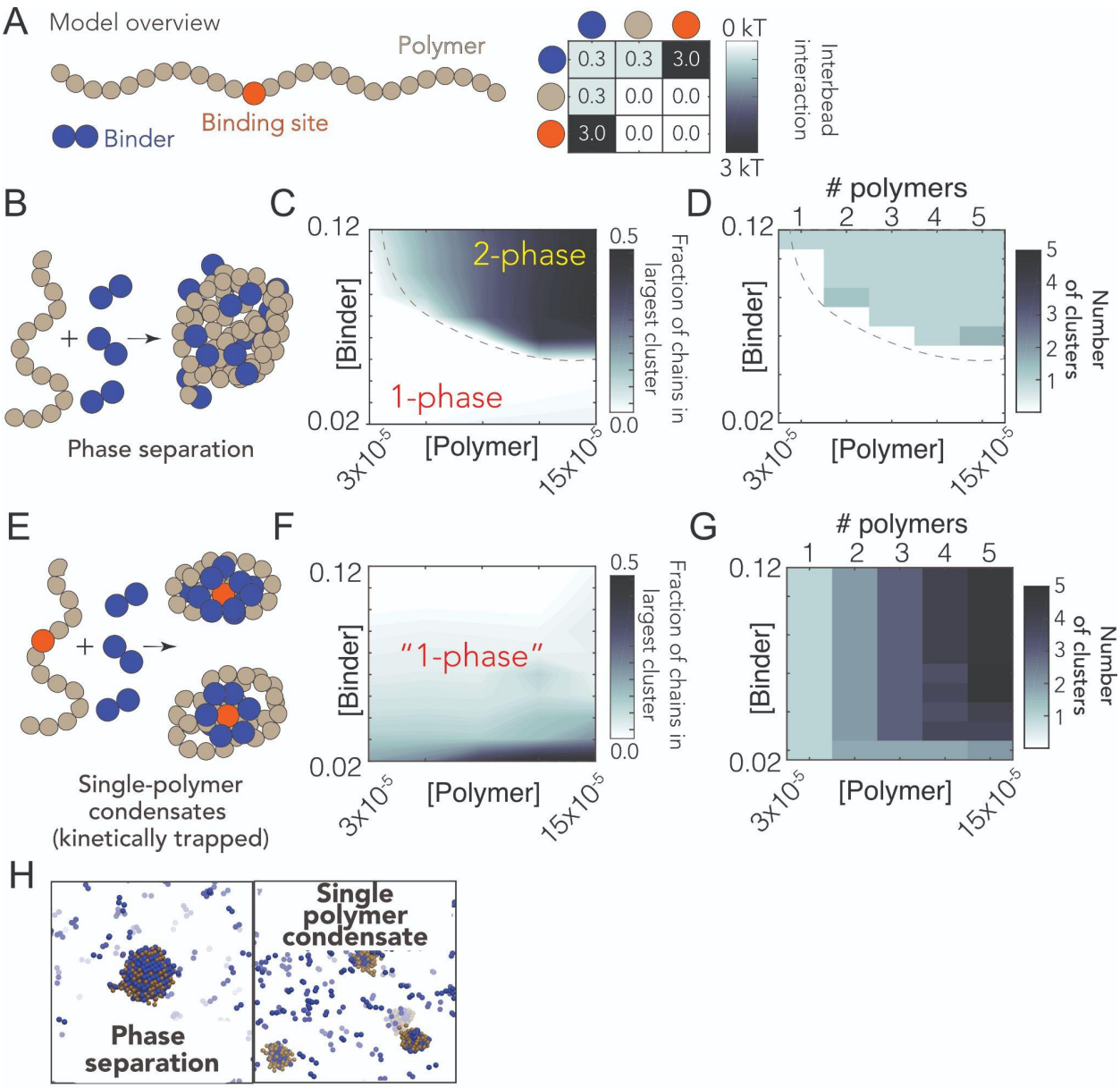


Figure 2. Phase separation with specific and non-specific packaging motif

Adapted from Cubuk, Alston *et al.*[441] **A.** Overview of the coarse grained lattice based simulation model. The model uses a 61-bead homopolymer (brown) and a 2-bead binder species (blue). Beads are multivalent and can interact with all lattice sites based on the pairwise interactions in the adjacent interaction matrix. **B.** Schematic representation of homopolymer and binder interactions,

in absence of a high affinity binding site. **C.** A concentration-dependent solution demixing occurs with enrichment of both homopolymer and binder within the condensate. Dashed line on the binder:polymer concentration plane denotes a qualitative estimate of the binodal. **D.** Number of clusters observed as a function of polymer and binder concentration: a single condensate forms for almost all conditions where condensates are observed. **E.** Schematics of the homopolymer with a single multivalent high affinity binding site (red bead) and binder species. **F.** Condensate formation is suppressed in presence of a high affinity binding site. **G.** The number of clusters observed (as a function of polymer and binder concentration) scales linearly with the number of homopolymers in the simulation **H.** Snapshots of simulations of the homopolymer and binder in absence (left) and presence (right) of the high affinity binding site.

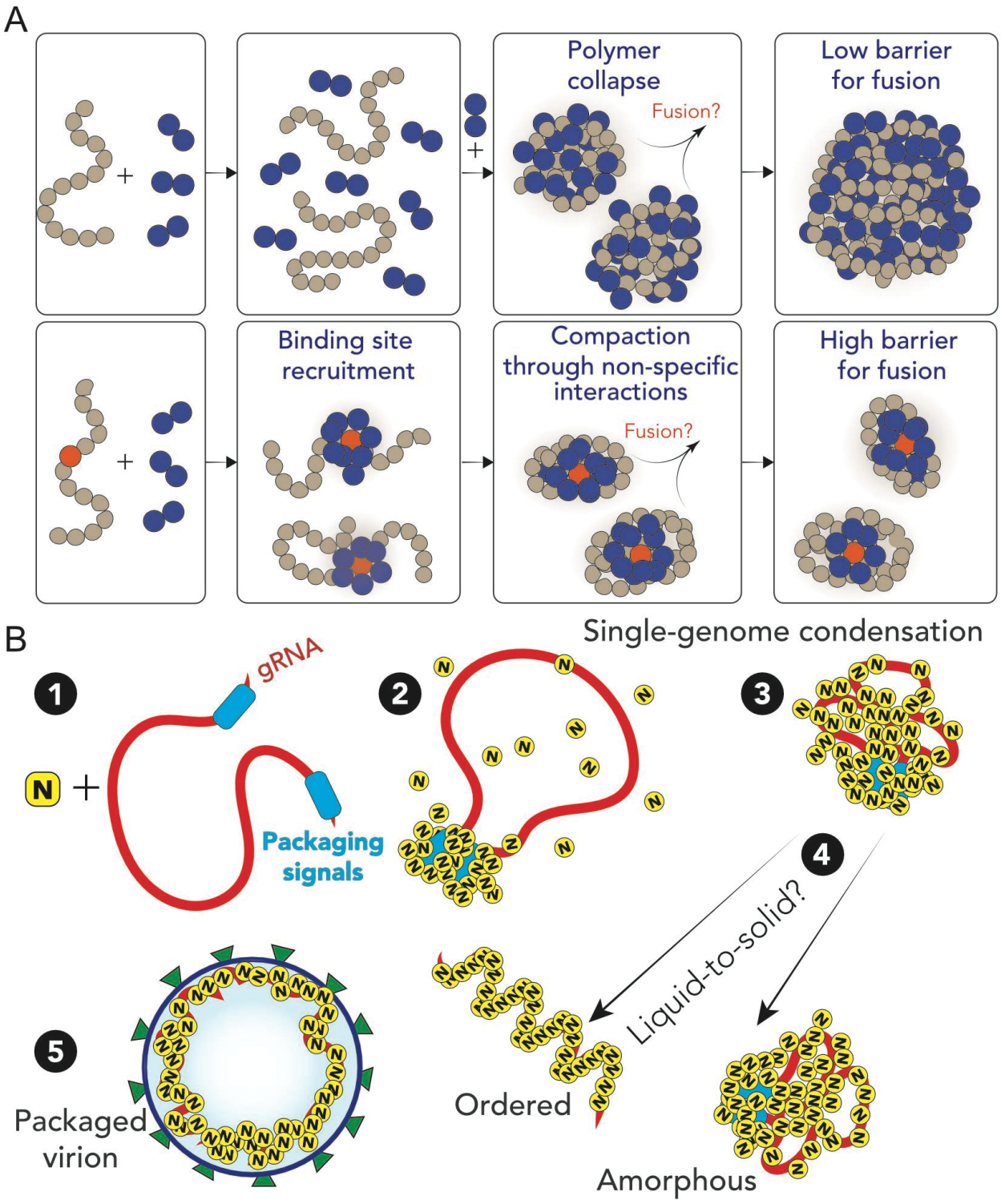

A
Polymer collapse
Fusion?
Low barrier for fusion
Binding site recruitment
Compaction through non-specific interactions
Fusion?
High barrier for fusion
B
Single-genome condensation
gRNA
N
Packaging signals
1
2
3
4
5
Liquid-to-solid?
Ordered
Amorphous
Packaged virion

Figure 3. Competition model of nucleic acid condensation: single chains *vs* phase separation

Adapted from Cubuk, Alston *et al.*[441] **A.** In absence of the high-affinity binding site, homopolymers collapse in presence of binder molecules and, under demixing conditions, can coalesce in larger condensates. Conversely, inclusion of a high affinity binding site leads to single polymer collapse with a high barrier for fusion. **B.** Proposed model for SARS-CoV-2 genome packaging. (1-2) A simple overview of potential SARS-CoV-2 genome packaging mediated by nucleocapsid protein. Packaging signals could mediate clustering of N protein at specific loci, here depicted at the 5' and 3'. (3) Clustering of nucleocapsid protein leads to condensation of the genome, enabling compaction of single genomes. (4) A liquid-to-solid transition of the single nucleocapsid protein-RNA condensate can lead to ordered or amorphous organization of the viral genome. The condensate is then able to be packaged *via* interaction with other SARS-CoV-2 structural proteins.

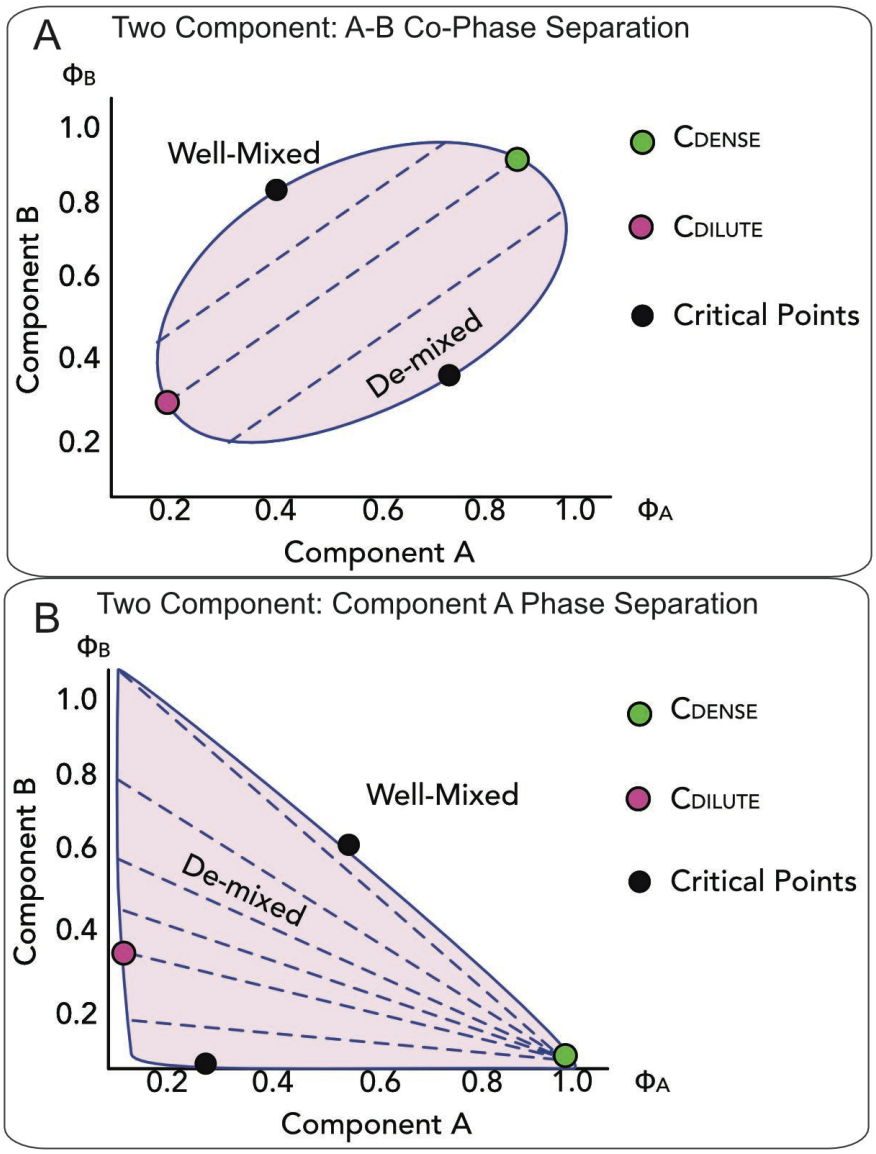


Figure 4. Examples of ternary phase diagrams comprising two polymers and a solvent

Phase diagrams as a function of the volume fraction ϕ_A and ϕ_B of the two polymer components A and B. In both panels, the shaded area indicates the region where the solution demixes, whereas black dots identify critical points. Dashed lines represent tie lines across which the concentrations in the dense (green dot) and dilute phase (magenta dot) remain constant. Note green and magenta dots

are shown for only a single tie line, but the locus of all their points, i.e. all the light and dense phase concentrations adopted when the solution demixes, delimitates the phase boundaries. **A.** In the first case, the two polymer components A and B co-phase separate. The positive slope of the tie lines indicates favorable interactions between the components and the two components are both enriched in the dense phase. **B.** In the second case, component A undergoes phase separation and component B is excluded from the demixed solution.

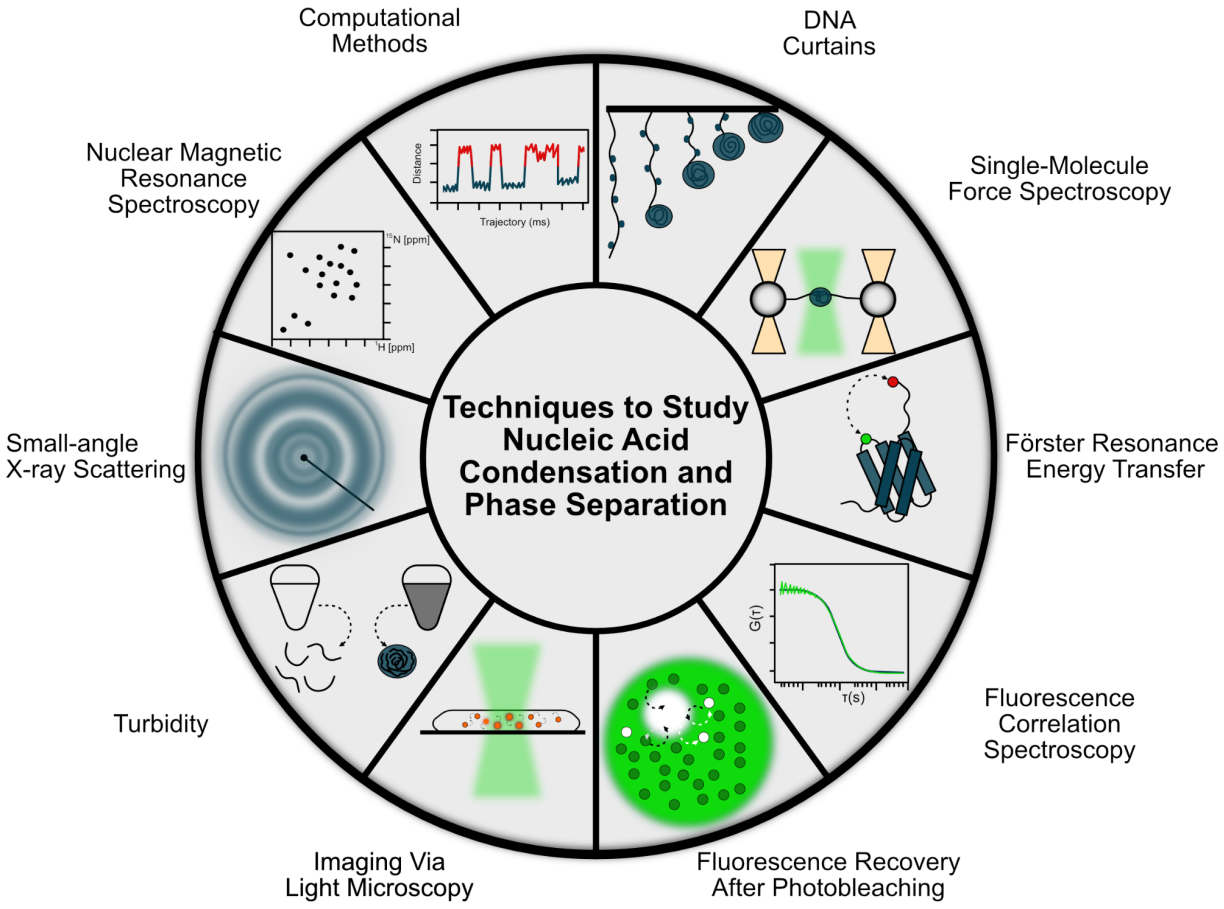


Figure 5. Overview of Techniques to Study Condensation and Phase Separation

Graphical summary of techniques that can be employed to study nucleic acid condensation and phase separation of proteins and protein-nucleic acid mixtures.

Chapter 4: The analytical Flory random coil is a simple-to-use reference model for unfolded and disordered proteins

This chapter was published in the Journal of Physical Chemistry B as:
Alston JJ, Ginell GM, Soranno A, Holehouse AS. The Analytical Flory Random Coil Is a Simple-to-Use Reference Model for Unfolded and Disordered Proteins. J Phys Chem B. 2023 Jun 1;127(21):4746-4760. doi: 10.1021/acs.jpcb.3c01619. Epub 2023 May 18. PMID: 37200094.

Author Contributions: J.J.A., G.M.G, and A.S.H. conceived the manuscript. A.S.H and J.J.A wrote code. G.M.G. collected data from literature. J.J.A., G.M.G, A.S., and A.S.H. wrote the manuscript.

4.1 Abstract

Denatured, unfolded, and intrinsically disordered proteins (collectively referred to here as unfolded proteins) can be described using analytical polymer models. These models capture various polymeric properties and can be fit to simulation results or experimental data. However, the model parameters commonly require users' decisions, making them useful for data interpretation but less clearly applicable as stand-alone reference models. Here we use all-atom simulations of polypeptides in conjunction with polymer scaling theory to parameterize an analytical model of unfolded polypeptides that behave as ideal chains ($\nu = 0.50$). The model, which we call the analytical Flory Random Coil (AFRC), requires only the amino acid sequence as input and provides direct access to probability distributions of global and local conformational order parameters. The model defines a specific reference state to which experimental and computational results can be compared and normalized. As a proof-of-concept, we use the AFRC to identify sequence-specific intramolecular interactions in simulations of disordered proteins. We also use the AFRC to contextualize a curated set of 145 different radii of gyration obtained from previously published small-angle X-ray scattering experiments of disordered proteins. The AFRC is implemented as a stand-alone software package and is also available via a Google colab notebook. In summary, the AFRC provides a simple-to-use reference polymer model that can guide intuition and aid in interpreting experimental or simulation results.

4.2 Introduction

Proteins are finite-sized heteropolymers, and the application of polymer physics has provided a useful toolkit for understanding protein structure and function[36,401,570,572–577]. In particular, there has been significant interest in unfolded proteins under both native and non-native conditions[74,186,400,573,578–582]. Depending on the experimental techniques employed, a variety of polymeric properties can be measured, including the radius of gyration (R_g), the hydrodynamic radius (R_h), the end-to-end distance (R_e), and the apparent scaling exponent (ν_{app}). These and many other parameters can be calculated directly from all-atom simulations, and the synergy of simulation and experiment has provided a powerful approach for constructing large ensembles of unfolded proteins for greater insight into the unfolded state[56,79,80,90,173,186,219,267,331,544,583,584].

Polymers can be described in terms of scaling laws, expressions that describe how chain dimensions vary as a function of chain length[398,585,586]. Polymer scaling laws typically have the format $D = R_0N^{\nu}$. Here, D reports on chain dimensions, R_0 is a prefactor in units of spatial distance, and N is the number of monomers, which in the case of proteins is typically written in terms of the number of amino acids. ν (or, more accurately, ν_{app} when applied to finite-sized heteropolymers like proteins) is the (apparent) Flory scaling exponent. In principle, ν_{app} lies between 0.33 (as is obtained for a perfect spherical globule) and 0.59 (as obtained for a self-avoiding chain). However, for finite-sized polymers, values beyond 0.59 can be obtainable for self-repulsive chains [35,231,272]. The applicability of polymer scaling laws to describe real proteins assumes they are sufficiently long to display *bona fide* polymeric behavior and that they are sufficiently self-similar over a certain length scale, analogous to fractals. While this assumption often holds true, it is worth noting that sequence-encoded patterns in

specific chemistries and/or secondary structure can lead to deviations from homopolymer-like behavior [90,317,318,587].

To what extent do polymer scaling laws apply to real proteins? For denatured polypeptides, Kohn *et al.* reported the ensemble-average radius of gyration using the scaling expressions $R_g = 1.927N^{0.598400}$. This result provides strong experimental evidence to support a model whereby denaturants unfolded proteins by uniformly weakening intramolecular protein-protein interactions[572]. A value for ν_{app} of 0.598 also agrees with the previously reported value of 0.57 by Wilkins *et al.* and earlier work by Damaschun[572,578,579]. In short, under strongly denaturing conditions, proteins appear to behave as polymers in a good solvent [35,271,572,588–590].

For proteins under native or native-like conditions, the apparent scaling exponents obtained for unfolded polypeptides are more variable. Marsh and Forman-Kay reported an average scaling expression of $R_h = 2.49N^{0.509}$, for a set of intrinsically disordered proteins, while Bernadó and Svergun found a similar average relationship in $R_g = 2.54N^{0.52}$ [45,591]. More recently, various means to estimate ν_{app} for individual proteins have enabled values of ν_{app} between 0.42 and 0.60 to be measured for a wide range of unfolded proteins of different lengths and compositions[80,90,173,186,233,271,308,589,592]. An emerging consensus suggests that ν_{app} depends on the underlying amino acid sequence[74,573,593]. If sequence-encoded chemical biases enable intramolecular interactions, then ν_{app} may be lower than 0.5. Notably, despite clear conceptual limitations, the physics of homopolymers remains a convenient tool through which unfolded proteins can be assessed[37,90,186,318,587].

Given the variety in scaling exponents for unfolded proteins under native conditions, we felt that a sequence-specific reference model would be helpful for the field. Such a model could provide a

touchstone for experimentally measurable polymeric parameters, including intermolecular distances, the radius of gyration, the end-to-end distance, and the hydrodynamic radius. Similarly, such a model would provide a simple reference state with which simulations could be directly compared and used to identify sequence-specific effects. Finally, a standard reference model could offer an easy way to compare unfolded proteins of different lengths to assess if they behave similarly despite different absolute dimensions.

Here, we perform sequence-specific numerical simulations for polypeptides as an ideal chain, so-called Flory Random Coil (FRC) simulations[573,586]. Under these conditions, chain-chain, chain-solvent, and solvent-solvent interactions are all equivalent, no long-range excluded volume contributions are included, and as such, the polypeptide behaves as a Gaussian chain with $\nu_{app} = 0.5$. Because our FRC implementation minimizes finite-chain artifacts, we can parameterize an analytical, sequence-specific model using standard approaches from scaling theory, a model we call the Analytical Flory Random Coil (AFRC). This model enables the calculation of distance distributions for the end-to-end distance and the radius of gyration, as well as a variety of additional parameters that become convenient for the analysis of all-atom simulations and experiments.

The AFRC is not a predictor of unfolded protein dimensions. Those dimensions depend on the complex interplay of chain:chain and chain:solvent interactions, which are themselves determined by sequence-encoded chemistry[178,180,424,594,595]. Instead, the AFRC provides a simple reference state that can aid in interpreting experimental and computational results without needing information other than the protein sequence. The AFRC is implemented in a stand-alone Python package and is also

provided as a simple Google Colab notebook. We demonstrate the utility of this model by comparing experimental data and computational results.

The remainder of this paper is outlined as follows. First, we discuss the implementation details of the model, including a comparison against existing polymer models. Next, we analyzed previously published all-atom simulations to demonstrate how the AFRC can identify signatures of sequence-specific intramolecular interactions in disordered ensembles. Finally, we use the AFRC model to re-interpret previously reported small-angle X-ray scattering data of intrinsically disordered proteins.

METHODS AND RESULTS

4.3 Implementation of a numerical model for sequence-specific ideal chain simulations

We used a Monte Carlo-based approach to construct sequence-specific atomistic ensembles of polypeptides as ideal chains. All-atom simulations with all non-bonded and solvation interactions scaled to zero were performed using a modified version of the CAMPARI Monte Carlo simulation engine using bond lengths and atomic radii defined by the ABSINTH-OPLS forcefield[104,132,573]. We modified CAMPARI to reproduce Flory's rotational isomeric state approximation[586,596]. In this method, an initial conformation of the polypeptide is randomly generated. Upon each Monte Carlo step, a residue is randomly selected, the backbone dihedrals are rearranged to one of a subset of allowed residue-specific psi/phi values (i.e., specific isomeric states), and the chain is rearranged accordingly (**Fig. 1A, B)**. Allowed phi/psi values are selected from a database of residue-specific allowed values as determined by all-atom simulations of peptide units, with the associated Ramachandran maps shown in (**Fig. S1**) Importantly, the Monte Carlo moves in these simulations approach are rejection-free. That is, only allowed phi/psi angles are proposed, and no consideration

of steric overlap in the resulting conformation is given. The ensemble generated by these simulations is referred to as the Flory Random Coil (FRC, **Fig. 1C**) and has been used as a convenient reference frame for comparing simulations of disordered and unfolded polypeptides for over a decade (as reviewed by Mao *et al.*[573])[186,193,306,319,597,598].

FRC simulations enable the construction of ensembles where each amino acid exists in a locally allowed configuration, yet no through-space interactions occur. This has two important implications for the construction of an ideal chain model. Firstly, each monomer has no preference for chain:chain vs. chain:solvent interactions (each monomer is "agnostic" to its surroundings). As a result, both internal and global dimensions show scaling behavior with an apparent scaling exponent (ν_{app}) of 0.5 (**Fig. 1D**), analogous to that of a polymer in a theta solvent. Secondly, terminal residues sample conformational space in the same way as residues internal to the chain (**Fig. S2**). This means that end-effects that emerge finite-chain effects are not experienced in terms of end effects (**Fig. 1E**). This is in contrast to finite-sized self-avoiding chains, in which internal scaling profiles reveal a noticeable and predictable "dangling end" finite-chain effect (**Fig 1E, Fig. S2**). In summary, FRC simulations enable us to generate ensembles at all-atom resolution that are nearly fully approximations of ideal chains, reproducing the behavior of a hypothetical "ideal" polypeptide.

4.4 Constructing an analytical description of the Flory Random Coil

Our FRC ensembles enable the calculation of a range of polymeric properties, including inter-residue distances, inter-residue contact probabilities, the hydrodynamic radius, or the radius of gyration. Comparing these properties with experiments or simulations is often convenient, offering a standard reference frame for normalization and biophysical context[74,186,318,573,587]. However,

performing and analyzing all-atom simulations with CAMPARI necessitates a level of computational sophistication that may make these calculations inaccessible to many scientists. To address this, we next sought to develop a set of closed-form analytical expressions to reproduce these properties and implement them as an easy-to-use package available both locally and – importantly – via a simple web interface (Google colab notebook).

FRC simulations generate ensembles that – by definition – reproduce the statistics expected for an ideal chain. As mentioned, polymer scaling behavior generally takes the form;

$$D = R_0 N^{\nu} \quad \text{(Eq. 1)}$$

For an ideal chain, ν_{app} should not depend on the amino acid sequence (as all chains should scale with $\nu_{app} = 0.5$). However, the prefactor R_0 can and will show sequence dependence. As such, computing polymeric properties from sequences necessitates a means to calculate sequence-specific prefactor values. Prefactor values were parameterized using homopolymer simulations of each amino acid (see supplementary information). The inter-residue distance prefactor A_0 was parameterized by fitting internal scaling profiles using equation (2);

$$\sqrt{\langle\langle r^2_{(i,j)}\rangle\rangle} = A_0|i-j|^{\nu} \quad \text{(Eq. 2)}$$

In equation 2, |i-j| is the number of residues between residues at position i and j, the left-hand-side reports on the root-mean-square (RMS) distance between residues i and j in the chain, ν is the scaling exponent (in our case this is equal to 0.5), and A_0 is a prefactor for which we can directly solve for. The double angle brackets around the RMS distance reflect the fact we are averaging over all pairs of residues that are |i-j| apart and doing so for all chain configurations. Plotting |i-j| vs. the RMSD generates the internal scaling profile shown in **Fig. 1E**. By fitting homopolymers of the

20 amino acids, a set of residue-specific A_0 prefactors was determined, as listed in **Supplementary Table 1**.

For our homopolymers, we can calculate the root-mean-squared end-to-end distances using equation (3);

$$\sqrt{\langle r_e^2 \rangle} = A_0 N^{\nu} \quad \text{(Eq. 3)}$$

From this, we can then use the standard function for P(r) of a Gaussian chain to calculate the end-to-end distance distribution;

$$P(r) = 4\pi r^2 \left(\frac{3}{2\pi\langle r_e^2 \rangle}\right)^{3/2} e^{-\left(\frac{3r^2}{2\langle r_e^2 \rangle}\right)} \quad \text{(Eq. 4)}$$

After determining residue-specific A_0, a comparison of analytical and numerical simulation distributions show excellent agreement when homopolymer end-to-end distance distributions are compared between FRC simulations and the AFRC-derived values (**Fig. 1f**).

We next took a similar route to define the radius of gyration (R_g) distribution. While no closed-form solution for the distribution of the radius of gyration exists, Lhuillier previously defined a closed-form approximation for this distribution for a fractal chain[599];

$$P_{Rg}(x) \sim N^{-\nu d} f(x) \left(\frac{x}{N^{\nu}}\right) \quad \text{(Eq. 5)}$$

Where;

$$f(x) \sim exp\left[-\left(\frac{N^{\nu}}{x}\right)^{\alpha d} - \left(\frac{x}{N^{\nu}}\right)^{\delta}\right] \quad \text{(Eq. 6)}$$

And the variables α and δ are defined as:

$$\alpha = \frac{1}{(\nu d - 1)} \quad \text{(Eq. 7)}$$

$$\delta = \frac{1}{(1 - \nu)} \quad \text{(Eq. 8)}$$

Here, x represents the distance in some arbitrary units (written as such to avoid confusion with r, which represents the distance in Angstroms [Å]), N and ν again represent the total number of residues and the scaling exponent (0.5.), while d is the dimensionality (d=3). This allows us to calculate α and δ exactly, given ν is fixed at 0.5. Consequently, we can recast equation 5 into units of Å using a sequence-specific normalization factor (X_0);

$$r = X_0 x \quad \text{(Eq. 8)}$$

To calculate X_0, we fit numerically-generated $P(R_g)$ distributions from homopolymer simulations with a series of analytically generated distributions to identify the best-fitting amino acid-specific X_0 values. These prefactors are listed in **Supplementary Table 1**. As with the end-to-end distances, a comparison of numerically-generated $P(R_g)$ with analytically-generated $P(R_g)$ values are in extremely good agreement (**Fig. 1g**). Comparing ensemble average end-to-end distance and radii of gyration for homopolymers of all 20 amino acids in lengths from 50 to 350 amino acids revealed a Pearson correlation coefficient of 0.999 and a root mean square error (RMSE) of 0.8 Å and 0.3 Å for the end-to-end distance and radius of gyration, respectively (**Fig. S2**).

With analytical expressions for computing the end-to-end distance and radius of gyration probability distributions in hand, we can calculate additional polymeric properties. Given the fractal nature of the Flory Random Coil and the absence of end effects, we can calculate all possible inter-residue distances and, correspondingly, contact frequencies between pairs of residues (**Fig. 2a, b**). Similarly, using either the Kirkwood-Riseman equation or a recently derived empirical relationship, we can

compute an approximation for the ensemble-average hydrodynamic radius[274,600,601]. In summary, the AFRC offers an analytic approach for calculating sequence-specific ensemble properties for unfolded homopolymers.

4.5 Generalization to heteropolymers

Our parameterization has thus far focused exclusively on homopolymer sequences. However, Flory's rotational isomeric state approach requires complete independence of each amino residue[586,596]. Consequently, we expected the prefactor associated with a given heteropolymer to reflect a weighted average of prefactors taken from homopolymers, where the sequence composition determines the weights.

To test this expectation, we compared numerical simulations with AFRC predictions for a set of different polypeptide sequences finding excellent agreement in both end-to-end distances and radii of gyration (**Fig. 3a, b** and **Fig. S3**). Similarly, given the absence of end-effects, our analytical end-to-end distance expression works equally well for intramolecular distances in addition to the end-to-end distance. To assess this, we compared internal scaling profiles between FRC simulations and AFRC predictions (**Fig. 3c**). These profiles compare the ensemble average distance between each possible inter-residue distance and offer a convenient means to assess both short and long-range intramolecular distances. We performed FRC simulations for 320 different polypeptide sequences ranging in length from 10 to 500 amino acids with a systematic variation in amino acid composition. Across all internal scaling profile comparisons between FRC and AFRC simulations, the overall average RMSE was 0.5 Å, with almost all (92%) of individual comparisons revealing an RMSE under 1 Å (**Fig. 3D**). Similarly, the Pearson's correlation coefficient between internal scaling profiles for

FRC vs. AFRC for all ten-residue chains was 0.9993, which was the worst correlation across all lengths (**Fig. S4**). In summary, the AFRC faithfully reproduces homo- and hetero-polymeric dimensions for polypeptides under the FRC assumptions.

4.6 Comparison with existing polymer models

For completeness, we compared the end-to-end distance distributions obtained from several other polymer models used throughout the literature for describing unfolded and disordered polypeptides. Previously-used polymer models offer a means to analytically fit experimental or computational results and benefit from taking one (or more) parameters that define the model's behavior. While the AFRC does not enable fitting to experimental or simulated data, it only requires an amino acid sequence as input. With this in mind, the AFRC serves a fundamentally different purpose than commonly used models.

We wondered if dimensions obtained from the AFRC would be comparable with dimensions obtained from other polymer models when using parameters used previously in the literature. We compared distributions obtained from the worm-like chain (WLC), the self-avoiding walk (SAW) model, and a recently-developed ν-dependent self-avoiding walk (SAW-ν)[120,173]. For the WLC model, we used a persistence length of 3.0 Å and an amino acid size of 3.8 Å (such that the contour length, l_c, is defined as $N \times 3.8$[120]). For the SAW model, we used a scaling prefactor of 5.5 Å (i.e., assuming $\langle R_e \rangle = 5.5N^{0.598}$)[35,120,173]. Finally, for SAW-ν, we computed distributions using a prefactor of 5.5 Å and using several different ν values[173,401]. These values were chosen because previous studies have used them to describe intrinsically disordered proteins.

Fig. 4A shows comparisons of the AFRC distance distribution obtained for a 100-mer polyalanine (A_{100}) vs. the WLC and SAW (top) and vs. ν-dependent distributions (bottom). The AFRC is slightly more expanded than the WLC model using the parameters provided, although the persistence length can, of course, be varied to explore more compact (lower l_p) or more extended (higher l_p) distributions (**Fig. S6A**). The AFRC is substantially more compact than the SAW model. The comparison with the SAW model is important, as with a prefactor of 5.5 Å the SAW model describes a polypeptide as a self-avoiding random coil (ν=0.588), whereas the AFRC describes a polypeptide as an ideal chain (ν = 0.5), such that we should expect the SAW to be more expanded than the AFRC. Finally, in comparing the AFRC with the SAW-ν model, we find that the AFRC distribution falls almost completely top of the ν = 0.50 distribution. This indicates that both models arrive at nearly identical distance distributions despite being developed independently. This result is both confirmatory and convenient, as it means the AFRC and SAW-ν models can be used to analyze the same data without concern for model incompatibility.

We emphasize that this comparison with the existing polymer model is not presented to imply the AFRC is better than existing models but to highlight their compatibility. One can tune input parameters for all three models to arrive at qualitatively matching end-to-end distributions (**Fig. S6B**). The major difference between these three models and the AFRC is simply that the AFRC requires only amino acid sequence as input, making it a convenient reference point. For completeness, all four models are implemented in our Google colab notebook.

We also compared ensemble-average radii of gyration obtained from the various models with those obtained from the AFRC. While the WLC, SAW, and SAW-ν models do not provide approximate

closed-form solutions for the radius of gyration distribution, they do enable an estimate of the ensemble-average radius of gyration to be calculated[120,173]. Using the same model parameters as was used in **Fig. 4A**, the AFRC falls between the SAW and the WLC. Moreover, the AFRC radii of gyration scale almost 1:1 with the SAW-ν derived radii as a function of chain length when $\nu = 0.50$. As such, we conclude that the AFRC is consistent with existing polymer models yet benefits from being both parameter-free (for the user) and offering full distributions for the radius of gyration and intramolecular distance distributions per-residue contact fractions, convenient properties for normalization in simulations and experiment.

4.7 Comparison with all-atom simulations

Our work thus far has focussed on developing and testing the robustness of the AFRC. Having done this, we next sought to ask how similar (or dissimilar) distributions obtained from the AFRC are compared to all-atom simulations. We used simulations generated via all-atom molecular dynamics with the Amber99-disp forcefield and all-atom Monte Carlo simulations with the ABSINTH-OPLS forcefield[80,91,93,99,104,602,603]. Specifically, we examined nine different fully disordered proteins: The unfolded Drosophila Drk N-terminal SH2 domain (DrkN, 59 residues)[99,604,605], the ACTR domain of p160 (ACTR, 71 residues)[99,271,589,606], a C-terminal disordered subregion of the yeast transcription factor Ash1 (Ash1, 83 residues)[93], the N-terminal disordered regions of p53 (p53, 91 residues)[91,607], the C-terminal IDR of p27 (p26, 107 residues)[603], the intrinsically disordered intracellular domain of the notch receptor (Notch, 132 residues)[602], the C-terminal disordered domain of the measles virus nucleoprotein (Ntail, 132 residues)[99,608], the C-terminal low-complexity domain of hnRNPA1 (A1-LCD, 137 residues)[80], and full-length alpha-synuclein (asyn, 140 residues)[99,609,610].

We compared distributions for the end-to-end distance and radius of gyration for our all-atom simulations with analogous distributions generated by the AFRC (**Fig. 5**). These comparisons revealed that while the general shape of the distributions recovered from simulations was not dissimilar from the AFRC-derived end-to-end distance and radius of gyration distributions, the width and mean were often different. This is hardly surprising, given that the global dimensions of an unfolded protein depend on the underlying amino acid sequence. The ratio of the mean end-to-end distance divided by the AFRC-derived mean end-to-end distance (or the corresponding ratio for the radius of gyration) was found to range between 0.7 and 1.4. In some cases, the end-to-end distance ratio or radius of gyration ratio varied within the same protein. For example, for the 132-residue intracellular-domain IDR from Notch (Notch), the end-to-end distance ratio was 0.8 (i.e., smaller than predicted by the AFRC), while the radius of gyration ratio was 1.0. Similarly, in alpha-synuclein (Asyn), the corresponding ratios were 0.7 and 0.9, again reporting a smaller end-to-end distance than radius of gyration. As suggested previously, discrepancies in end-to-end distance vs. radius of gyration vs. expectations from homopolymer models are diagnostic of sequence-encoded conformational biases[90,317,587,611].

We also used the AFRC to calculate scaling maps. Scaling maps are non-redundant matrices of inter-residue distances obtained from simulations and normalized by the expected inter-residue distances obtained by the AFRC (**Fig. 6**)[93]. We compared these scaling maps (top left triangle of each panel) against absolute distances (bottom right triangle). This comparison highlights the advantage that using a reference polymer model offers. Long-range sequence-specific conformational biases are much more readily visualized as deviations from an expected polymer model. Moreover, the same

dynamic range of values can be used for chains of different lengths, normalizing the units from Å to a unitless ratio.

Returning to the notch simulations, both types of intramolecular distance analysis clearly illustrate a strong long-range interaction between the N-terminal residues 1-30 and the remainder of the sequence. The long-range interaction between chain ends influences the end-to-end distance much more substantially than it does the radius of gyration (**Fig. 6**). Similarly, in alpha-synuclein, we observed long-range interactions between the negatively charged C-terminus and the positively-charged residues 20-50, leading to a reduction in the end-to-end distance. In short, the AFRC provides a convenient approach to enable direct interrogation of sequence-to-ensemble relationships in all-atom simulations.

Finally, we calculated per-residue contact scores for each residue in our nine proteins (**Fig. 7**). These contact scores sum the length-normalized fraction of the simulation in which each residue is in contact with any other residue in the sequence[80]. While this collapses information on residue-specificity into a single number, it integrates information from the typically-sparse contact maps for IDR ensembles to identify residues that may have an outside contribution towards short (<6 Å) range molecular interactions. We and others have previously used this approach to identify "stickers" - regions or residues in IDRs that have an outsized contribution to intra- and inter-molecular interactions[80,598,612,613].

In some proteins, specific residues or subregions were identified as contact hotspots. This includes the aliphatic residues in ACTR, and hydrophobic residues in the p53 transactivation domains, in line

with recent work identifying aliphatic residues as driving intramolecular interactions[598,614]. Most visually noticeable, aromatic residues in the A1-LCD appear as spikes that uniformly punctuate the sequence, highlighting their previously-identified role as evenly-spaced stickers[80]. Intriguingly, in alpha-synuclein, several regions in the aggregation-prone non-amyloid core (NAC) region (residues 61-95) appear as contact score spikes, potentially highlighting the ability of intramolecular interactions to guide regions or residues that may mediate inter-molecular interaction.

4.8 Comparison with SAXS-derived radii of gyration

Having compared AFRC-derived parameters with all-atom simulations, we next sought to determine if AFRC-derived polymeric properties compared reasonably with experimentally-measured values. As a reminder, the AFRC is not a predictor of IDR behavior; instead, it offers a null model against which IDR dimensions can be compared. To perform a comparison with experimentally derived data, we curated a dataset of 145 examples of radii of gyration measured by small-angle X-ray scattering (SAXS) of disordered proteins. We choose to use SAXS data because SAXS-derived radii of gyration offer a label-free, model-free means to determine the overall dimensions of a disordered protein. That said, SAXS-derived measurements are not without their caveats (see discussion), and where possible, we re-analyzed primary scattering data to ensure all radii of gyration reported here are faithful and accurate.

To assess our SAXS-derived radii of gyration, we calculated expected dimensions for denatured proteins, folded globular domains, or AFRC chains by fitting scaling laws with the form $R_g = R_0N^{\nu}$ against different polymer models. We used a denatured-state polymer model ($\nu = 0.59$, $R_0 = 1.98$, as defined by Kohn *et al.*) and a folded globular domain model ($\nu = 0.33$, $R_0 = 2.86$, as obtained from

PDBSELECT25 originally plotted by Holehouse & Pappu)[37,400,615]. We also calculated the AFRC-derived radii of gyration for all 145 chains and fitted a polymer scaling model to the resulting data where the only free parameter was R_0 ($\nu = 0.50$, $R_0 = 2.50$). This analysis showed that the majority of the 145 proteins have a radius of gyration above that of the AFRC-derived radius of gyration (see discussion), with some even exceeding the expected radius of gyration of a denatured protein (**Fig. 8A**). Based on these data, we determined an empirical upper and lower bound for the biologically accessible radii of gyration given a chain length (see discussion). This threshold suggests that, for a sequence of a given length, there is a wide range of possible IDR dimensions accessible (**Fig. 8B**, **Fig. S5**).

Finally, we wondered how well the AFRC-derived radii of gyration would correlate with experimentally-measured values. Based on the upper and lower bounds shown in **Fig. 8B**, we excluded four radii of gyration that appear to be spuriously large, leaving 141 data points. For these 141 points, we calculated the Pearson correlation coefficient (r) and the RMSE between the experimentally-measured radii of gyration and the AFRC-derived radii of gyration. This analysis yielded a correlation coefficient of 0.91 and an RMSE of 6.4 Å (**Fig. 8C**). To our surprise, these metrics outperform several established coarse-grained models for assessing intrinsically disordered proteins, as reported recently[405]. We again emphasize that the AFRC is not a predictor of IDR dimensions. However, we tentatively suggest that this result demonstrates that a reasonably good correlation between amino acid sequence and global dimensions can be obtained solely by recognizing that disordered proteins are flexible polymers. With this in mind, we conclude that the AFRC provides a convenient and easily-accessible control for experimentalists measuring the global dimensions of disordered proteins.

4.9 Reference implementation and distribution

Computational and theoretical tools are only as useful as they are usable. To facilitate the adoption of the AFRC as a convenient reference ensemble, we provide the AFRC as a stand-alone Python package distributed through PyPI (**pip install afrc**). We also implemented the additional polymer modes described in **Fig. 4** with a consistent programmatic interface, making it relatively straightforward to apply these models to analyze and interpret computational and experimental data. Finally, to further facilitate access, we provide an easy-to-use Google colab notebook for calculating expected parameters for easy comparison with experiments and simulations. All information surrounding access to the AFRC model is provided at https://github.com/idptools/afrc.

4.10 Discussion & Conclusion

In this work, we have developed and presented the Analytical Flory Random Coil (AFRC) as a simple-to-use reference model for comparing against simulations and experiments of unfolded and disordered proteins. We demonstrated that the AFRC behaves as a truly ideal chain and faithfully reproduces homo- and hetero-polymeric inter-residue and radius of gyration distributions obtained from explicit numerical simulations. We also compared the AFRC against several previously-established analytical polymer models, showing that ensemble-average or distribution data obtained from the AFRC are interoperable with existing models. Finally, we illustrated how the AFRC could be used as a null model for comparing data obtained from simulations and from experiments.

The AFRC differs from established polymer models in two key ways. While existing models define functional forms for polymeric properties, they do not prescribe specific length scales or parameters

for those models. This is not a weakness - it simply reflects how analytical models work. However, the need to provide 'appropriate' parameters to ensure these models recapitulate behaviors expected for polypeptides places the burden on selecting and/or justifying those parameters on the user. The AFRC combines several existing analytical models (the Gaussian chain and the Lhuillier approximation for the radius of gyration distribution) with specific parameters obtained from numerical simulations to provide a "parameter-free" polymer model defined by its reference implementation (as opposed to the mathematical form of the underlying distributions). We place parameter free in quotation marks because the freedom from parameters is at the user level - the model itself is explicitly parameterized to reproduce polypeptides dimensions. However, from the user's perspective, no information is needed other than the amino acid sequence.

Although the AFRC was explicitly parameterized to recapitulate numerical FRC simulations, sequence-specific effects do not generally have a major impact on the resulting dimensions. For example, **Fig. S6** illustrates the radius of gyration or end-to-end distance obtained for varying lengths of poly-alanine and poly-glycine. This behavior is not a weakness of the model - it *is* the model. This relatively modest sequence dependence reflects the fact that for an ideal chain, both the second and third virial coefficients are set to zero (*i.e.*, the integral of Mayer f-function should equal 0)[616]. As such, the AFRC does not enable explicitly excluded volume contributions to the chain's dimensions from sidechain volume, although this is captured implicitly based on the allowed isomeric states (compare glycine to alanine in **Fig. S1**). In summary, the AFRC does not offer any new physics, but it does encapsulate previously derived physical models along with numerically-derived sequence-specific parameters to make it easy to construct null models explicitly for comparison with polypeptides.

In comparing AFRC-derived polymeric properties with those obtained from all-atom simulations, we recapitulate sequence-to-ensemble features identified previously [80,99,584,602]. When comparing the normalized radii of gyration (R_g^{Sim}/ R_g^{AFRC}), we noticed the lower and upper bounds obtained here appear to be approximately 0.8 and 1.4, respectively. To assess if this trend held true for experimentally derived radii of gyration, we calculated the normalized radii of gyration for the 141 values reported in **Fig. 8C**, recapitulating a similar range (0.8 to 1.46). Based on these values, we defined an empirical boundary condition for the anticipated range in which we would expect to see a disordered chain's radius of gyration as between $0.8R_g^{AFRC}$ and $1.45R_g^{AFRC}$ (**Fig. 8B**). We emphasize this is not a hard threshold. However, it offers a convenient rule-of-thumb, such that measured radii of gyration can be compared against this value to assess if a potentially spurious radius of gyration has been obtained (either from simulations or experiments). Such a spurious value does not necessarily imply a problem, but may warrant further investigation to explain its physical origins.

Our comparison with experimental data focussed on radii of gyration obtained from SAXS experiments. We chose this route given the wealth of data available and the label-free and model-free nature in which SAXS data are collected and analyzed. Given the AFRC offers the expected dimensions for a polypeptide behaving qualitatively as if it is in a theta solvent, it may be tempting to conclude from these data that the vast majority of disordered proteins are found in a good solvent environment (**Fig. 9A**). The solvent environment reflects the mean-field interaction between a protein and its environment. In the good solvent regime, protein:solvent interactions are favored, while in the poor solvent regime protein:protein interactions are favored [37,308,401,573]. However, it is worth bearing in mind that SAXS experiments generally require relatively high concentrations of

protein to obtain reasonable signal-to-noise[591]. Recent advances in size exclusion chromatography (SEC) coupled SAXS have enabled the collection of scattering data for otherwise aggregation-prone proteins with great success[617]. However, there is still a major acquisition bias in the technical need of these experiments to work with high concentrations of soluble proteins when integrated over all existing measured data. By definition, such highly soluble proteins experience a good solvent environment. Given this acquisition bias, we remain agnostic as to whether these results can be used to extrapolate to the solution behavior of all IDRs.

Prior work has implicated the presence of charged and proline residues as mediating IDR chain expansion [45,93,231,272,319,592,594,618–621]. We took advantage of the fact that the AFRC enables a length normalization of experimental radii of gyration and assessed the normalized radius of gyration vs. the fraction of charged and proline residue (**Fig. 9B**). Our data support this conclusion as a first approximation, but also clearly demonstrate that while this trend is true on average, there is variance in this relationship. Notably, for IDRs with a fraction of charged and proline residues between 0.2 and 0.4, the full range of possible IDR dimensions are accessible. The transition from (on average) more compact to (on average) more expanded chains occurs around a fraction of proline and charged residues of around 0.25 – 0.30, in qualitative agreement with prior work exploring the fraction of charge residues required to drive chain expansion [45,231,272]. However, we emphasize that there is massive variability observed on a per-sequence basis. In summary, while the presence of charged and proline residues clearly influences IDR dimensions, complex patterns of intramolecular interactions can further tune this behavior [74,573,584].

In summary, the AFRC offers a convenient, analytical approach to obtain a well-defined reference state for comparing and contrasting simulations and experiments of unfolded and disordered proteins. It can be easily integrated into complex analysis pipelines, or used for one-off analysis via a Google Colab notebook without requiring any computational expertise at all.

4.11 Acknowledgements

We thank members of the Pappu lab and Holehouse lab for many useful discussions over the years. We are indebted to Dr. Nick Lyle for the original implementation of the CAMPARI-based FRC engine. We thank Dr. Erik Martin for bringing the work of Lhuillier to our attention. Funding for this work was provided by the National Institute on Allergic and Infectious Diseases with R01AI163142 to A.S.H. and A.S., by the National Science Foundation with 2128068 to A.S.H., by the Longer Life Foundation, an RGA/Washington University in St. Louis Collaboration to A.S.H., and by the National Cancer Institute with F99CA264413 to J.J.A. We also thank members of the Water and Life Interface Institute (WALII), supported by NSF DBI grant #2213983, for helpful discussions.

4.12 Figures

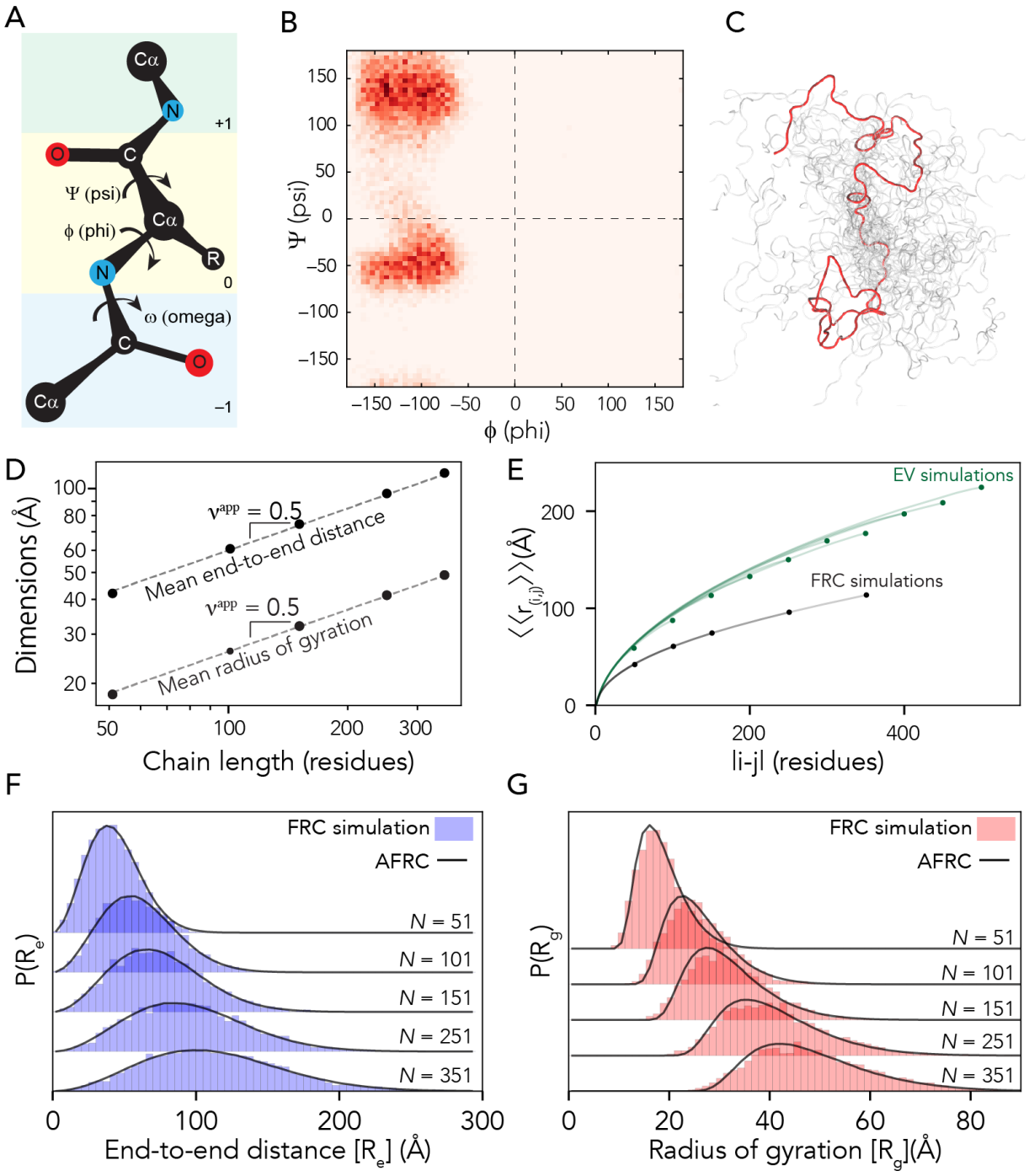

A
Cα
N
+1
O
C
Ψ (psi)
ϕ (phi)
Cα
R
N
0
ω (omega)
C
O
Cα
−1
B
Ψ (psi)
ϕ (phi)
C
D
Dimensions (Å)
Chain length (residues)
ν^{app} = 0.5
Mean end-to-end distance
ν^{app} = 0.5
Mean radius of gyration
E
EV simulations
FRC simulations
$\langle\langle r_{(i,j)} \rangle\rangle$ (Å)
li-jl (residues)
F
FRC simulation
AFRC
N = 51
N = 101
N = 151
N = 251
N = 351
P(R_e)
End-to-end distance [R_e] (Å)
G
FRC simulation
AFRC
N = 51
N = 101
N = 151
N = 251
N = 351
P(R_g)
Radius of gyration [R_g](Å)

Figure 1. The AFRC is a pre-parameterized polymer model based on residue-specific polypeptide behavior.

A. Schematic of the amino acid dihedral angles. **B.** Ramachandran map for alanine used to select acceptable backbone conformations for the FRC simulations. All twenty amino acids are shown in **Fig. S1**. **C.** Graphical rendering of an FRC ensemble for a 100-residue homopolymer. **D.** Flory Random Coil (FRC) simulations performed using a modified version of the ABSINTH implicit model and CAMPARI simulation engine yield ensembles that scale as ideal chains (i.e., R_e and R_g scale with the number of monomers to the power of 0.5). **E.** Internal scaling profiles for FRC simulations and Excluded Volume (EV) simulations for poly-alanine chains of varying lengths (filled circles demark the end of profiles for different polymer lengths). Internal scaling profiles map the average distance between all pairs of residues |i-j| apart in sequence space, where i and j define two residues. This double average reports on the fact we average over both all pairs of residues that are |i-j| apart and do so over all possible configurations. EV simulations show a characteristic tapering ("dangling end" effect) for large values of |i-j|. All FRC simulation profiles superimpose on top of one another, reflecting the absence of finite chain effects. **F.** Histograms of end-to-end distances (blue) taken from FRC simulations vs. corresponding probability density profiles generated by the Analytical FRC (AFRC) model (black line) show excellent agreement. **G.** Histograms of radii of gyration (red) taken from FRC simulations vs. corresponding probability density profiles generated by the AFRC model (black line) also show excellent agreement.

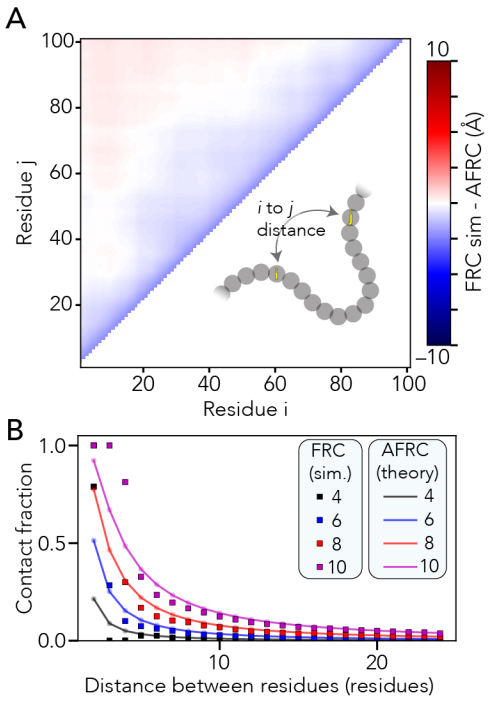


Figure 2. The AFRC enables the calculation of intra-residue distance distributions and expected distance-dependent contact fractions

A. We compared all-possible mean inter-residue distances obtained from FRC simulations with predictions from the AFRC. The maximum deviation across the entire chain is around 2.5 Å, with 92% of all distances having a deviation of less than 1 Å. **B.** Using the inter-residue distance, we can calculate the average fraction of an ensemble in which two residues are in contact (i.e., within some threshold distance). Here, we assess how that fractional contact varies with the contact threshold (different lines) and distance between the two residues. The AFRC does a somewhat poor job of estimating contact fractions for pairs of residues separated by 1,2 or 3 amino acids due to the discrete nature of the FRC simulations vis the continuous nature of the Gaussian chain distribution. However, the agreement is excellent above a sequence separation of three or more amino acids, suggesting that the AFRC offers a reasonable route to normalize expected contact frequencies.

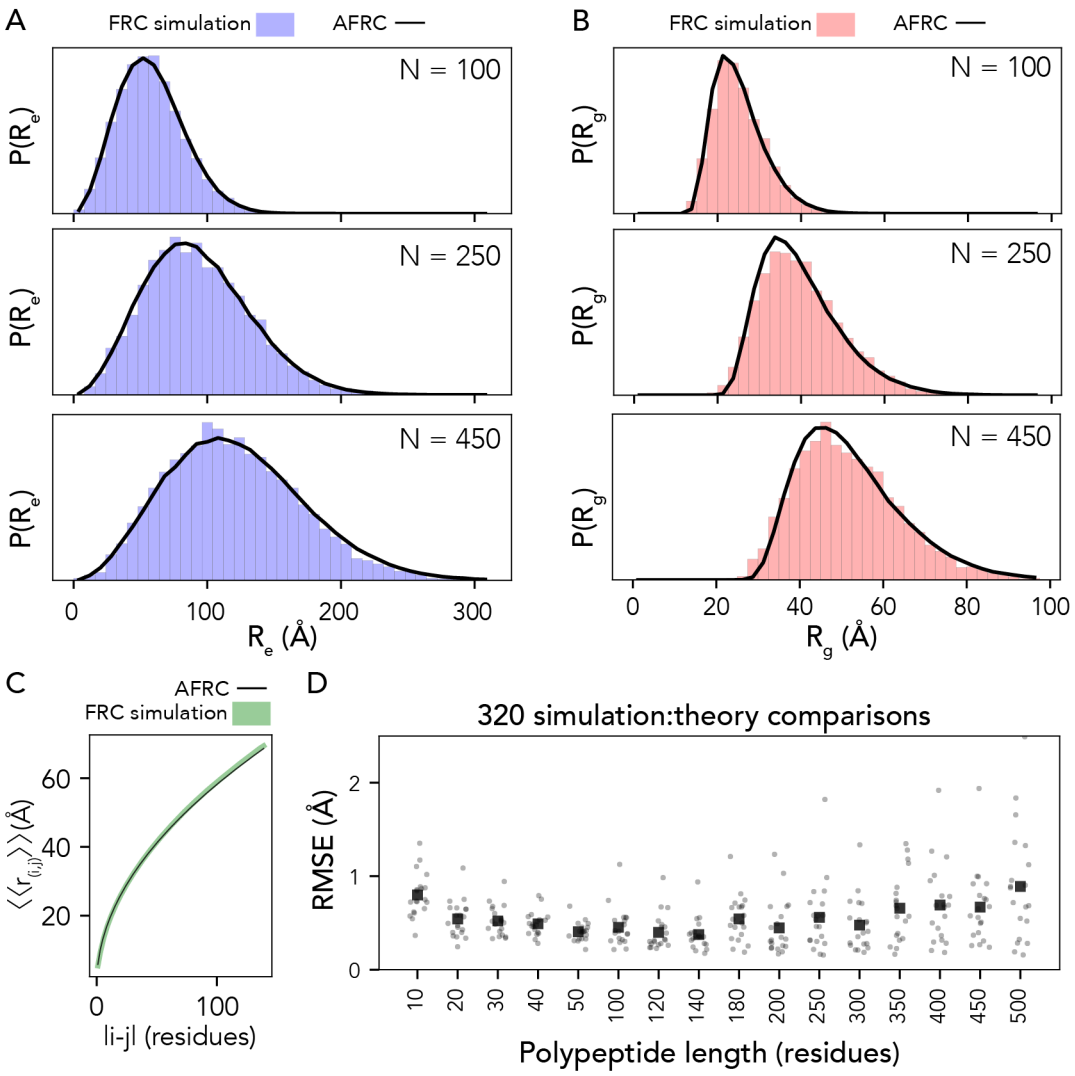


Figure 3. The AFRC generalizes to arbitrary heteropolymeric sequences with the same precision and accuracy as it does for homopolymeric sequences

A. Representative examples of randomly polypeptide heteropolymers of lengths 100, 250, and 450, comparing the AFRC-derived end-to-end distance distribution (black curve) with the empirically-determined end-to-end distance histogram from FRC simulations (blue bars). **B**. The same three polymers, as shown in A, now compare the AFRC-derived radius of gyration distance distribution (black curve) with the empirically-determined radius of gyration histogram from FRC simulations

(blue bars). **C.** Comparison of AFRC vs. FRC simulation-derived internal scaling profiles for a 150-amino acid random heteropolymer. The deviation between FRC and AFRC for these profiles offers a measure of agreement across all length scales. **D.** Comparison of root-mean-square error (RMSE) obtained from internal scaling profile comparisons (i.e., as shown in C) for 320 different heteropolymers straddling 10 to 500 amino acids in length. In all cases, the agreement with theory and simulations is excellent.

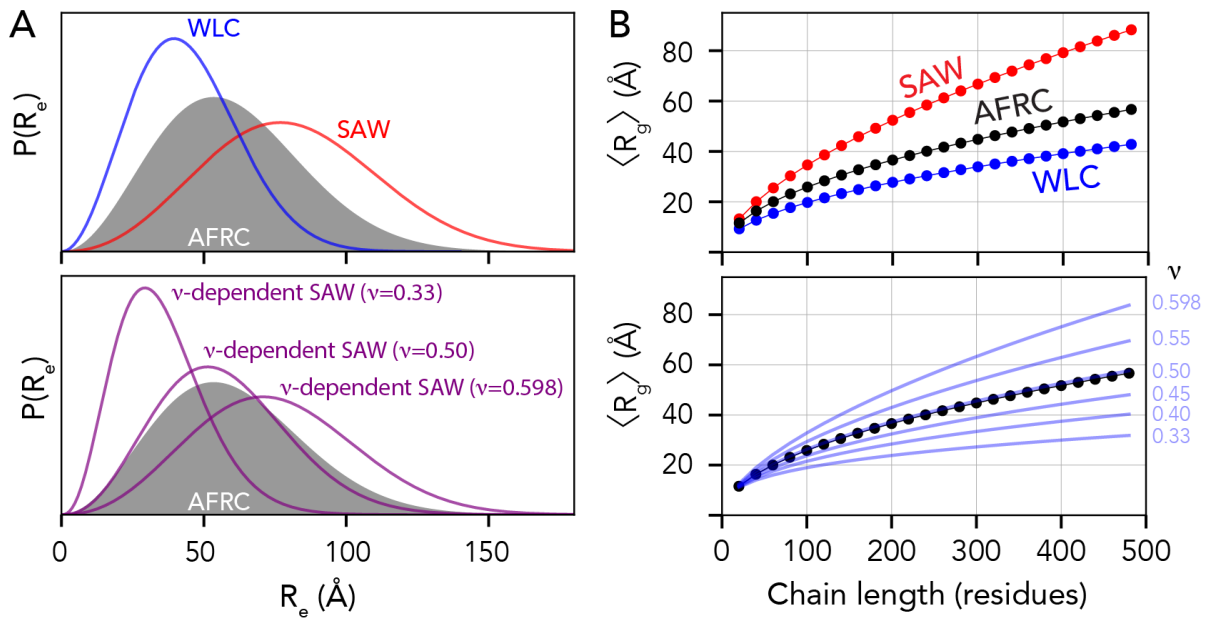


Figure 4. The AFRC is complementary to existing polymer models

A. Comparison of end-to-end distance distributions for several other analytical models, including the Wormlike Chain (WLC), the self-avoiding walk (SAW), and the ν-dependent SAW model (SAW-ν). The AFRC behaves like a ν-dependent SAW with a scaling exponent of 0.5. **B.** Comparisons of ensemble-average radii of gyration as a function of chain length for the same sets of polymer models. The AFRC behaves as expected and again is consistent with a ν-dependent SAW with a scaling exponent of 0.5.

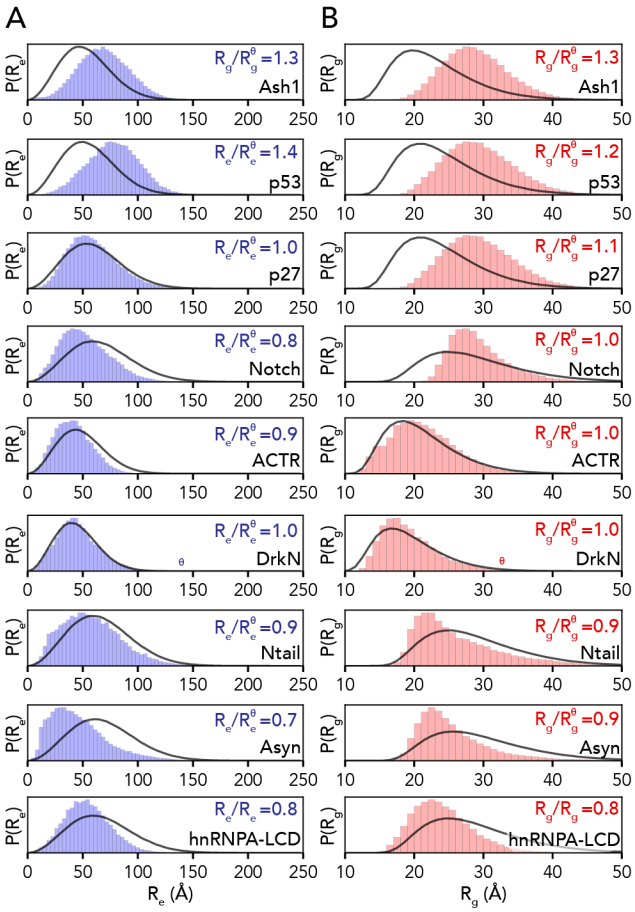


Figure 5. AFRC-derived distance distributions enable simulations to be qualitatively compared against a null model

A. Comparison of the AFRC-derived end-to-end distance distributions (black line) with the simulation-derived end-to-end distribution (blue bars) for all-atom simulations of nine different disordered proteins. **B.** Comparison of the AFRC-derived radius of gyration distributions (black line) with the simulation-derived radius of gyration distribution (red bars) for all-atom simulations of nine different disordered proteins.

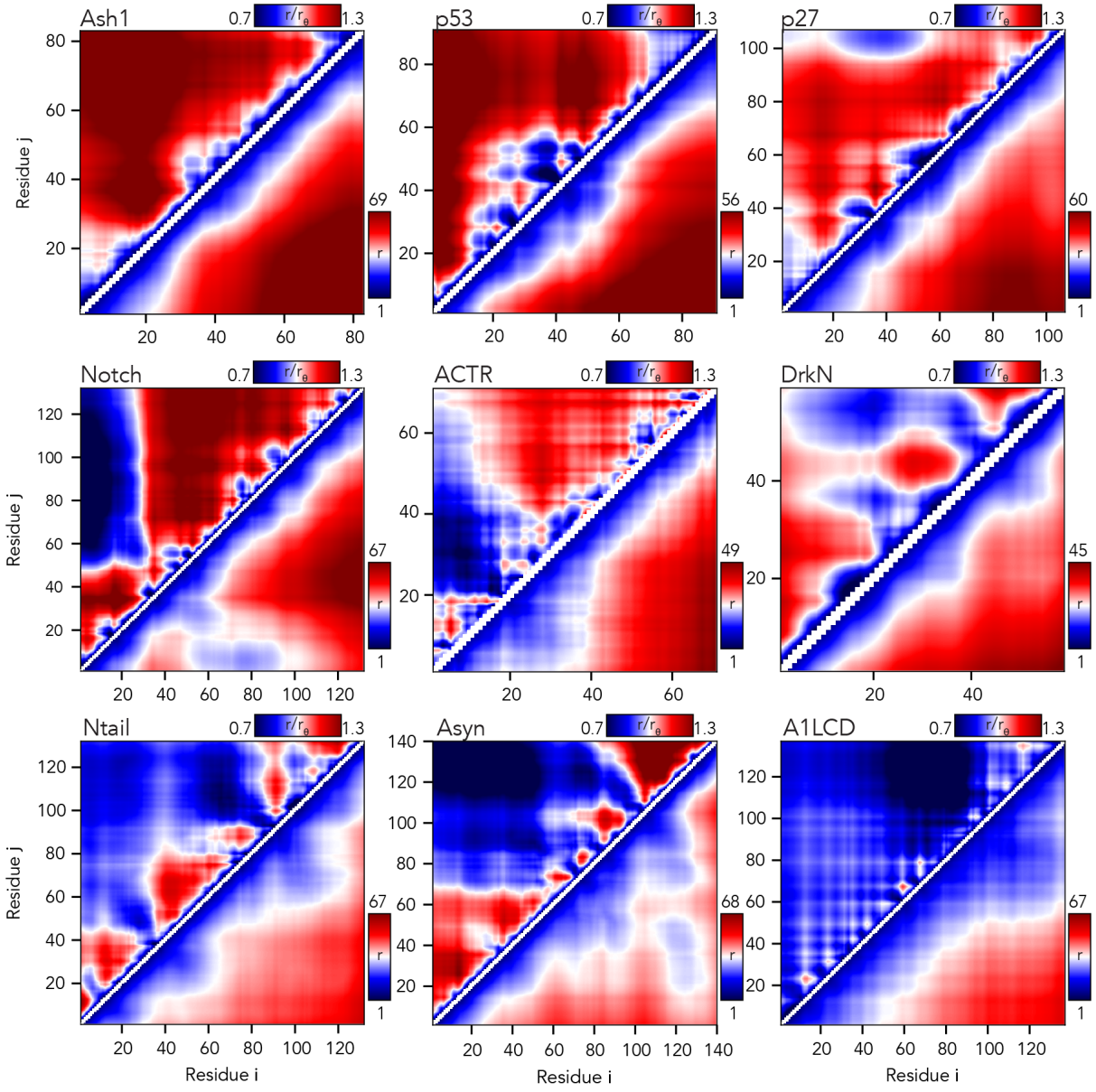


Figure 6. The AFRC enables a consistent normalization of intra-chain distances to identify specific sub-regions that are closer or further apart than expected

Inter-residue scaling maps (top left) and distance maps (bottom right) reveal the nuance of intramolecular interactions. Scaling maps (top left) report the average distance between each pair of

residues (i,j) divided by the distance expected for an AFRC-derived distance map, providing a unitless parameter that varies between 0.7 and 1.3 in these simulations. Distance maps (bottom right) report the absolute distance between each pair of residues in angstroms. While distance maps provide a measure of absolute distance in real space, scaling maps provide a cleaner, normalized route to identify deviations from expected polymer behavior, offering a convenient means to identify sequence-specific effects. For example, in Notch and alpha-synuclein, scaling maps clearly identify end-to-end distances as close than expected. Scaling maps also offer a much sharper resolution for residue-specific effects - for example, in p53, residues embedded in the hydrophobic transactivation domains are clearly identified as engaging in transient intramolecular interactions, leading to sharp deviations from expected AFRC distances.

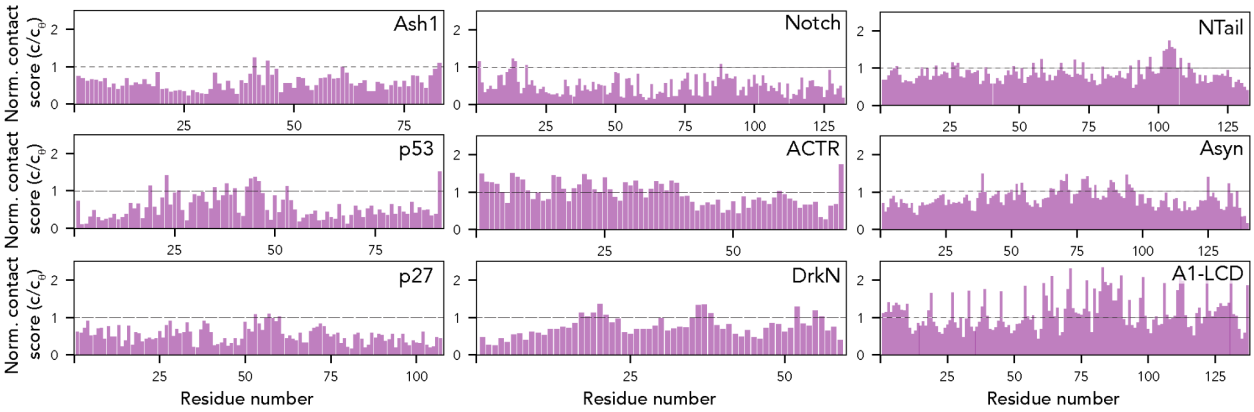


Figure 7. The AFRC enables an expected contract fraction to be calculated, such that normalized contact frequencies can be easily calculated for simulations

Across the nine different simulated disordered proteins, we computed the contact fraction (i.e., the fraction of simulations each residue is in contact with any other residue) and divided this value by the expected contact fraction from the AFRC model. This analysis revealed subregions within IDRs that contribute extensively to intramolecular interactions, mirroring finer-grain conclusions obtained in **Fig. 6**.

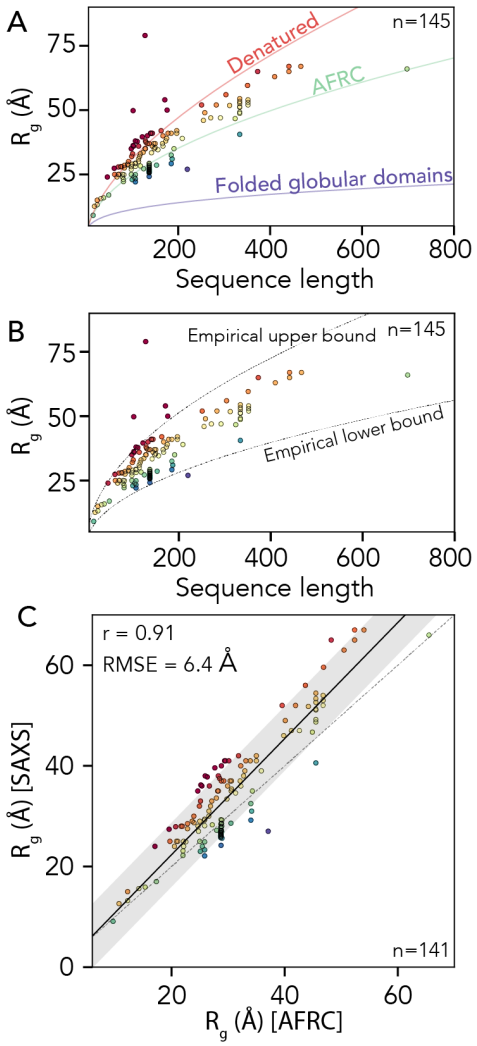


Figure 8. Comparison of AFRC-derived radii of gyration with experimentally-measured values

A. We compared 145 experimentally-measured radii of gyration against three empirical polymer scaling models that capture the three classes of polymer scaling ($\nu = 0.33$ [globular domains], $\nu = 0.5$ [AFRC], and $\nu = 0.59$ [denatured state]). Individual points are colored by their normalized radius of gyration (SAXS-derived radius of gyration divided by AFRC-derived radius of gyration). **B.** The same data as in panel A with the empirically defined upper and lower bound. As with panel A, individual points are colored by their normalized radius of gyration. **C.** Comparison of SAXS-

derived radii of gyration and AFRC-derived radii of gyration, as with panels A and B, individual points are colored by their normalized radius of gyration.

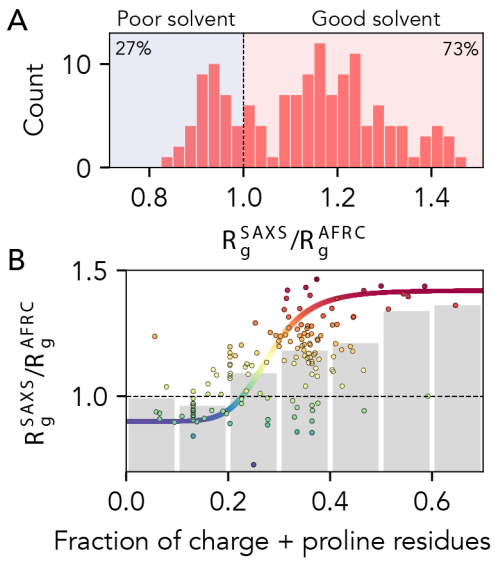


Figure 9. AFRC-normalized radii of gyration from experimentally-measured proteins

A. Histogram showing the normalized radii of gyration for 141 different experimentally-measured sequences. B. Comparison of normalized radii of gyration for 141 different experimentally-measured sequences against the fraction of charge and proline residues in those sequences. Individual points are colored by their normalized radius of gyration. Grey bars reflect the average radius of gyrations obtained by binning sequences with the corresponding fraction of charge and proline residues. The colored sigmoidal curve is included to guide the eye across the transition region, suggesting that – on average – the midpoint of this transition is at a fraction of charged and proline residues of ~0.25. The Pearson correlation coefficient (r) for the fraction of charged and proline residues vs. normalized radius of gyration is 0.58).

4.13 Supplementary Methods

Flory Random Coil (FRC) simulations, excluded volume (EV) simulations and quantification of finite size effects

Flory Random Coil (FRC) Monte Carlo simulations were run using a customized version of CAMPARI (V1). Simulations were run in a simulation droplet with a radius of 500 Å for 25 x 10^6 steps with 50 x 10^3 steps discarded as equilibration. Conformers were saved every 5 x 10^3 steps, generating 5 x 10^3 independent conformations. Because FRC simulations are rejection free, these ensembles are sufficiently well-sampled and enable calibration for FRC fitting parameters (**Table S1**).

Homopolymeric FRC simulations were run for length of 51, 101, 151, 251 and 351 residues for all twenty amino acids (*i.e.* 100 independent sequences in total). Heteropolymeric simulations were run for lengths 10, 20, 30, 40, 50, 100, 120, 140, 180, 200, 250, 300, 350, 400, 450, 500 (*i.e.* 320 independent sequences in total). For each length series, twenty separate simulations were run where, for each sequence, one of the twenty amino acids is enriched (30% of the sequence) while the remaining residues are randomly selected. All FRC simulations were analyzed using SOURSOP[584].

Excluded volume (EV) simulations were run using CAMPARI (V2). In EV simulations, the underlying energy function for the ABSINTH forcefield is altered such that solvation, attractive Lennard-Jones, and polar (charge) interactions are set to zero, as has been described previously[306]. EV simulations were used solely to compare finite-size effects for ensembles constructed for real chains. Excluded volume (EV) Monte Carlo simulations were run for homopolymers of 50, 100,

150, 200, 250, 300, 350, 400, 450, and 500 residue poly-alanine chains as a reference model to quantify finite-size effects. Simulations were run in a simulation droplet with a radius of 500 Å for 21 x 10^6 steps, with 1 x 10^6 steps discarded as equilibration. It is worth noting that given chains are generated in a random non-overlapping starting configuration and the only criterion for move acceptance or rejection is steric overlap, strictly speaking, no equilibration is needed as the chain begins the simulation "equilibrated" in the context of the underlying Hamiltonian. Conformers were saved every 2 x 10^4 steps, generating 1 x 10^3 independent conformations, a sufficiently large ensemble for our purposes of calculating internal scaling profiles, although we suggest these ensembles would not be large enough for other types of quantification (**Fig. 1E**).

For quantifying dangle end effects of internal vs. external inter-residue distances (**Fig. S1D**), we ran extensive additional simulations of an A_{151} homopolymer (to match FRC simulations). For these simulations, ten independent replicas were run for 8.05 x 10^7 steps, with the first 5 x 10^5 discarded as equilibration. Conformers were saved every 2 x 10^4 steps. These simulations generated an ensemble of 4 x 10^4 conformations, enabling a robust assessment of finite-size effects.

We assessed finite-size effects for FRC simulations in several ways, comparing against excluded volume (EV) simulations as a real-chain reference model. First, we compared internal scaling profiles. For real chains, residues at or near the ends have a great volume of space they can explore than residues internal to chain due to excluded volume of the chain. This manifests for internal scaling profiles whereby super-imposing a series of homopolymers of different lengths reveals the distance between residue 1 and n when 1 and *n* are the first and terminal residues is shorter than residue 1 and n when n is an internal residue (**Fig. 1E**). In contrast, because FRC simulations lack

any excluded volume contribution, there is no difference between internal and external residues, such that all inter-residue distances of the same residue spacing are equivalent, regardless of where in the chain the two residues lie. This is even more clearly shown by calculating the normalized distance for different inter-residue spacing as a function of starting residue (**Fig. S2C, D**).

Second, we calculated the Flory characteristic ratio as;

$$C_n = \frac{\langle R^2 \rangle}{nl^2} \quad \text{(Eq. 1)}$$

Where n is the number of residues, l is the monomer size, and $\langle R^2 \rangle$ is the ensemble-average squared end-to-end distance (or inter-residue distance)[398]. Given both the FRC and AFRC models describe ideal chains, we can empirically define l as using the standard ideal chain relationship;

$$l = \sqrt{\frac{\langle R^2 \rangle}{n}} \quad \text{(Eq. 2)}$$

in the limit of n tending to ∞[398].

By defining l empirically from our FRC simulations or AFRC model, finite size effects emerge upon plotting n vs. C_n (**Fig. S2E,F**). In FRC simulations, C_n is less than 1 for shorter chains. This is expected in that the rotational isomeric state means local chain geometry is not truly ideal but instead limited to the inter-residue vector path defined by the Ramachandran isomeric states. In contrast, the AFRC is a true ideal chain model, such that the Flory characteristic ratio is always 1 regardless of n. This difference between the AFRC and FRC models manifests as a very slight (1-2 Å) difference in intramolecular distances visible in **Fig. 2A**.

All-atom simulations

All-atom simulations were analyzed as described previously, and all the all-atom trajectories can be obtained as described previously[584]. Specifically, all-atom simulations included both Monte Carlo and molecular dynamics simulations. Monte Carlo simulations include those of Ash1[93], p53[91], p27[603], the notch intracellular domain[602], the hnRNPA1 low complexity domain[80]. Molecular dynamics simulations include alpha-synuclein, DrkN, ACTR and NTail [99].

SAXS data

Experimental SAXS data includes 145 separate radius of gyration values. All values and associated references are included in table S4. In addition, all data are tabulated at the main GitHub directory for this paper (https://github.com/holehouse-lab/supportingdata/tree/master/2023/alston_ginell_2023) and available as an Excel spreadsheet and Pandas-compatible CSV file.

Amino acid sequence analysis

Sequence analysis to calculate the fraction of charged residues and proline residues was done using localCIDER [455] and sparrow (https://github.com/idptools/sparrow).

AFRC implementation

The AFRC is implemented as a stand-alone Python package. All code is open-sourced and available at https://github.com/idptools/afrc. All documentation is available at https://afrc.readthedocs.io/. The package itself can be downloaded from https://pypi.org/project/afrc and installed using the command **pip install afrc**. A Google colab notebook that implements the AFRC along with the

other three analytical models described in this work are linked from https://github.com/idptools/afrc. The afrc package uses numpy and scipy, and in addition to the AFRC implements the Worm-like chain (WLC), the self-avoiding random walk (SAW), and the ν-dependent self-avoiding random walk (SAW-ν) [120,173].

Figures and analysis in this paper

Jupyter notebooks to recreate all figures in this paper are available at https://github.com/holehouse-lab/supportingdata/tree/master/2023/alston_ginell_2023.

4.14 Supplementary Figures

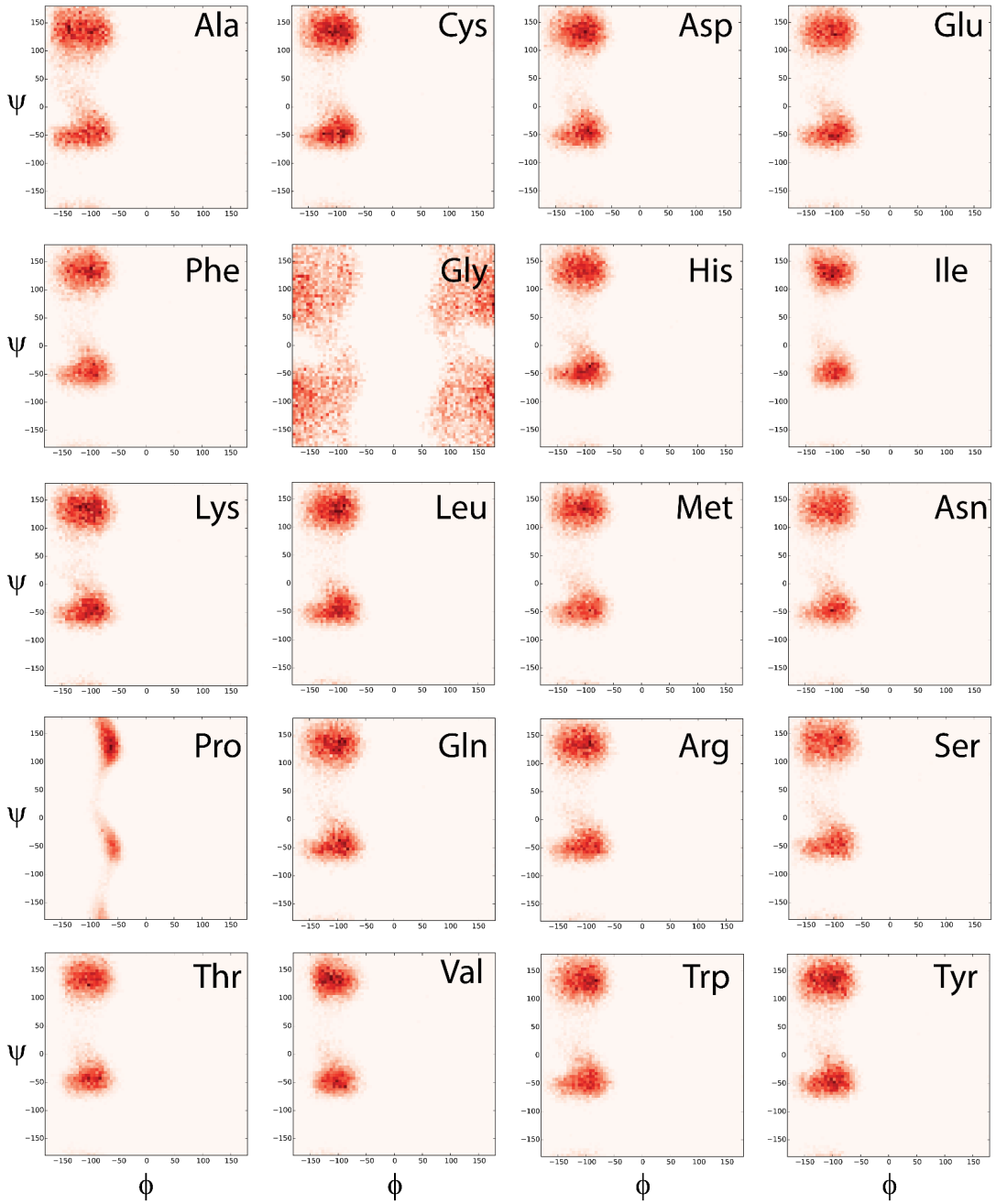


Fig. S1 Residue-specific Ramachandran maps used for FRC simulations

Ramachandran maps for all twenty amino acids performed as excluded volume simulations define the allowed isomeric states and are used by FRC simulations to construct the FRC ensembles.

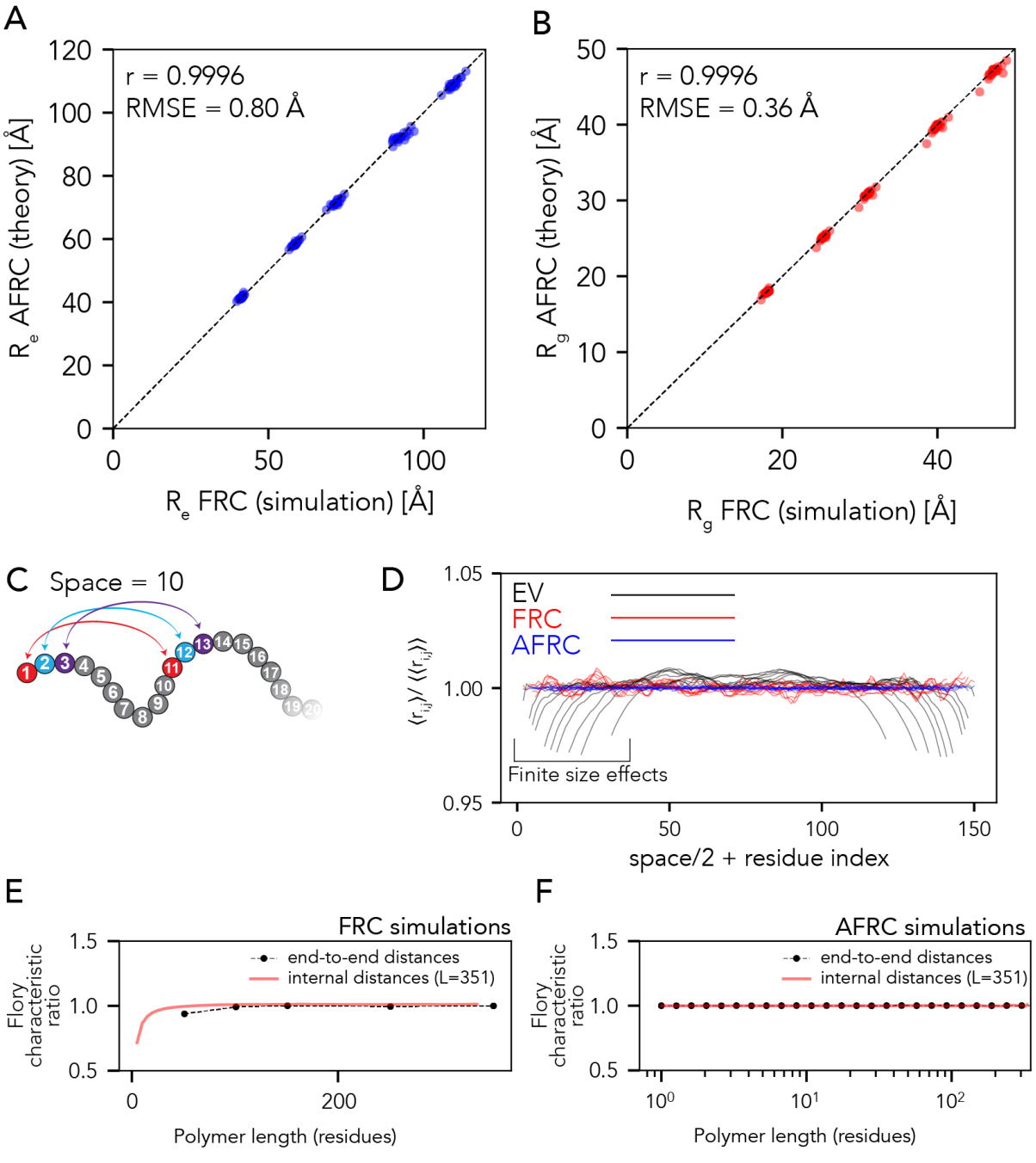
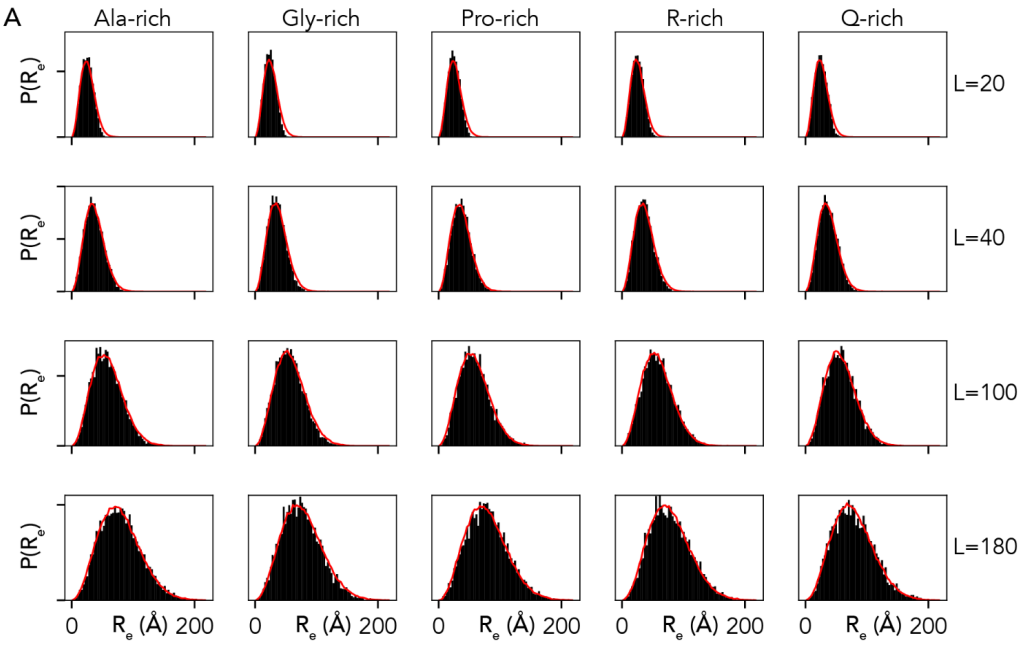


Fig. S2 Comparison between global dimensions from simulations vs. AFRC

A. The correlation between the end-to-end distance (R_e) obtained from FRC simulations and AFRC analysis is shown. The comparisons here are for ensemble-average values for homopolymers derived from the twenty different amino acids for lengths of 51, 101, 151, 251, and 351 residues. **B.** The correlation between radius of gyration (R_g) values obtained from FRC simulations and AFRC analysis. Again, the comparisons here are for ensemble-average values for homopolymers derived from the twenty different amino acids for lengths of 51, 101, 151, 251 and 351 residues. **C.** Schematic of the approach taken in panel D. **D.** For a 151-residue homopolymer, we calculated the average distance between all pairs of residues that are a fixed spacing apart for EV and FRC simulations and for the AFRC model. The inter-residue spacing used were 2, 6, 8, 10, 16, 20, 24, 32, 40, and 60 residues, and each spacing yields a different line. For example, for a spacing of 6 residues, we calculated the average distance between the following pairs of residues $\langle r_{1,7} \rangle$, $\langle r_{2,8} \rangle$, ..., $\langle r_{145,151} \rangle$. Note the angle brackets here denote the ensemble-average distance. Each line represents the profile revealed by the set of inter-residue distances. For every point along the line, the y-axis position reports on the average distance normalized by the overall average distance for all residues of a given spacing. In contrast, the x-axis position is the location of the first residue of the two in a pair, to which half of the inter-residue spacing is added. For example, if we examined positions for $\langle r_{1,7} \rangle$, $\langle r_{2,8} \rangle$, ..., $\langle r_{145,151} \rangle$ then the corresponding x-axis positions would be ($1 + 0.5\times6 = 4$, $2 + 0.5\times6 = 5$, …, $145 + 0.5^*\times6 = 148$). We take this approach such that the middle of the x-axis in the figure always corresponds to the central position in the polymer. For EV simulations, when one of the two residues in a pair falls near the end of the chain, we see a suppression of the inter-residue distances compared to the same inter-residue distance when both positions are internal to the chain. This is the expected result and reflects the fact that internal residues are 'repelled' by steric overlap with other residues, whereas end residues are less constrained. For FRC simulations and AFRC models, no such end effects are observed, reflecting the finite-size end effects do not influence ideal chains. **E**. We also calculated the Flory characteristic ratio (C_n) for chains of different lengths (black circles) and for intramolecular distances (red lines) for FRC simulations. The characteristic ratio enables correlations in chain dimensions to be assessed, and for FRC simulations, we see the expected deviation from 1 at shorter chain lengths (see supplemental methods). While these deviations are expected finite-size effects, their impact when comparing inter-residue distances is minimal (**Fig. 2**).


A
Ala-rich
Gly-rich
Pro-rich
R-rich
Q-rich
P(R_e)
L=20
L=40
L=100
L=180
0
R_e (Å) 200

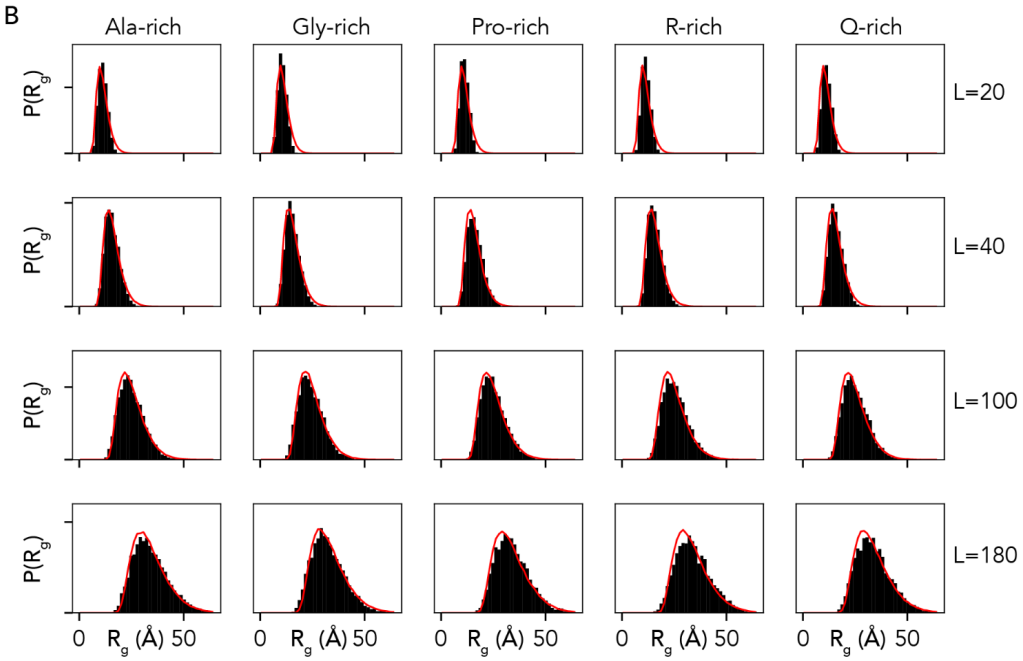

B
Ala-rich
Gly-rich
Pro-rich
R-rich
Q-rich
P(R_g)
L=20
L=40
L=100
L=180
0
R_g (Å) 50

Fig. S3 Comparison of end-to-end distance distributions and radii of gyration distributions for select heteropolymers of variable composition and length

A. Comparison of end-to-end distance distributions. Empirical distributions obtained from simulations are shown in black, while predictions of the distribution from the AFRC are shown as red lines. **B.** Comparison of radii of gyration distributions. Empirical distributions obtained from simulations shown in black, while predictions of the distribution from the AFRC are shown as red lines.

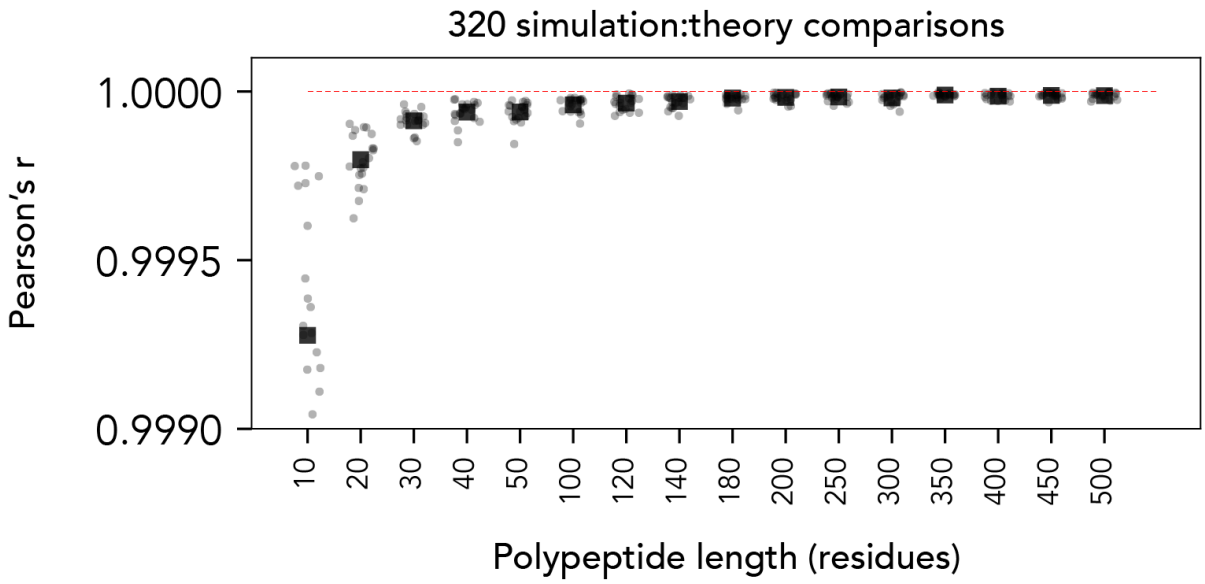


Fig. S4. Correlation between internal scaling profiles for random heteropolymers from FRC simulations vs. AFRC-derived internal scaling profiles

For each length (10,20,30, …, 500) 20 different heteropolymers, were generated where each heteropolymer is enriched (30%) in one of the twenty amino acids while the remaining residues are randomly selected. This yields 320 different internal scaling comparisons (16 lengths with 20 amino acids).

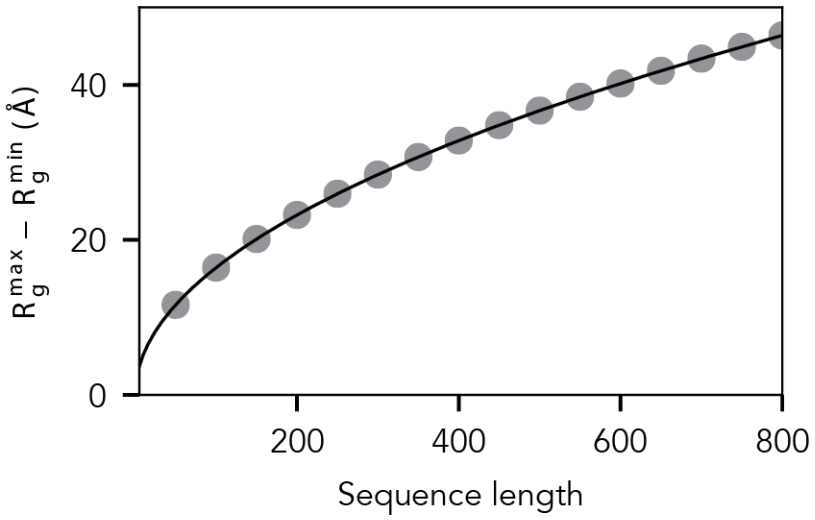


Fig. S5.

Difference in radii of gyration based on empirical min and max values reveals the length-dependent variation in expected accessible radii of gyration values.

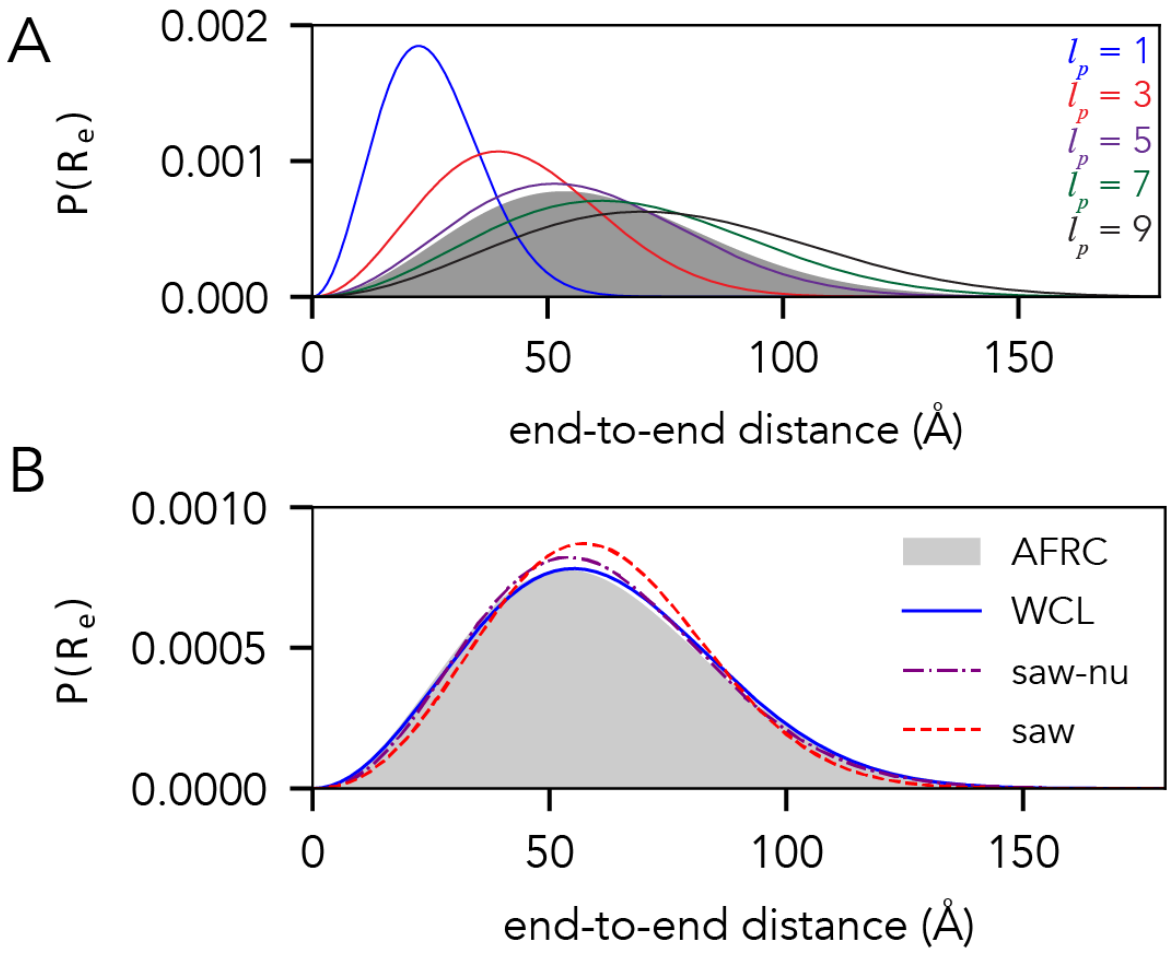


Fig. S6. Comparison of the end-to-end distance distributions for the AFRC with existing polymer models

A. Comparison of the AFRC model (grey shaded area) for 100-residue polyalanine chain (A_{100}) with Worm-Like chain (WLC)-derived distributions, where the WLC monomer size is fixed at 3.8 Å, and the persistence length varies from 1 Å to 9 Å. **B.** Comparison of AFRC, WLC, SAW-ν, and SAW models in which model input parameters were selected to reproduce the AFRC end-to-end distance distribution for an A_{100} chain. The WLC model uses an amino acid size of 3.8 Å and a persistence length of 5.7 Å. The SAW-ν model uses a prefactor of 5.8 Å and a ν of 0.5. The SAW model uses a prefactor of 4.1 Å.

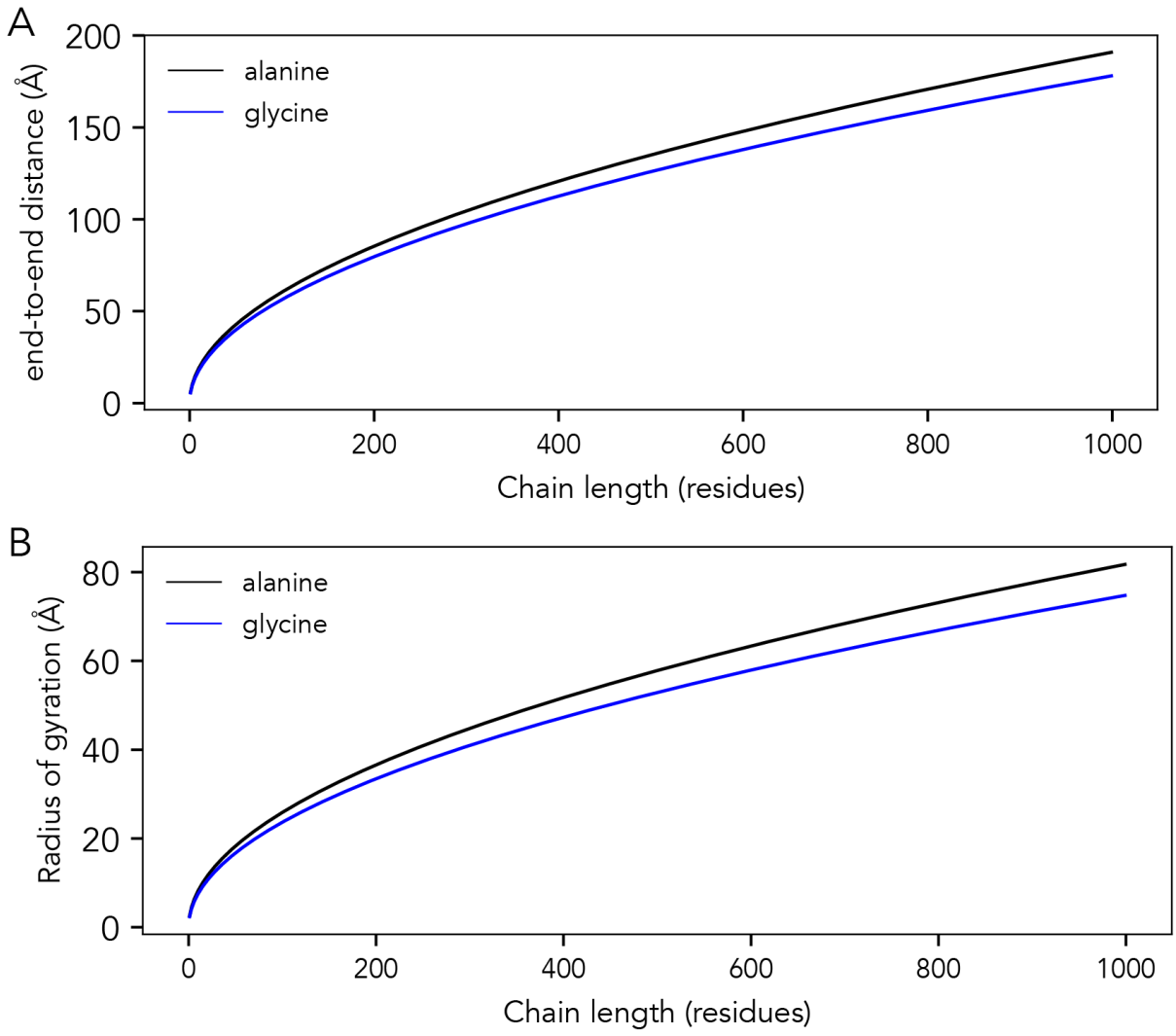


Fig. S7. Comparison of chain dimensions obtained from the AFRC model:

for poly-alanine vs. poly-glycine, examining end-to-end distance (**A**) and radius of gyration (**B**).

4.15 Supplementary Tables

Amino acid	R_{ij} RMS (Å)	R_{ij} (Å)	X_0 ($Å^{-1}$)
A	6.5463	6.0381	0.5405
C	6.2676	5.7826	0.5635
D	6.3994	5.911	0.5567
E	6.2649	5.768	0.5613
F	6.2519	5.7612	0.5571
G	6.1045	5.6324	0.5911
H	6.2156	5.7262	0.5645
I	6.4353	5.9361	0.5483
K	6.306	5.8272	0.5533
L	6.2636	5.7801	0.5605
M	6.3813	5.8894	0.5501
N	6.2652	5.773	0.5598
P	6.4323	5.9388	0.5599
Q	6.2547	5.7719	0.5617
R	6.279	5.7921	0.5531
S	6.3161	5.8364	0.5553
T	6.1995	5.7242	0.5695
V	6.3204	5.8409	0.5571
W	6.3	5.814	0.5539
Y	6.3188	5.8266	0.5543

Table S1 Model parameters obtained by fitting against FRC simulations.

Name	Sequence
Ash1	GASASSSPSP STPTKSGKMR SRSSSPVRPK AYTPSPRSPN YHRFALDSPP QSPRRSSNSS ITKKGSRRSS GSSPTRHTTR VCV
p53	MEEPQSDPSV EPPLSQETFS DLWKLLPENN VLSPLPSQAM DDLMLSPDDI EQWFTEDPGP DEAPRMPEAA PPVAPAPAAP TPAAPAPAPS W
p27	GSHMKGACKV PAQESQDVSG SRPAAPLIGA PANSEDTHLV DPKTDPSDSQ TGLAEQCAGI RKRPATDDSS TQNKRANRTE ENVSDGSPNA GSVEQTPKKP GLRRRQT
Notch	MARKRRRQHG QLWFPEGFKV SEASKKKRRE PLGEDSVGLK PLKNASDGAL MDDNQNEWGD EDLETKKFRF EEPVVLPDLD DQTDHRQWTQ QHLDAADLRM SAMAPTPPQG EVDADCMDVN VRGPDGFTPL LE
ACTR	GTQNRPLLRN SLDDLVGPPS NLEGQSDERA LLDQLHTLLS NTDATGLEEI DRALGIPELV NQGQALEPKQ D
drkN	MEAIAKHDFS ATADDELSFR KTQILKILNM EDDSNWYRAE LDGKEGLIPS NYIEMKNHD
Ntail	MHHHHHHTTE DKISRAVGPR QAQVSFLHGD QSENELPRLG GKEDRRVKQS RGEARESYRE TGPSRASDAR AAHLPTGTPL DIDTASESSQ DPQDSRRSAD ALLRLQAMAG ISEEQGSDTD TPIVYNDRNL LD
asyn	MDVFMKGLSK AKEGVVAAAE KTKQGVAEAA GKTKEGVLYV GSKTKEGVVH GVATVAEKTK EQVTNVGGAV VTGVTAVAQK TVEGAGSIAA ATGFVKKDQL GKNEEGAPQE GILEDMPVDP DNEAYEMPSE EGYQDYEPEA
A1-LCD	GSMASASSSQ RGRSGSGNFG GGRGGGFGGN DNFGRGGNFS GRGGFGGSRG GGGYGGSGDG YNGFGNDGSN FGGGGSYNDF GNYNNQSSNF GPMKGGNFGG RSSGPYGGGG QYFAKPRNQG GYGGSSSSSS YGSGRRF

Table S2. Sequences from simulations

Full sequences used from all-atom simulations. Amino acids are colored by chemical type as per localCIDER[455].

Name	N	R_g (Å)	R_g/R_g^θ	R_e (Å)	R_e/R_e^θ	ν^{app} [a]	Quality of ν^{app} fit [b]
Ash1	83	28.9	1.27	68.95	1.30	0.61	GOOD
p53	91	29.4	1.23	77.73	1.39	0.66	GOOD
p27	107	28.3	1.09	59.15	0.98	0.49	POOR
Notch	132	29.3	1.02	52.16	0.78	0.34	POOR
ACTR	71	21.1	1.01	41.45	0.85	0.50	GOOD
drkN	59	19.3	1.00	45.26	1.01	0.43	GOOD
Ntail	132	26.3	0.92	58.11	0.87	0.39	POOR
asyn	140	25.6	0.87	46.47	0.67	0.23	POOR
A1-LCD	137	24.1	0.84	54.37	0.81	0.47	GOOD

[a] Estimated ν^{app} based on linear fitting of the internal scaling regime using SOURSOP.
[b] Quality of fit based on the reduced chi-squared from the fit.

Table S3: Simulation and AFRC-derived parameters for all-atom simulations

Table S4: SAXS sequences and values

(note table caption comes before table as table is 36 pages long).

Protein name	R_g (Å)	R_g error (Å)	Amino acid sequence	Reference
Nucleoporin Nup49 (N49)	15.9	1.3	GCQTSRGLFGNNNTNNINNSSSGMNNASAGLFGSKPCA	Fuertes, et al. PNAS (2017) 114, E6342–E6351.
Heh2 (NLS)	24	3	ACETNKRKREQISTDNEAKMQIQEEKSPKKKRKKRSSKANKPPECA	Fuertes, et al. PNAS (2017) 114, E6342–E6351.
VSV Protein Phosphoprotein P	24	1	HHHHHELMDNLTKVREYLKSYSRLDQAVGEIDEIEAQRAEKSNYELFQEDGVEEHTKPSYFQAADDS	Leyrat, C., Jensen, M.R., Ribeiro, E.A., Gérard, F.C.A., Ruigrok, R.W.H., Blackledge, M., and Jamin, M. (2011). The N0-binding region of the vesicular stomatitis virus phosphoprotein is globally disordered but contains transient α-helices. Protein Sci. 20, 542–556.
LS	27.9	1	SPPGKPQGPPQQEGNKPQGPPPPGKPQGPPPAGGNPQQPQAPPAGKPQGPPPPPQGGRPPRPAQGQQPPQ	Boze, H., Marlin, T., Durand, D., Pérez, J., Vernhet, A., Canon, F., Sarni-Manchado, P., Cheynier, V., and Cabane, B. (2010). Proline-rich salivary proteins have extended conformations. Biophys. J. 99, 656–

				665.
Nup153_NUS	24.9	1.3	GCPSASPAFGANQTPTFGQSQGASQPNPPGFGSISSSTALFPTGSQPAPPTFGTVSSSSQPPVFGQQPSQSAFGSGTTPNCA	Fuertes, et al. PNAS (2017) 114, E6342–E6351.
Sic1	30	4	GSMTPSTPPRSRGTRYLAQPSGNTSSSALMQGQKTPQKPSQNLVPVTPSTTKSFKNAPLLAPPNSNMGMTSPFNGLTSPQRSPFPKSSVKRT	Gomes G-NW, Krzeminski M, Namini A, Martin EW, Mittag T, Head-Gordon T, et al. Conformational Ensembles of an Intrinsically Disordered Protein Consistent with NMR, SAXS, and Single-Molecule FRET. J Am Chem Soc. 2020;142: 15697–15710.
chloroplastic calvin cycle protein	23		HHHHHHHHHHSSGHIEGRHMSGQPAVDLNKKVQDAVKEAEDACAKGTSADCAVAWDTVEELSAAVSHKKDAVKADVTLTDPLEAFCKDAPDADECRVYED	Launay H, Barré P, Puppo C, Zhang Y, Maneville S, Gontero B, Receveur-Bréchot V, J Mol Biol 430(8):1218-1234 (2018)
Antitermination protein N (from lambda phage)	38	3.5	MDAQTRRRERRAEKQAQWKAANPLLVGVSAKPVNRPILSLNRKPKSRVESALNPIDLTVLAEYHKQIESNLQRIERKNQRTWYSKPGERGITCSGRQKIKGKSIPLI	Johansen, D., Trewhella, J., and Goldenberg, D.P. (2011). Fractal dimension of an intrinsically disordered protein: small-angle X-ray scattering and computational study of the bacteriophage λ N protein. Protein Sci. 20, 1955–1970.

Nup153_NUL	30	3	GCGFKGFDTSSSSSNSAASSSFKFGVSSSSSGPSQTLTSTGNFKFGDQGGFKIGVSSDSGSINPMSEGFKFSKPIGDFKFGVSSESKPEEVKKDSKNDNFKFGLSSGLSNPVCA	Fuertes, et al. PNAS (2017) 114, E6342–E6351.
DARPP-32 (aka Protein phosphatase 1 regulatory subunit 1B)	28.28		MDPKDRKKIQFSVPAPPSQLDPRQVEMIRRRRPTPALLFRVSEHSSPEEESSPHQRTSGEGHHPKSKRPNPCAYTPPSLKAVQRIAESHLQTISNLSENQASEEEDELGELRELGYPQ	Marsh, J.A., Dancheck, B., Ragusa, M.J., Allaire, M., Forman-Kay, J.D., and Peti, W. (2010). Structural diversity in free and bound states of intrinsically disordered protein phosphatase 1 regulators. Structure 18, 1094–1103.
II-1	41		GKPVGRRPQGGNQPQRPPPPPGKPQGPPPQGGNQSQGPPPPPGKPEGRPPQGRNQSQGPPPHPGKPERPPPQGGNQSQGTPPPPGKPERPPPQGGNQSHRPPPPPGKPERPPPQGGNQSRGPPPHRGKPEGPPPQEGNKSR	Boze, H., Marlin, T., Durand, D., Pérez, J., Vernhet, A., Canon, F., Sarni-Manchado, P., Cheynier, V., and Cabane, B. (2010). Proline-rich salivary proteins have extended conformations. Biophys. J. 99, 656–665.
Fhua	33.4		ESAWGPAATIAARQSATGTKTDTPIQKVPQSISVVTAEEMALHQPKSVKEALSYTPGVSVGTRGASNTYDHLIIRGFAAEGQSQNNYLNGLKLQGNFYNDAVIDPYMLERAEIMRGPVSVLYGKSSPGGLLNMV	Riback, J.A., Bowman, M.A., Zmyslowski, A.M., Knoverek, C.R., Jumper, J.M., Hinshaw, J.R., Kaye, E.B., Freed, K.F., Clark, P.L., and Sosnick, T.R. (2017). Innovative scattering analysis shows that hydrophobic disordered proteins are

			SKRPTTEP	expanded in water. Science 358, 238–241.
N98	28.6	1.3	GCFNKSFGTPFGGGTGGFGTTSTFGQNTGFGTTSGGAFGTSAFGSSNNTGGLFGNSQTKPGGLFGTSSFSQPATSTSTGFGFGTSTGTANTLFGTASTGTSLFSSQNNAFAQNKPTGFGNFGTSTSSGGLFGTTNTTSNPFGSTSGSLFGPCA	Fuertes, et al. PNAS (2017) 114, E6342–E6351.
Protein Phosphatase Inhibitor 2	34.6		PIKGILKNKTSTTSSMVASAEQPRGNVDEELSKKSQKWDEMNILATYHPADKDYGLMKIDEPSTPYHSMMGDDEDACSDTEATEAMAPDILARKLAAAEGLEPKYRIQEQESSGEEDSDLSPEEREKKRQFEMKRKLHYNEGLNIKLARQLISKDL	Marsh, J.A., Dancheck, B., Ragusa, M.J., Allaire, M., Forman-Kay, J.D., and Peti, W. (2010). Structural diversity in free and bound states of intrinsically disordered protein phosphatase 1 regulators. Structure 18, 1094–1103.
Nsp1	41	3	GCNFNTPQQNKTPFSFGTANNNSNTTNQNSSTGAGAFGTGQSTFGFNNSAPNNTNNANSSITPAFGSNNTGNTAFGNSNPTSNVFGSNNSTTNTFGSNSAGTSLFGSSSAQQTKSNGTAGGNTFGSSSLFNNSTNSNTTKPAFGGLNFGGGNNTTPSSTGNANTSNNLFGATANANCA	Fuertes, et al. PNAS (2017) 114, E6342–E6351.

IBB	32	2	GCTNENANTPAARLHRFK NKGKDSTEMRRRRIEVNVE LRKAKKDDQMLKRRNVSS FPDDATSPLQENRNNQGT VNWSVDDIVKGINSSNVEN QLQATCA	Fuertes, et al. PNAS (2017) 114, E6342–E6351.
Ash1	28.5	3.4	GASASSSPSPSTPTKSGKMRS RSSSPVRPKAYTPSPRSPNYH RFALDSPPQSPRRSSNSSITK KGSRRSSGSSPTRHTTRVCV	Martin, E.W., Holehouse, A.S., Grace, C.R., Hughes, A., Pappu, R.V., and Mittag, T. (2016). Sequence Determinants of the Conformational Properties of an Intrinsically Disordered Protein Prior to and upon Multisite Phosphorylation. J. Am. Chem. Soc. 138, 15323–15335.
pAsh1	27.5	1.2	GASASSSPSPSTPTKSGKMRS RSSSPVRPKAYTPSPRSPNYH RFALDSPPQSPRRSSNSSITK KGSRRSSGSSPTRHTTRVCV	Martin, E.W., Holehouse, A.S., Grace, C.R., Hughes, A., Pappu, R.V., and Mittag, T. (2016). Sequence Determinants of the Conformational Properties of an Intrinsically Disordered Protein Prior to and upon Multisite Phosphorylation. J. Am. Chem. Soc. 138, 15323–15335.
PIR domain (GRB14)	27		YGMQLYQNYMHPYQGRSG CSSQSISPMRSISENSLVAMD FSGQKSRVIENPTEALSVAV EEGLAWRKKGCLRLGTHG SPTASSQSSATNMAIHRSQP W	Moncoq, K., Broutin, I., Craescu, C.T., Vachette, P., Ducruix, A., and Durand, D. (2004). SAXS study of the PIR domain from the Grb14 molecular adaptor: a natively unfolded protein with a transient structure

				primer? Biophys. J. 87, 4056–4064.
RpII215_gibbs	28	0.7	YSPGNAYSPSSSNYSPNSPSYSPTSPSYSPSSPSYSPTSPCYSPTSPSYSPTSPNYTPVTPSYSPTSPNYSASPQ	Gibbs, E.B., Lu, F., Portz, B., Fisher, M.J., Medellin, B.P., Laremore, T.N., Zhang, Y.J., Gilmour, D.S., and Showalter, S.A. (2017). Phosphorylation induces sequence-specific conformational switches in the RNA polymerase II C-terminal domain. Nat. Commun. 8, 15233.
RpII215_portz	51.8		SPSYSPTSPNYTASSPGGASPNYSPSSPNYSPTSPLYASPRYASTTPNFNPQSTGYSPSSSGYSPTSPVYSPTVQFQSSPSFAGSGSNIYSPGNAYSPSSSNYSPNSPSYSPTSPSYSPSSPSYSPTSPCYSPTSPSYSPTSPNYTPVTPSYSPTSPNYSASPQYSPASPAYSQTGVKYSPTSPTYSPPSPSYDGSPGSPQYTPGSPQYSPASPKYSPTSPLYSPSSPQHSPSNQYSPTGSTYSATSPRYSPNMSIYSPSSTKYSPTSPTYTPTARNYSPTSPMYSPTAPSHYSPTSPAYSPSSPTFEESED	Portz, B., Lu, F., Gibbs, E.B., Mayfield, J.E., Rachel Mehaffey, M., Zhang, Y.J., Brodbelt, J.S., Showalter, S.A., and Gilmour, D.S. (2017). Structural heterogeneity in the intrinsically disordered RNA polymerase II C-terminal domain. Nat. Commun. 8, 15231.
ACTR	25		GPSGTQNRPLLRNSLDDLVGPPSNLEGQSDERALLDQLHTLLSNTDATGLEEIDRALGIPELVNQGQALEPKQDSGGPR	Borgia, A., Zheng, W., Buholzer, K., Borgia, M.B., Schüler, A., Hofmann, H., Soranno, A., Nettels, D., Gast, K., Grishaev, A., et al. (2016). Consistent View of Polypeptide Chain Expansion

				in Chemical Denaturants from Multiple Experimental Methods. J. Am. Chem. Soc. 138, 11714–11726.
Msh6	56	2	MAPATPKTSKTAHFENGSTSSQKKMKQSSLLSFFSKQVPSGTPSKKVQKPTPATLENTATDKITKNPQGGKTGKLFVDVDEDNDLTIAEETVSTVRSDIMHSQEPQSDTMLNSNTTEPKSTTTDEDLSSSQSRRNHKRRVNYAESDDDDSDTTFTAKRKKGKVVDSESDEDEYLPDKNDGDEDDDIADDKEDIKGELAEDSGDDDDLISLAETTSKKKFSYNTSHSSSPFTRNISRDNSKKKSRPNQAPSRSYNPSHSQPSATSKSSKFNKQNEERYQWLVDERDAQRRPKSDPEYDPRTLYIP	Shell, S.S., Putnam, C.D., and Kolodner, R.D. (2007). The N terminus of Saccharomyces cerevisiae Msh6 is an unstructured tether to PCNA. Mol. Cell 26, 565–578.
AN16	50	2	AQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGAP	Nairn, K.M., Lyons, R.E., Mulder, R.J., Mudie, S.T., Cookson, D.J., Lesieur, E., Kim, M., Lau, D., Scholes, F.H., and Elvin, C.M. (2008). A synthetic resilin is largely unstructured. Biophys. J. 95, 3358–3365.

HrpO	35		MEDTLEDDPQRAALEQVISLLTPVRQHRQASAERAHRHAQVELKSMLDHLSKIRASLDQERDNHKRRREGLSQEHLEKTISPNDIDRWHEKEKHMLDRLACIRQDVQQQQLRVAEQQALLEQKRLQAKASQRAVEKLACMEETLNEEG	Gazi, A.D., Bastaki, M., Charova, S.N., Gkougkoulia, E.A., Kapellios, E.A., Panopoulos, N.J., and Kokkinidis, M. (2008). Evidence for a Coiled-coil Interaction Mode of Disordered Proteins from Bacterial Type III Secretion Systems. J. Biol. Chem. 283, 34062–34068.
alpha-syn	41	1	MDVFMKGLSKAKEGVVAAAEKTKQGVAEAAGKTKEGVLYVGSKTKEGVVHGVATVAEKTKEQVTNVGGAVVTGVTAVAQKTVEGAGSIAAATGFVKKDQLGKNEEGAPQEGILEDMPVDPDNEAYEMPSEEGYQDYEPEA	Uversky, V.N., Li, J., Souillac, P., Millett, I.S., Doniach, S., Jakes, R., Goedert, M., and Fink, A.L. (2002). Biophysical properties of the synucleins and their propensities to fibrillate: inhibition of alpha-synuclein assembly by beta- and gamma-synucleins. J. Biol. Chem. 277, 11970–11978.
NTail	27.2	0.5	TTEDKISRAVGPRQAQVSFLHGDQSENELPRLGGKEDRRVKQSRGEARESYRETGPSRASDARAAHLPTGTPLDIDTASESSQDPQDSRRSADALLRLQAMAGISEEQGSDTDTPIVYNDRNLLD	Longhi, S., Receveur-Bréchot, V., Karlin, D., Johansson, K., Darbon, H., Bhella, D., Yeo, R., Finet, S., and Canard, B. (2003). The C-terminal domain of the measles virus nucleoprotein is intrinsically disordered and folds upon binding to the C-terminal moiety of the phosphoprotein. J. Biol. Chem. 278, 18638–18648.
ERM	39.6	0.7	MDGFYDQQVPFMVPGKSR	Lens, Z., Dewitte, F., Monté, D.,

			SEECRGRPVIDRKRKFLDT DLAHDSEELFQDLSQLQEA WLAEAQVPDDEQFVPDFQ SDNLVLHAPPPTKIKRELHS PSSELSSCSHEQALGANYGE KCLYNYCA	Baert, J.-L., Bompard, C., Sénéchal, M., Van Lint, C., de Launoit, Y., Villeret, V., and Verger, A. (2010). Solution structure of the N-terminal transactivation domain of ERM modified by SUMO-1. Biochem. Biophys. Res. Commun. 399, 104–110.
Neuroligin-3	33	3	YRKDKRRQEPLRQPSPQRG AGAPELGAAPEEELAALQL GPTHHECEAGPPHDTLRLT ALPDYTLTLRRSPDDIPLMT PNTITMIPNSLVGLQTLHPY NTFAAGFNSTGLPHSHSTT RV	Paz, A., Zeev-Ben-Mordehai, T., Lundqvist, M., Sherman, E., Mylonas, E., Weiner, L., Haran, G., Svergun, D.I., Mulder, F.A.A., Sussman, J.L., et al. (2008). Biophysical characterization of the unstructured cytoplasmic domain of the human neuronal adhesion protein neuroligin 3. Biophys. J. 95, 1928–1944.
Prothymosin alpha	37.8	0.9	MSDAAVDTSSEITTKDLKE KKEVVEEAENGRDAPANG NAENEENGEQEADNEVD EEEEEGGEEEEEEEEGDG EEEDGDEDEEAESATGKR AAEDDEDDDVDTKKQKT DEDD	Uversky, V.N., Gillespie, J.R., Millett, I.S., Khodyakova, A.V., Vasiliev, A.M., Chernovskaya, T.V., Vasilenko, R.N., Kozlovskaya, G.D., Dolgikh, D.A., Fink, A.L., et al. (1999). Natively Unfolded Human Prothymosin α Adopts Partially Folded Collapsed Conformation at Acidic pH. Biochemistry 38, 15009–15016.
Fez1	36	1	QIQEEEETLQDEEVWDAL TDNYIPSLSEDWRDPNIEAL	Alborghetti, M.R., Furlan, A.S., Silva, J.C., Paes Leme, A.F., Torriani, I.C.L.,

			NGNCSDTEIHEKEEEEFNEKSENDSGINEEPLLTADQVIEEIEEMMQNSPDPEEEEEVLEEEDGG	and Kobarg, J. (2010). Human FEZ1 Protein Forms a Disulfide Bond Mediated Dimer: Implications for Cargo Transport. J. Proteome Res. 9, 4595–4603.
HIV-TAT	33	1.05	MEPVDPRLEPWKHPGSQPRTACTNCYCKKCCFHCQVCFIRKALGISYGRKKRRQRRRAPQDSETHQVSPPKQPASQPRGDPTGPKESKKKVERETETHPVN	Foucault, M., Mayol, K., Receveur-Bréchot, V., Bussat, M.-C., Klinguer-Hamour, C., Verrier, B., Beck, A., Haser, R., Gouet, P., and Guillon, C. (2010). UV and X-ray structural studies of a 101-residue long Tat protein from a HIV-1 primary isolate and of its mutated, detoxified, vaccine candidate. Proteins 78, 1441–1456.
p531-91	28.7	0.3	MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGPDEAPRMPEAAPPVAPAPAAPTPAAPAPAPSW	Wells, M., Tidow, H., Rutherford, T.J., Markwick, P., Jensen, M.R., Mylonas, E., Svergun, D.I., Blackledge, M., and Fersht, A.R. (2008). Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. Proc. Natl. Acad. Sci. U. S. A. 105, 5762–5767.
Tau - ht40	65	3	MAEPRQEFEVMEDHAGTYGLGDRKDQGGYTMHQDQEGDTDAGLKESPLQTPTEDGSEEPGSETSDAKSTPTAEDVTAPLVDEGAPGKQAAAQPHTEIPEGTTAEEAGIGDTP	E. Mylonas, A. Hascher, P. Bernado´, M. Blackledge, E. Mandelkow and D. I. Svergun, Biochemistry, 2008, 47, 10345–10353.

			SLEDEAAGHVTQARMVSKS KDGTGSDDKKAKGADGK TKIATPRGAAPPGQKGQAN ATRIPAKTPPAPKTPPSSGEP PKSGDRSGYSSPGSPGTPGS RSRTPSLPTPPTREPKKVAV VRTPPKSPSSAKSRLQTAPV PMPDLKNVKSKIGSTENLK HQPGGGKVQIINKKLDLSN VQSKCGSKDNIKHVPGGGS VQIVYKPVDLSKVTSKCGSL GNIHHKPGGGQVEVKSEK LDFKDRVQSKIGSLDNITH VPGGGNKKIETHKLTFRE NAKAKTDHGAEIVYKSPVV SGDTSPRHLSNVSSTGSIDM VDSPQLATLADEVSASLAK QGL	
Tau - K32	42	3	SSPGSPGTPGSRSRTPSLPTP PTREPKKVAVVRTPPKSPSS AKSRLQTAPVPMPDLKNVK SKIGSTENLKHQPGGGKV QIINKKLDLSNVQSKCGSK DNIKHVPGGGSVQIVYKPV DLSKVTSKCGSLGNIHHKP GGGQVEVKSEKLDFKDRV QSKIGSLDNITHVPGGGNK KIETHKLTFRENAKAKTDH GAEIVY	E. Mylonas, A. Hascher, P. Bernado´, M. Blackledge, E. Mandelkow and D. I. Svergun, Biochemistry, 2008, 47, 10345–10353.
Tau - K16	39	3	SSPGSPGTPGSRSRTPSLPTP PTREPKKVAVVRTPPKSPSS	E. Mylonas, A. Hascher, P. Bernado´, M. Blackledge, E. Mandelkow and D.

			AKSRLQTAPVPMPDLKNVK SKIGSTENLKHQPGGGKV QIINKKLDLSNVQSKCGSK DNIKHVPGGGSVQIVYKPV DLSKVTSKCGSLGNIHHKP GGGQVEVKSEKLDFKDRV QSKIGSLDNITHVPGGGNK KIE	I. Svergun, Biochemistry, 2008, 47, 10345–10353.
Tau - K18	38	3	QTAPVPMPDLKNVKSKIGS TENLKHQPGGGKVQIINK KLDLSNVQSKCGSKDNIKH VPGGGSVQIVYKPVDLSKV TSKCGSLGNIHHKPGGGQ VEVKSEKLDFKDRVQSKIG SLDNITHVPGGGNKKIE	E. Mylonas, A. Hascher, P. Bernado´, M. Blackledge, E. Mandelkow and D. I. Svergun, Biochemistry, 2008, 47, 10345–10353.
Tau - ht23	53	3	MAEPRQEFEVMEDHAGTY GLGDRKDQGGYTMHQDQ EGDTDAGLKAEEAGIGDT PSLEDEAAGHVTQARMVSK SKDGTGSDDKKAKGADG KTKIATPRGAAPPGQKGQA NATRIPAKTPPAPKTPPSSG EPPKSGDRSGYSSPGSPGTP GSRSRTPSLPTPPTREPKKV AVVRTPPKSPSSAKSRLQTA PVPMPDLKNVKSKIGSTEN LKHQPGGGKVQIVYKPVD LSKVTSKCGSLGNIHHKPG GGQVEVKSEKLDFKDRVQ SKIGSLDNITHVPGGGNKK IETHKLTFRENAKAKTDHG	E. Mylonas, A. Hascher, P. Bernado´, M. Blackledge, E. Mandelkow and D. I. Svergun, Biochemistry, 2008, 47, 10345–10353.

			AEIVYKSPVVSGDTSPRHLS NVSSTGSIDMVDSPQLATLA DEVSASLAKQGL	
Tau - K27	37	2	SSPGSPGTPGSRSRTPSLPTP PTREPKKVAVVRTPPKSPSS AKSRLQTAPVPMPDLKNVK SKIGSTENLKHQPGGGSVQ IVYKPVDLSKVTSKCGSLG NIHHKPGGGQVEVKSEKL DFKDRVQSKIGSLDNITHV PGGGNKKIETHKLTFREN AKAKTDHGAEIVY	E. Mylonas, A. Hascher, P. Bernado´, M. Blackledge, E. Mandelkow and D. I. Svergun, Biochemistry, 2008, 47, 10345–10353.
Tau - K17	36	2	SSPGSPGTPGSRSRTPSLPTP PTREPKKVAVVRTPPKSPSS AKSRLQTAPVPMPDLKNVK SKIGSTENLKHQPGGGSVQ IVYKPVDLSKVTSKCGSLG NIHHKPGGGQVEVKSEKL DFKDRVQSKIGSLDNITHV PGGGNKKIE	E. Mylonas, A. Hascher, P. Bernado´, M. Blackledge, E. Mandelkow and D. I. Svergun, Biochemistry, 2008, 47, 10345–10353.
Tau - K19	35	1	QTAPVPMPDLKNVKSKIGS TENLKHQPGGGSVQIVYKP VDLSKVTSKCGSLGNIHHK PGGGQVEVKSEKLDFKDR VQSKIGSLDNITHVPGGGN KKIE	E. Mylonas, A. Hascher, P. Bernado´, M. Blackledge, E. Mandelkow and D. I. Svergun, Biochemistry, 2008, 47, 10345–10353.
Tau - K44	52	2	MAEPRQEFEVMEDHAGTY	E. Mylonas, A. Hascher, P. Bernado´,

			GLGDRKDQGGYTMHQDQ EGDTDAGLKAEEAGIGDT PSLEDEAAGHVTQARMVSK SKDGTGSDDKKAKGADG KTKIATPRGAAPPGQKGQA NATRIPAKTPPAPKTPPSSG EPPKSGDRSGYSSPGSPGTP GSRSRTPSLPTPPTREPKKV AVVRTPPKSPSSAKSRLQTA PVPMPDLKNVKSKIGSTEN LKHQPGGGKVQIVYKPVD LSKVTSKCGSLGNIHHKPG GGQVEVKSEKLDFKDRVQ SKIGSLDNITHVPGGGNKK IE	M. Blackledge, E. Mandelkow and D. I. Svergun, Biochemistry, 2008, 47, 10345–10353.
Tau - K10	40	1	QTAPVPMPDLKNVKSKIGS TENLKHQPGGGSVQIVYKP VDLSKVTSKCGSLGNIHHK PGGGQVEVKSEKLDFKDR VQSKIGSLDNITHVPGGGN KKIETHKLTFRENAKAKTD HGAEIVYKSPVVSGDTSPR HLSNVSSTGSIDMVDSPQLA TLADEVSASLAKQGL	E. Mylonas, A. Hascher, P. Bernado´, M. Blackledge, E. Mandelkow and D. I. Svergun, Biochemistry, 2008, 47, 10345–10353.
Tau - K25	41	2	MAEPRQEFEVMEDHAGTY GLGDRKDQGGYTMHQDQ EGDTDAGLKAEEAGIGDT PSLEDEAAGHVTQARMVSK SKDGTGSDDKKAKGADG KTKIATPRGAAPPGQKGQA NATRIPAKTPPAPKTPPSSG	E. Mylonas, A. Hascher, P. Bernado´, M. Blackledge, E. Mandelkow and D. I. Svergun, Biochemistry, 2008, 47, 10345–10353.

			EPPKSGDRSGYSSPGSPGTP GSRSRTPSLPTPPTREPKKV AVVRTPPKSPSSAKSRL	
Tau - K23	49	2	MAEPRQEFEVMEDHAGTY GLGDRKDQGGYTMHQDQ EGDTDAGLKAEEAGIGDT PSLEDEAAGHVTQARMVSK SKDGTGSDDKKAKGADG KTKIATPRGAAPPGQKGQA NATRIPAKTPPAPKTPPSSG EPPKSGDRSGYSSPGSPGTP GSRSRTPSLPTPPTREPKKV AVVRTPPKSPSSAKSRLKKIE THKLTFRENAKAKTDHGA EIVYKSPVVSGDTSPRHLSN VSSTGSIDMVDSPQLATLAD EVSASLAKQGL	E. Mylonas, A. Hascher, P. Bernado´, M. Blackledge, E. Mandelkow and D. I. Svergun, Biochemistry, 2008, 47, 10345–10353.
Tau - K32 AT8 AT100	41	3	SEPGEPGEPGSRSREPELPT PPTREPKKVAVVRTPPKSPS SAKSRLQTAPVPMPDLKNV KSKIGSTENLKHQPGGGK VQIINKKLDLSNVQSKCGS KDNIKHVPGGGSVQIVYKP VDLSKVTSKCGSLGNIHHK PGGGQVEVKSEKLDFKDR VQSKIGSLDNITHVPGGGN KKIETHKLTFRENAKAKTD HGAEIVY	E. Mylonas, A. Hascher, P. Bernado´, M. Blackledge, E. Mandelkow and D. I. Svergun, Biochemistry, 2008, 47, 10345–10353.
Tau - ht23	54	3	MAEPRQEFEVMEDHAGTY	E. Mylonas, A. Hascher, P. Bernado´,

S214E			GLGDRKDQGGYTMHQDQ EGDTDAGLKAEEAGIGDT PSLEDEAAGHVTQARMVSK SKDGTGSDDKKAKGADG KTKIATPRGAAPPGQKGQA NATRIPAKTPPAPKTPPSSG EPPKSGDRSGYSSPGSPGTP GSRSRTPELPTPPTREPKKV AVVRTPPKSPSSAKSRLQTA PVPMPDLKNVKSKIGSTEN LKHQPGGGKVQIVYKPVD LSKVTSKCGSLGNIHHKPG GGQVEVKSEKLDFKDRVQ SKIGSLDNITHVPGGGNKK IETHKLTFRENAKAKTDHG AEIVYKSPVVSGDTSPRHLS NVSSTGSIDMVDSPQLATLA DEVSASLAKQGL	M. Blackledge, E. Mandelkow and D. I. Svergun, Biochemistry, 2008, 47, 10345–10353.
Tau - ht23 AT8 AT100	52	3	MAEPRQEFEVMEDHAGTY GLGDRKDQGGYTMHQDQ EGDTDAGLKAEEAGIGDT PSLEDEAAGHVTQARMVSK SKDGTGSDDKKAKGADG KTKIATPRGAAPPGQKGQA NATRIPAKTPPAPKTPPSSG EPPKSGDRSGYSEPGEPGE PGSRSREPELPTPPTREPKK VAVVRTPPKSPSSAKSRLQT APVPMPDLKNVKSKIGSTE NLKHQPGGGKVQIVYKPV DLSKVTSKCGSLGNIHHKP GGGQVEVKSEKLDFKDRV	E. Mylonas, A. Hascher, P. Bernado´, M. Blackledge, E. Mandelkow and D. I. Svergun, Biochemistry, 2008, 47, 10345–10353.

			QSKIGSLDNITHVPGGGNK KIETHKLTFRENAKAKTDH GAEIVYKSPVVSGDTSPRHL SNVSSTGSIDMVDSPQLATL ADEVSASLAKQGL	
Tau - K18 P301L	35	2	QTAPVPMPDLKNVKSKIGS TENLKHQPGGGKVQIINK KLDLSNVQSKCGSKDNIKH VLGGGSVQIVYKPVDLSKV TSKCGSLGNIHHKPGGGQ VEVKSEKLDFKDRVQSKIG SLDNITHVPGGGNKKIE	E. Mylonas, A. Hascher, P. Bernado´, M. Blackledge, E. Mandelkow and D. I. Svergun, Biochemistry, 2008, 47, 10345–10353.
Tau - K18 ΔK280	79	10	QTAPVPMPDLKNVKSKIGS TENLKHQPGGGKVQIINK LDLSNVQSKCGSKDNIKHV LGGGSVQIVYKPVDLSKVT SKCGSLGNIHHKPGGGQV EVKSEKLDFKDRVQSKIGS LDNITHVPGGGNKKIE	E. Mylonas, A. Hascher, P. Bernado´, M. Blackledge, E. Mandelkow and D. I. Svergun, Biochemistry, 2008, 47, 10345–10353.
Tau - K18 ΔK280 I277P I308P	35	2	QTAPVPMPDLKNVKSKIGS TENLKHQPGGGKVQPINK LDLSNVQSKCGSKDNIKHV LGGGSVQPVYKPVDLSKVT SKCGSLGNIHHKPGGGQV EVKSEKLDFKDRVQSKIGS LDNITHVPGGGNKKIE	E. Mylonas, A. Hascher, P. Bernado´, M. Blackledge, E. Mandelkow and D. I. Svergun, Biochemistry, 2008, 47, 10345–10353.
Histatin	13.2	0.01	DSHAKRHHGYKRKFHEKH	Cragnell, C., Durand, D., Cabane, B.,

			HSHRGY	and Skepö, M. (2016). Coarse-grained modeling of the intrinsically disordered protein Histatin 5 in solution: Monte Carlo simulations in combination with SAXS. Proteins 84, 777–791.
CortactinCR	46.7		GPLGSGYGGKFGVEQDRM DKSAVGHEYQSKLSKHCSQ VDSVRGFGGKFGVQMDRV DQSAVGFEYQGKTEKHAS QKDYSSGFGGKYGVQADR VDKSAVGFDYQGKTEKHE SQRDYSKGFGGKYGIDKD KVDKSAVGFEYQGKTEKH ESQKDYVKGFGGKFGVQT DRQDKCALGWDHQEKLQ LHESQKDYKTGFGGKFGV QSERQDSAAVGFDYKEKL AKHESQQDYSKGFGGKYG VQKDRMDKNASTFEDVTQ VSSAYQKTVPVEAVTSKTSN IRANFENLAKEKEQEDRRK AEAERAQRMAKERQEQEE ARRKLEEQARAKTQT	Li, X., Tao, Y., Murphy, J.W., Scherer, A.N., Lam, T.T., Marshall, A.G., Koleske, A.J., and Boggon, T.J. (2017). The repeat region of cortactin is intrinsically disordered in solution. Sci. Rep. 7, 16696.
Pertactin-NTD	51.3	0.1	DWNNQSIVKTGERQHGIHI QGSDPGGVRTASGTTIKVS GRQAQGILLENPAAELQFR NGSVTSSGQLSDDGIRRFLG TVTVKAGKLVADHATLAN VGDTWDDDGIALYVAGEQ AQASIADSTLQGAGGVQIE	Riback, J.A., Bowman, M.A., Zmyslowski, A.M., Knoverek, C.R., Jumper, J.M., Hinshaw, J.R., Kaye, E.B., Freed, K.F., Clark, P.L., and Sosnick, T.R. (2017). Innovative scattering analysis shows that hydrophobic disordered proteins are

			RGANVTVQRSAIVDGGLHI GALQSLQPEDLPPSRVVLR DTNVTAVPASGAPAAVSVL GASELTLDGGHITGGRAAG VAAMQGAVVHLQRATIRR GEALAGGAVPGGAVPGGA VPGGFGPGGFGPVLDGWY GVDVSGSSVELAQSIVEAPE LGAAIRVGRGARVTVPGGS LSAPHGNVIETGGARRFAP QAAPLSITLQAGAH	expanded in water. Science 358, 238–241.
Reduced_RnaseH	33.6	0.1	KETAAAKFERQHMDSSTSA ASSSNYCNQMMKSRNLTKD RCKPVNTFVHESLADVQAV CSQKNVACKNGQTNCYQS YSTMSITDCRETGSSKYPNC AYKTTQANKHIIVACEGNP YVPVHFDASV	Riback, J.A., Bowman, M.A., Zmyslowski, A.M., Knoverek, C.R., Jumper, J.M., Hinshaw, J.R., Kaye, E.B., Freed, K.F., Clark, P.L., and Sosnick, T.R. (2017). Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water. Science 358, 238–241.
Nup1573_frag	24	5	GCPSASPAFGANQTPTFGQ SQGASQPNPPGFSISSSTALF PTGSQPAPPTFGTVSSSSQPP VFGQQPSQSAFGSTTPNA	Mercadante, D., Milles, S., Fuertes, G., Svergun, D.I., Lemke, E.A., and Gräter, F. (2015). Kirkwood-Buff Approach Rescues Overcollapse of a Disordered Protein in Canonical Protein Force Fields. J. Phys. Chem. B 119, 7975–7984.
LOX-PP	37	0.4	APPAAGQQQPPREPPAAPG AWRQQIQWENNGQVFSLL	Vallet, S.D., Miele, A.E., Uciechowska-Kaczmarzyk, U., Liwo,

			SLGSQYQPQRRRDPGAAVP GAANASAQQPRTPILLIRDN RTAAARTRTAGSSGVTAGR PRPTARHWFQAGYSTSRAR EAGASRAENQTAPGEVPAL SNLRPPSRVDGMVG	A., Duclos, B., Samsonov, S.A., and Ricard-Blum, S. (2018). Insights into the structure and dynamics of lysyl oxidase propeptide, a flexible protein with numerous partners. Sci. Rep. 8, 11768.
H1_CTD	25	0.2	KGDEPKRSVAFKKTKKEV KKVATPKKAAKPKKAASK APSKKPKATPVKKAKKKPA ATPKKAKKPKVVKVKPVK ASKPKKAKTVKPKAKSSAK RASKKK	Roque, A., Ponte, I., and Suau, P. (2007). Macromolecular crowding induces a molten globule state in the C-terminal domain of histone H1. Biophys. J. 93, 2170–2177.
p27_WT (v31)	28.1	1.8	GSHMKGACKVPAQESQDV SGSRPAAPLIGAPANSEDTH LVDPKTDPSDSQTGLAEQC AGIRKRPATDDSSTQNKRA NRTEENVSDGSPNAGSVEQ TPKKPGLRRRQT	Das, R.K., Huang, Y., Phillips, A.H., Kriwacki, R.W., and Pappu, R.V. (2016). Cryptic sequence features within the disordered protein p27Kip1 regulate cell cycle signaling. Proc. Natl. Acad. Sci. U. S. A. 113, 5616–5621.
p27_v14	29.4	1.3	GSHMKGACKSSSPPSNDQG RPGDPKQVIDKTEVERTQ DTSNIQETQSANNSGPDKP SRCDLAVSGVAAAALPAPG HANSTARDLTRDEEAGSVE QTPKKPGLRRRQT	Das, R.K., Huang, Y., Phillips, A.H., Kriwacki, R.W., and Pappu, R.V. (2016). Cryptic sequence features within the disordered protein p27Kip1 regulate cell cycle signaling. Proc. Natl. Acad. Sci. U. S. A. 113, 5616–5621.
p27_v15	29.2	1	GSHMKGACIVANSPPDDVK	Das, R.K., Huang, Y., Phillips, A.H.,

			SKEDVPQTDPRLTGGDRD NARASRTGNDPAGASTQSA EVACSNPILSTPDAQEKQA GTSNSKERPHEQLSAGSVE QTPKKPGLRRRQT	Kriwacki, R.W., and Pappu, R.V. (2016). Cryptic sequence features within the disordered protein p27Kip1 regulate cell cycle signaling. Proc. Natl. Acad. Sci. U. S. A. 113, 5616–5621.
p27_v44	24.9	1.3	GSHMKGACRKPANAEADS SSCQNVPRGKSKQAPETPT GSPLGDATLNQVKPRRPSS ASTNIGQLEDADEDDAED HVGSAVTSQTIPNDRAGSV EQTPKKPGLRRRQT	Das, R.K., Huang, Y., Phillips, A.H., Kriwacki, R.W., and Pappu, R.V. (2016). Cryptic sequence features within the disordered protein p27Kip1 regulate cell cycle signaling. Proc. Natl. Acad. Sci. U. S. A. 113, 5616–5621.
p27_v56	23.3	1	GSHMKGACGSSVLGTGNP RNQAHVSDTSLEEDDDEQ DDSTPDEVSQACTIVASALD INAATPRSPKASPKRKRKRQ STAPAQGNEPPGNAGSVEQ TPKKPGLRRRQT	Das, R.K., Huang, Y., Phillips, A.H., Kriwacki, R.W., and Pappu, R.V. (2016). Cryptic sequence features within the disordered protein p27Kip1 regulate cell cycle signaling. Proc. Natl. Acad. Sci. U. S. A. 113, 5616–5621.
p27_v78	22.1	0.3	GSHMKGACALPSGVVPAE DDDDDEEEEDDQDPAQP QAVQGAAPSSGTNNSQPIL PSIAVNSTTGPNSTAGKKKR KRRRTRHSNCATLSSAGSVE QTPKKPGLRRRQT	Das, R.K., Huang, Y., Phillips, A.H., Kriwacki, R.W., and Pappu, R.V. (2016). Cryptic sequence features within the disordered protein p27Kip1 regulate cell cycle signaling. Proc. Natl. Acad. Sci. U. S. A. 113, 5616–5621.

Ki-1/57	47	2	PRRGEQQGWNDSRGPEGM LERAERRSYREYRPYETERQ ADFTAEKFPDEKPGDRFDR DRPLRGRGGPRGGMRGRG RGGPGNRVFDAFDQRGKR EFERYGGNDKIAVRTEDN MGGCGVRTWGSGKDTSDV EPTAPMEEPTVVEESQGTP EEESPAKVPELEVEEETQV QEMTLDEWKNLQEQTRPK PEFNIRKPESTVPSKAVVIH KSKYRDDMVKDDYEDDSH VFRKPANDITSQLEINFGNL PRPGRGARGGTRGGRGRIR RAENYGPRAEVVMQDVAP NPDDPEDFPALS	Bressan, G.C., Silva, J.C., Borges, J.C., Dos Passos, D.O., Ramos, C.H.I., Torriani, I.L., and Kobarg, J. (2008). Human regulatory protein Ki-1/57 has characteristics of an intrinsically unstructured protein. J. Proteome Res. 7, 4465–4474.
CTCF-R domain (WT)	32.5	1.8	SAERRNSILTETLHRFSLEG DAPVSWTETKKQSFKQTG EFGEKRKNSILNPINSIRKFS IVQKTPLQMNGIEEDSDEP LERRLSLVPDSEQGEAILPRI SVISTGPTLQARRRQSVLNL MTHSVNQGQNIHRKTTAST RKVSLAPQANLTELDIYSRR LSQETGLEISEEINEEDLKE CFFDDME	Marasini, C., Galeno, L., and Moran, O. (2013). A SAXS-based ensemble model of the native and phosphorylated regulatory domain of the CFTR. Cell. Mol. Life Sci. 70, 923–933.
CTCF-R domain (phosphorylate	29.2	0.4	SAERRNSILTETLHRFSLEG DAPVSWTETKKQSFKQTG EFGEKRKNSILNPINSIRKFS IVQKTPLQMNGIEEDSDEP	Marasini, C., Galeno, L., and Moran, O. (2013). A SAXS-based ensemble model of the native and phosphorylated regulatory domain of

d)			LERRLSLVPDSEQGEAILPRISVISTGPTLQARRRQSVLNLMTHSVNQGQNIHRKTTASTRKVSLAPQANLTELDIYSRRLSQETGLEISEEINEEDLKECFFDDME	the CFTR. Cell. Mol. Life Sci. 70, 923–933.
hNHE1cdt	37.5	0	VPAHKLDSPTMSRARIGSDPLAYEPKEDLPVITIDPASPQSPESVDLVNEELKGKVLGLSRDPAKVAEEDEDDDGGIMMRSKETSSPGTDDVFTPAPSDSPSSQRIQRCLSDPGPHPEPGEGEPFFPKGQ	Kjaergaard, M., Nørholm, A.-B., Hendus-Altenburger, R., Pedersen, S.F., Poulsen, F.M., and Kragelund, B.B. (2010). Temperature-dependent structural changes in intrinsically disordered proteins: Formation of α-helices or loss of polyproline II? Protein Sci. 19, 1555–1564.
pMBP	54	0	ASQKRPSQRHGSKYLASASTMDHARHGFLPRHRDTGIDSLGRFFGADRGAPKRGSGKDGHHAARTTHYGSLPQKAQHGRPQDENPVVHFFKNIVTPRTPPPSQGKGRGLSLSRFSWGAEGQKPGFGYGGRAPDYKPAHKGLKGAQDAQGTLSKIFKLGGRDSRSGSPMARR	Majava, V., Wang, C., Myllykoski, M., Kangas, S.M., Kang, S.U., Hayashi, N., Baumgärtel, P., Heape, A.M., Lubec, G., and Kursula, P. (2010). Structural analysis of the complex between calmodulin and full-length myelin basic protein, an intrinsically disordered molecule. Amino Acids 39, 59–71.
HMPV	27.4	0.5	MSFPEGKDILFMGNEAAKLAEAFQKSLRKPSHKRSQSIIGEKVNTVSETLELPTISRPTKP	Renner, M., Paesen, G.C., Grison, C.M., Granier, S., Grimes, J.M., and Leyrat, C. (2017). Structural dissection of human metapneumovirus phosphoprotein using small angle x-

				ray scattering. Sci. Rep. 7, 14865.
redAFP	22.2	0.1	CKGADGAHGVNGCPGTAGAAGSVGGPGCDGGHGGNGGNGNPGCAGGVGGAGGASGGTGVGGRGGKGGSGTPKGADGAPGAP	Gates, Z.P., Baxa, M.C., Yu, W., Riback, J.A., Li, H., Roux, B., Kent, S.B.H., and Sosnick, T.R. (2017). Perplexing cooperative folding and stability of a low-sequence complexity, polyproline 2 protein lacking a hydrophobic core. Proc. Natl. Acad. Sci. U. S. A. 114, 2241–2246.
CSD1 (with overhang)	35.4	0	MAMITNSSSVPAESKSSKPSGKSDMDAALDDLIDTLGGPEETEEDNTTYTGPEVLDPMSSTYIEELGKREVTLPPKYRELLDKKEGIPVPPPDTSKPLGPDDAIDALSLDLTCSSPTADGKKTEKEKSTGEVLKAQSVGVIKSDPLESLN	Konno, T., Tanaka, N., Kataoka, M., Takano, E., and Maki, M. (1997). A circular dichroism study of preferential hydration and alcohol effects on a denatured protein, pig calpastatin domain I. Biochim. Biophys. Acta 1342, 73–82.
PAGE4_WT	36.2	1.1	MSARVRSRSRGRGDGQEAPDVVAFVAPGESQQEEPPTDNQDIEPGQEREGTPPIEERKVEGDCQEMDLEKTRSERGDGSDVKEKTPPNPKHAKTKEAGDGQP	Kulkarni, P., Jolly, M.K., Jia, D., Mooney, S.M., Bhargava, A., Kagohara, L.T., Chen, Y., Hao, P., He, Y., Veltri, R.W., et al. (2017). Phosphorylation-induced conformational dynamics in an intrinsically disordered protein and potential role in phenotypic heterogeneity. Proc. Natl. Acad. Sci. U. S. A. 114, E2644–E2653.

PAGE4_WT_ phosphorylated	49.8	1.9	MSARVRSRSRGRGDGQEAP DVVAFVAPGESQQEEPPTD NQDIEPGQEREGTPPIEER KVEGDCQEMDLEKTRSER GDGSDVKEKTPPNPKHAK TKEAGDGQP	Kulkarni, P., Jolly, M.K., Jia, D., Mooney, S.M., Bhargava, A., Kagohara, L.T., Chen, Y., Hao, P., He, Y., Veltri, R.W., et al. (2017). Phosphorylation-induced conformational dynamics in an intrinsically disordered protein and potential role in phenotypic heterogeneity. Proc. Natl. Acad. Sci. U. S. A. 114, E2644–E2653.
ERalpha-NTD	31	0.2	SNAMTMTLHTKASGMALL HQIQGNELEPLNRPQLKIP LERPLGEVYLDSSKPAVYN YPEGAAYEFNAAAAANAQ VYGQTGLPYGPGSEAAAFG SNGLGGFPPLNSVSPSPLML LHPPPQLSPFLQPHGQQVP YYLENEPSGYTVREAGPPA FYRPNSDNRRQGGRERLAS TNDKGSMAMESAKETRY	Peng, Y., Cao, S., Kiselar, J., Xiao, X., Du, Z., Hsieh, A., Ko, S., Chen, Y., Agrawal, P., Zheng, W., Shi, W., Jiang, W., Yang, L., Chance, M. R., Surewicz, W. K., Buck, M., & Yang, S. (2019). A Metastable Contact and Structural Disorder in the Estrogen Receptor Transactivation Domain. Structure , 27(2), 229–240.e4.
A1-LCD-NLS	27.6	0.16	GSMASASSSQRGRSGSGNF GGGRGGGFGGNDNFGRG GNFSGRGGFGGSRGGGGY GGSGDGYNGFGNDGSNF GGGGSYNDFGNYNNQSSN FGPMKGGNFGGRSSGGSG GGGQYFAKPRNQGGYGGS SSSSSYGSGRRF	Bremer, A., Farag, M., Borcherds, W. M., Peran, I., Martin, E. W., Pappu, R. V., & Mittag, T. (2022). Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. Nature Chemistry, 14(2), 196–207.

A1-LCD+NLS	25.83	0.11	GSMASASSSQRGRSGSGNF GGGRGGGFGGNDNFGRG GNFSGRGGFGGSRGGGGY GGSGDGYNGFGNDGSNF GGGGSYNDFGNYNNQSSN FGPMKGGNFGGRSSGPYG GGGQYFAKPRNQGGYGGS SSSSSYGSGRRF	Bremer, A., Farag, M., Borcherds, W. M., Peran, I., Martin, E. W., Pappu, R. V., & Mittag, T. (2022). Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. Nature Chemistry, 14(2), 196–207.
A1-LCD-12F+12Y	26.04	0.2	GSMASASSSQRGRSGSGNY GGGRGGGYGGNDNYGRG GNYSGRGGYGGSRGGGGY GGSGDGYNGYGNDGSNY GGGGSYNDYGNYNNQSSN YGPMKGGNYGGRSSGGSG GGGQYYAKPRNQGGYGG SSSSSSYGSGRRY	Bremer, A., Farag, M., Borcherds, W. M., Peran, I., Martin, E. W., Pappu, R. V., & Mittag, T. (2022). Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. Nature Chemistry, 14(2), 196–207.
A1-LCD+7F-7Y	27.18	0.13	GSMASASSSQRGRSGSGNF GGGRGGGFGGNDNFGRG GNFSGRGGFGGSRGGGGF GGSGDGFNGFGNDGSNFG GGGSFNDFGNFNNQSSNF GPMKGGNFGGRSSGGSGG GGQFFAKPRNQGGFGGSSS SSSFGSGRRF	Bremer, A., Farag, M., Borcherds, W. M., Peran, I., Martin, E. W., Pappu, R. V., & Mittag, T. (2022). Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. Nature Chemistry, 14(2), 196–207.
A1-LCD-9F+6Y	26.55	0.1	GSMASASSSQRGRSGSGNF GGGRGGGYGGNDNYGRG GNYSGRGGFGGSRGGGGY GGSGDGYNGGGNDGSNY	Bremer, A., Farag, M., Borcherds, W. M., Peran, I., Martin, E. W., Pappu, R. V., & Mittag, T. (2022). Deciphering how naturally occurring sequence

			GGGGSYNDSGNYNNQSSN FGPMKGGNYGGRSSGGSG GGGQYGAKPRNQGGYGG SSSSSSYGSGRRY	features impact the phase behaviours of disordered prion-like domains. Nature Chemistry, 14(2), 196–207.
A1-LCD-8F+4Y	27.07	0.07	GSMASASSSQRGRSGSGNF GGGRGGGYGGNDNGGRG GNYSGRGGFGGSRGGGGY GGSGDGYNGGGNDGSNY GGGGSYNDSGNYNNQSSN FGPMKGGNYGGRSSGGSG GGGQYGAKPRNQGGYGG SSSSSSYGSGRRF	Bremer, A., Farag, M., Borcherds, W. M., Peran, I., Martin, E. W., Pappu, R. V., & Mittag, T. (2022). Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. Nature Chemistry, 14(2), 196–207.
A1-LCD-9F+3Y	26.83	0.13	GSMASASSSQRGRSGSGNF GGGRGGGYGGNDNGGRG GNYSGRGGFGGSRGGGGY GGSGDGYNGGGNDGSNY GGGGSYNDSGNGNNQSSN FGPMKGGNYGGRSSGGSG GGGQYGAKPRNQGGYGG SSSSSSYGSGRRS	Bremer, A., Farag, M., Borcherds, W. M., Peran, I., Martin, E. W., Pappu, R. V., & Mittag, T. (2022). Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. Nature Chemistry, 14(2), 196–207.
A1-LCD-10R	26.71	0.07	GSMASASSSQGGSSGSGNF GGGGGGGFGGNDNFGGG GNFSGSGGFGGSGGGGGY GGSGDGYNGFGNDGSNF GGGGSYNDFGNYNNQSSN FGPMKGGNFGGSSSGPYG GGGQYFAKPGNQGGYGG SSSSSSYGSGGGF	Bremer, A., Farag, M., Borcherds, W. M., Peran, I., Martin, E. W., Pappu, R. V., & Mittag, T. (2022). Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. Nature Chemistry, 14(2), 196–207.

A1-LCD-6R	25.73	0.09	GSMASASSSQGGRSGSGNF GGGRGGGFGGNDNFGGG GNFSGSGGFGGSRGGGGY GGSGDGYNGFGNDGSNF GGGGSYNDFGNYNNQSSN FGPMKGGNFGGSSSGPYG GGGQYFAKPGNQGGYGG SSSSSSYGSGGRF	Bremer, A., Farag, M., Borcherds, W. M., Peran, I., Martin, E. W., Pappu, R. V., & Mittag, T. (2022). Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. Nature Chemistry, 14(2), 196–207.
A1-LCD+2R	26.23	0.23	GSMASASSSQRGRSGSGNF GGGRGGGFGGNDNFGRG GNFSGRGGFGGSRGGGGY GGSGDGYNGFRNDGSNFG GGGRYNDFGNYNNQSSNF GPMKGGNFGGRSSGPYGG GGQYFAKPRNQGGYGGSS SSSSYGSGRRF	Bremer, A., Farag, M., Borcherds, W. M., Peran, I., Martin, E. W., Pappu, R. V., & Mittag, T. (2022). Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. Nature Chemistry, 14(2), 196–207.
A1-LCD+7R	27.09	0.07	GSMASASSSQRGRSGRGNF GGGRGGGFGGNDNFGRG GNFSGRGGFGGSRGGGRY GGSGDRYNGFGNDGRNF GGGGSYNDFGNYNNQSSN FGPMKGGNFRGRSSGPYGR GGQYFAKPRNQGGYGGSS SSRSYGSGRRF	Bremer, A., Farag, M., Borcherds, W. M., Peran, I., Martin, E. W., Pappu, R. V., & Mittag, T. (2022). Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. Nature Chemistry, 14(2), 196–207.
A1-LCD-3R+3K	26.34	0.15	GSMASASSSQRGKSGSGNF GGGRGGGFGGNDNFGRG GNFSGRGGFGGSKGGGGY GGSGDGYNGFGNDGSNF	Bremer, A., Farag, M., Borcherds, W. M., Peran, I., Martin, E. W., Pappu, R. V., & Mittag, T. (2022). Deciphering how naturally occurring sequence

			GGGGSYNDFGNYNNQSSN FGPMKGGNFGGRSSGGSG GGGQYFAKPRNQGGYGGS SSSSSYGSGRKF	features impact the phase behaviours of disordered prion-like domains. Nature Chemistry, 14(2), 196–207.
A1-LCD-6R+6K	27.87	0.08	GSMASASSSQKGKSGSGNF GGGRGGGFGGNDNFGKG GNFSGRGGFGGSKGGGGY GGSGDGYNGFGNDGSNF GGGGSYNDFGNYNNQSSN FGPMKGGNFGGKSSGGSG GGGQYFAKPRNQGGYGGS SSSSSYGSGRKF	Bremer, A., Farag, M., Borcherds, W. M., Peran, I., Martin, E. W., Pappu, R. V., & Mittag, T. (2022). Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. Nature Chemistry, 14(2), 196–207.
A1-LCD-10R+10K	28.49	0.05	GSMASASSSQKGKSGSGNF GGGKGGGFGGNDNFGKG GNFSGKGGFGGSKGGGGY GGSGDGYNGFGNDGSNF GGGGSYNDFGNYNNQSSN FGPMKGGNFGGKSSGGSG GGGQYFAKPKNQGGYGG SSSSSSYGSGKKF	Bremer, A., Farag, M., Borcherds, W. M., Peran, I., Martin, E. W., Pappu, R. V., & Mittag, T. (2022). Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. Nature Chemistry, 14(2), 196–207.
A1-LCD-4D	26.42	0.12	GSMASASSSQRGRSGSGNF GGGRGGGFGGNGNFGRG GNFSGRGGFGGSRGGGGY GGSGGGYNGFGNSGSNFG GGGSYNGFGNYNNQSSNF GPMKGGNFGGRSSGPYGG GGQYFAKPRNQGGYGGSS SSSSYGSGRRF	Bremer, A., Farag, M., Borcherds, W. M., Peran, I., Martin, E. W., Pappu, R. V., & Mittag, T. (2022). Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. Nature Chemistry, 14(2), 196–207.

A1-LCD+4D	27.18	0.3	GSMASASSSQRDRSGSGNF GGGRGGGFGGNDNFGRG GNFSGRGDFGGSRGGGGY GGSGDGYNGFGNDGSNF GGGGSYNDFGNYNNQSSN FGPMKGGNFGGRSSDPYG GGGQYFAKPRNQGGYGGS SSSSSYDSGRRF	Bremer, A., Farag, M., Borcherds, W. M., Peran, I., Martin, E. W., Pappu, R. V., & Mittag, T. (2022). Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. Nature Chemistry, 14(2), 196–207.
A1-LCD+8D	26.85	0.07	GSMASASSSQRDRSGSGNF GGGRDGGFGGNDNFGRG DNFSGRGDFGGSRDGGGY GGSGDGYNGFGNDGSNF GGGGSYNDFGNYNNQSSN FGPMKGGNFGGRSSDPYG GGGQYFAKPRNQDGYGGS SSSSSYDSGRRF	Bremer, A., Farag, M., Borcherds, W. M., Peran, I., Martin, E. W., Pappu, R. V., & Mittag, T. (2022). Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. Nature Chemistry, 14(2), 196–207.
A1-LCD+12D	28.01	0.12	GSMASADSSQRDRDDSGNF GDGRGGGFGGNDNFGRG GNFSDRGGFGGSRGDGGY GGDGDGYNGFGNDGSNF GGGGSYNDFGNYNNQSSN FDPMKGGNFGDRSSGPYD GGGQYFAKPRNQGGYGGS SSSSSYGSDRRF	Bremer, A., Farag, M., Borcherds, W. M., Peran, I., Martin, E. W., Pappu, R. V., & Mittag, T. (2022). Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. Nature Chemistry, 14(2), 196–207.
A1-LCD+12E	28.52	0.05	GSMASAESSQREREESGNF GEGRGGGFGGNDNFGRG GNFSERGGFGGSRGEGGY GGEGDGYNGFGNDGSNF	Bremer, A., Farag, M., Borcherds, W. M., Peran, I., Martin, E. W., Pappu, R. V., & Mittag, T. (2022). Deciphering how naturally occurring sequence

			GGGGSYNDFGNYNNQSSNFEPMKGGNFGERSSGPYEGGGQYFAKPRNQGGYGGSSSSSSYGSERRF	features impact the phase behaviours of disordered prion-like domains. Nature Chemistry, 14(2), 196–207.
A1-LCD+7R+10D	29.21	0.08	GSMASADSSQRDRDGRGNFGDGRGGGFGGNDNFGRGGNFSDRGGFGGSRGGGRYGGDGDRYNGFGNDGRNFGGGGSYNDFGNYNNQSSNFDPMKGGNFRDRSSGPYDRGGQYFAKPRNQGGYGGSSSSRSYGSDRRF	Bremer, A., Farag, M., Borcherds, W. M., Peran, I., Martin, E. W., Pappu, R. V., & Mittag, T. (2022). Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. Nature Chemistry, 14(2), 196–207.
A1-LCD+7K+12Dblocky	25.62	0.14	GSMASAKSSQRDRDDDGNFGKGRGGGFGGNKNFGRGGNFSKRGGFGGSRGKGKYGGKGDDYNGFGNDGDNFGGGGSYNDFGNYNNQSSNFDPMDGGNFDDRSSGPYDDGGQYFADPRNQGGYGGSSSSKSYGSKRRF	Bremer, A., Farag, M., Borcherds, W. M., Peran, I., Martin, E. W., Pappu, R. V., & Mittag, T. (2022). Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. Nature Chemistry, 14(2), 196–207.
A1-LCD-12F+12Y10R	26.07	0.2	GSMASASSSQGGSSGSGNYGGGGGGGGYGGNDNYGGGGNYSGSGGYGGSGGGGGYGGSGDGYNGYGNDGSNYGGGGSYNDYGNYNNQSSNYGPMKGGNYGGSSSGPYGGGGQYYAKPGNQGGYGGSSSSSSYGSGGGY	Bremer, A., Farag, M., Borcherds, W. M., Peran, I., Martin, E. W., Pappu, R. V., & Mittag, T. (2022). Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. Nature Chemistry, 14(2), 196–207.

A1-LCD10F+7R+12D	28.6	0.04	GSMASADSSQRDRDDRGNFGDGRGGGGGGNDNFGRGGNGSDRGGGGGSRGDGRYGGDGDRYNGGGNDGRNGGGGGSYNDGGNYNNQSSNGDPMKGGNGRDRSSGPYDRGGQYGAKPRNQGGYGGSSSSRSYGSDRRG	Bremer, A., Farag, M., Borcherds, W. M., Peran, I., Martin, E. W., Pappu, R. V., & Mittag, T. (2022). Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. Nature Chemistry, 14(2), 196–207.
PNt	51.1	0.13	DWNNQSIVKTGERQHGIHIQGSDPGGVRTASGTTIKVSGRQAQGILLENPAAELQFRNGSVTSSGQLSDDGIRRFLGTVTVKAGKLVADHATLANVGDTWDDDGIALYVAGEQAQASIADSTLQGAGGVQIERGANVTVQRSAIVDGGLHIGALQSLQPEDLPPSRVVLRDTNVTAVPASGAPAAVSVLGASELTLDGGHITGGRAAGVAAMQGAVVHLQRATIRRGEALAGGAVPGGAVPGGAVPGGFGPGGFGPVLDGWYGVDVSGSSVELAQSIVEAPELGAAIRVGRGARVTVPGGSLSAPHGNVIETGGARRFAPQAAPLSITLQAGAH	Bowman, M. A., Riback, J. A., Rodriguez, A., Guo, H., Li, J., Sosnick, T. R., & Clark, P. L. (2020). Properties of protein unfolded states suggest broad selection for expanded conformational ensembles. Proceedings of the National Academy of Sciences, 117(38), 23356–23364.
Swap1	49.2	0.59	DWNNQSIVKTGERQHGIHIQGSDPGGVRTASGTTIKVSGRQAQGILLENPAAELQFRNGSVTSSGQKSDDGIRRFL	Bowman, M. A., Riback, J. A., Rodriguez, A., Guo, H., Li, J., Sosnick, T. R., & Clark, P. L. (2020). Properties of protein unfolded states

			GTVTVLAGKLVADHATLA NVGDTWDDDGIALYVAGE QAQASIADSTLQGAGGVQI ERGANVTVQRSAIVLGGLH IGALQSLQPEDDPPSRVVLR DTNVTAVPASGAPAAVSVL GASLLTLDGGHITGGRAAG VAAMQGAVVHEQRATIRR GEALAGGAVPGGAVPGGA VPGGFGPGGFGPVLDGWY GVDVSGSSVELAQSIVEAPE LGAAIRVGRGARVTVPGGS LSAPHGNVIETGGARRFAP QAAPLSITLQAGAH	suggest broad selection for expanded conformational ensembles. Proceedings of the National Academy of Sciences, 117(38), 23356–23364.
Swap3	40.58	1.07	DWNNQSIVKTGERQHGIHI QGSDPGGVRTASGTTIKVS GRQAQGILLENPAAELQFR NGSVTSSGQKSTDGTRRFL GDVIVKAGLLVADHATLA NVGDTWDDDGIALYVAGE QAQASIADSTLQGAGGVQI ERGANVDVLRLAIVDGGL HIGALQSQQPETSPPSRVVL RDTNVTAVPASGAPAAVSV QGASEQTLDGGAITGGRA AGVAAMLGHVVHLLRATIR RGEALAGGAVPGGAVPGG AVPGGFGPGGFGPVLDGW YGVDVSGSSVELAQSIVEAP ELGAAIRVGRGARVTVPGG SLSAPHGNVIETGGARRFAP QAAPLSITLQAGAH	Bowman, M. A., Riback, J. A., Rodriguez, A., Guo, H., Li, J., Sosnick, T. R., & Clark, P. L. (2020). Properties of protein unfolded states suggest broad selection for expanded conformational ensembles. Proceedings of the National Academy of Sciences, 117(38), 23356–23364.

Swap4	53.37	0.17	DWNNQSIVKTGERQHGIHI QGSDPGGVRTASGTTIKVS GRQAQGILLENPAAELQFR NGSVTSSGQLSFVGITRDLG RDTVKAGKLVADHATLAN VGDTWDDDGIALYVAGEQ AQASIADSTLQGAGGVQIE RGADVRVQREAIVDGGLH NGALQSLQPSILPPSTVVLR DTNVTAVPASGAPAAVLVS GASGLRLDGGHIHEGRAA GVAAMQGAVVTLQTATIRR GEALAGGAVPGGAVPGGA VPGGFGPGGFGPVLDGWY GVDVSGSSVELAQSIVEAPE LGAAIRVGRGARVTVPGGS LSAPHGNVIETGGARRFAP QAAPLSITLQAGAH	Bowman, M. A., Riback, J. A., Rodriguez, A., Guo, H., Li, J., Sosnick, T. R., & Clark, P. L. (2020). Properties of protein unfolded states suggest broad selection for expanded conformational ensembles. Proceedings of the National Academy of Sciences, 117(38), 23356–23364.
Swap4.1	54.45	0.14	DWNNQSIVKTGERQHGIHI QGSDPGGVRTASGTTIKVS GRQAQGILLENPAAELQFR NGSVTSSGQLSFVGITRRLG DDTVKAGKLVADHATLAN VGDTWDDDGIALYVAGEQ AQASIADSTLQGAGGVQIE RGADVEVQRRAIVDGGLH NGALQSLQPSILPPSTVVLR DTNVTAVPASGAPAAVLVS GASGLELDGGHIHRGRAA GVAAMQGAVVTLQTATIRR GEALAGGAVPGGAVPGGA	Bowman, M. A., Riback, J. A., Rodriguez, A., Guo, H., Li, J., Sosnick, T. R., & Clark, P. L. (2020). Properties of protein unfolded states suggest broad selection for expanded conformational ensembles. Proceedings of the National Academy of Sciences, 117(38), 23356–23364.

			VPGGFGPGGFGPVLDGWY GVDVSGSSVELAQSIVEAPE LGAAIRVGRGARVTVPGGS LSAPHGNVIETGGARRFAP QAAPLSITLQAGAH	
Swap5	48.71	0.34	DWNNQSIVKTGERQHGIHI QGSDPGGVRTASGTTIKVS GRQAQGILLENPAAELQFR NGSVTSSGQLSDDGIEDFL GTVTVDAGELVADHATLA NVGDTWDDDGIALYVAGE QAQASIADSTLQGAGGVQI EDGANVTVQESAIVDGGL HIGALQSLQPRRLPPSRVVL RKTNVTAVPASGAPAAVSV LGASKLTLRGGHITGGRAA GVAAMQGAVVHLQRATIR RGRALAGGAVPGGAVPGG AVPGGFGPGGFGPVLDGW YGVDVSGSSVELAQSIVEAP ELGAAIRVGRGARVTVPGG SLSAPHGNVIETGGARRFAP QAAPLSITLQAGAH	Bowman, M. A., Riback, J. A., Rodriguez, A., Guo, H., Li, J., Sosnick, T. R., & Clark, P. L. (2020). Properties of protein unfolded states suggest broad selection for expanded conformational ensembles. Proceedings of the National Academy of Sciences, 117(38), 23356–23364.
Swap6	52.61	0.27	DWNNQSIVKTGERQHGIHI QGSDPGGVRTASGTTIKVS GRQAQGILLENPAAELQFR NGSVTSSGQLSDRGIDRFLG TVTVEAGKLVADHATLAN VGDTWDKDGIALYVAGRQ AQASIADSTLQGAGGVQIR EGANVTVQRSAIVDGGLHI	Bowman, M. A., Riback, J. A., Rodriguez, A., Guo, H., Li, J., Sosnick, T. R., & Clark, P. L. (2020). Properties of protein unfolded states suggest broad selection for expanded conformational ensembles. Proceedings of the National Academy

			GALQSLQPERLPPSDVVLR DTNVTAVPASGAPAAVSVL GASRLTLDGGHITGGDAA GVAAMQGAVVHLQRATIE RGEALAGGAVPGGAVPGG AVPGGFGPGGFGPVLDGW YGVDVSGSSVELAQSIVEAP ELGAAIRVGRGARVTVPGG SLSAPHGNVIETGGARRFAP QAAPLSITLQAGAH	of Sciences, 117(38), 23356–23364.
sfAFP	23.1	2	CKGADGAHGVNGCPGTA GAAGSVGGPGCDGGHGG NGGNGNPGCAGGVGGAG GASGGTGVGGRGGKGGS GTPKGADGAPGAP	Gates ZP, Baxa MC, Yu W, Riback JA, Li H, Roux B, et al. Perplexing cooperative folding and stability of a low-sequence complexity, polyproline 2 protein lacking a hydrophobic core. Proc Natl Acad Sci U S A. 2017;114: 2241–2246.
FCP1	15.6	0.12	ESSRESSNEDEGSSSEADEM AKALEAELNDLM	Gibbs, Eric B., and Scott A. Showalter. 2016. "Quantification of Compactness and Local Order in the Ensemble of the Intrinsically Disordered Protein FCP1." The Journal of Physical Chemistry. B 120 (34): 8960–69.
RS-peptide	12.62	0.07	MYRSRSRSRSRSRSRSRS	**SAXS data – NMR data - Xiang, S., Gapsys, V., Kim, H.-Y., Bessonov, S., Hsiao, H.-H., Möhlmann, S., Klaukien, V., Ficner, R., Becker, S., Urlaub, H., Lührmann, R., de**

				Groot, B., & Zweckstetter, M. (2013). Phosphorylation drives a dynamic switch in serine/arginine-rich proteins. Structure , 21(12), 2162–2174.
P1_100	29	0	MAEEQARHVKNGLECIRALKAEPIGSLAIEEAMAAWSEISDNPGQERATCREEKAGSSGLSKPCLSAIGSTEGGAPRIRGQGPGESDDDAETLGIPPRNL	Naudi-Fabra, S., Tengo, M., Jensen, M. R., Blackledge, M., & Milles, S. (2021). Quantitative Description of Intrinsically Disordered Proteins Using Single-Molecule FRET, NMR, and SAXS. Journal of the American Chemical Society, 143(48), 20109–20121.
DSS1	25	0.1	MSRAALPSLENLEDDDEFEDFATENWPMKDTELDTGDDTLWENNWDDEDIGDDDFSVQLQAELKKKGVAAC	Pesce, F., Newcombe, E. A., Seiffert, P., Tranchant, E. E., Olsen, J. G., Grace, C. R., Kragelund, B. B., & Lindorff-Larsen, K. (2022). Assessment of models for calculating the hydrodynamic radius of intrinsically disordered proteins. Biophysical Journal. https://doi.org/10.1016/j.bpj.2022.12.013
GHR_ICD	59.59	0.38	SKQQRIKMLILPPVPVPKIKGIDPDLLKEGKLEEVNTILAIHDSYKPEFHSDDSWVEFIELDIDEPDEKTEESDTDRLLSSDHEKSHSNLGVKDGDSGRTSCCEPDILETDFNANDI	Pesce, F., Newcombe, E. A., Seiffert, P., Tranchant, E. E., Olsen, J. G., Grace, C. R., Kragelund, B. B., & Lindorff-Larsen, K. (2022). Assessment of models for calculating the hydrodynamic radius of

			HEGTSEVAQPQRLKGEADLLCLDQKNQNNSPYHDACPATQQPSVIQAEKNKPQPLPTEGAESTHQAAHIQLSNPSSLSNIDFYAQVSDITPAGSVVLSPGQKNKAGMSQCDMHPEMVSLCQENFLMDNAYFCEADAKKCIPVAPHIKVESHIQPSLNQEDIYITTESLTTAAGRPGTGEHVPGSEMPVPDYTSIHIVQSPQGLILNATALPLPDKEFLSSCGYVSTDQLNKIMP	intrinsically disordered proteins. Biophysical Journal. https://doi.org/10.1016/j.bpj.2022.12.013
NHE6cmdd	32	0.2	GPPLTTTLPACCGPIARCLTSPQAYENQEQLKDDDSDLILNDGDISLTYGDSTVNTEPATSSAPRRFMGNSSEDALDRELAFGDHELVIRGTRLVLPMDDSEPPLNLLDNTRHGPA	Pesce, F., Newcombe, E. A., Seiffert, P., Tranchant, E. E., Olsen, J. G., Grace, C. R., Kragelund, B. B., & Lindorff-Larsen, K. (2022). Assessment of models for calculating the hydrodynamic radius of intrinsically disordered proteins. Biophysical Journal. https://doi.org/10.1016/j.bpj.2022.12.013
ANAC046	36	0.3	NAPSTTITTTKQLSRIDSLDNIDHLLDFSSLPPLIDPGFLGQPGPSFSGARQQHDLKPVLHHPTTAPVDNTYLPTQALNFPYHSVHNSGSDFGYGAGSGNNNKGMIKLEHSLVSVSQETGLSSDVNTTATPEISSYPMMMNPAMMDGSKSACDG	Pesce, F., Newcombe, E. A., Seiffert, P., Tranchant, E. E., Olsen, J. G., Grace, C. R., Kragelund, B. B., & Lindorff-Larsen, K. (2022). Assessment of models for calculating the hydrodynamic radius of intrinsically disordered proteins. Biophysical Journal.

			LDDLIFWEDLYTS	https://doi.org/10.1016/j.bpj.2022.12.013
stath_NTD	9.1	0.3	DSSEEKFLRRIGRFG	Rieloff, E., & Skepö, M. (2020). Phosphorylation of a disordered peptide—Structural effects and force field inconsistencies. Journal of Chemical Theory and Computation. https://pubs.acs.org/doi/abs/10.1021/acs.jctc.9b01190
A1_Aro_minus	27.9	0.8	GSMASASSSQRGRSGSGNSGGGRGGGFGGNDNFGRGGNSSGRGGFGGSRGGGGYGGSGDGYNGFGNDGSNSGGGGSSNDFGNYNNQSSNFGPMKGGNFGGRSSGGSGGGGQYSAKPRNQGGYGGSSSSSSGSGRRF	Martin, E. W., Holehouse, A. S., Peran, I., Farag, M., Incicco, J. J., Bremer, A., Grace, C. R., Soranno, A., Pappu, R. V., & Mittag, T. (2020). Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. Science, 367(6478), 694–699.
A1_Aro_minus_minus	29.3	0.5	GSMASASSSQRGRSGSGNSGGGRGGGFGGNDNSGRGGNSSGRGGFGGSRGGGGSGGSGDGYNGSGNDGSNSGGGGSSNDFGNSNNQSSNSGPMKGGNFGGRSSGGSGGGGQYSAKPRNQGGSGGSSSSSSGSGRRS	Martin, E. W., Holehouse, A. S., Peran, I., Farag, M., Incicco, J. J., Bremer, A., Grace, C. R., Soranno, A., Pappu, R. V., & Mittag, T. (2020). Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. Science, 367(6478), 694–699.
A1_Aro_plus	24.2	1.5	GSMAFASSFQRGRYGSGNFGGGRGGGFGGNDNFGRG	Martin, E. W., Holehouse, A. S., Peran, I., Farag, M., Incicco, J. J.,

			GNFSGRGGFGGSRGGGGY GGSGDGYNGFGNDGSNF GGGGSYNDFGNYNNQSSN FGPMKGGNFGGRSSGGSY GGGQYFAKPRNQGGYGGS SFSSSYGSGRRF	Bremer, A., Grace, C. R., Soranno, A., Pappu, R. V., & Mittag, T. (2020). Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. Science, 367(6478), 694–699.
HeV_PNT3_CTD_200_254	28	0	MSYYHHHHHHLESTSLYKK AGFTPTEEPPVIPEYYYGSG RRGDLSKSPPRGNVNLDSIK IYTSDDEDENQLEYEDEF	Nilsson, J. F., Baroudi, H., Gondelaud, F., Pesce, G., Bignon, C., Ptchelkine, D., Chamieh, J., Cottet, H., Kajava, A. V., & Longhi, S. (2022). Molecular Determinants of Fibrillation in a Viral Amyloidogenic Domain from Combined Biochemical and Biophysical Studies. International Journal of Molecular Sciences, 24(1). https://doi.org/10.3390/ijms24010399
HeV_PNT3_200_310_YYY_AAA	40	0	MSYYHHHHHHLESTSLYKK AGFTPTEEPPVIPEAAAGSG RRGDLSKSPPRGNVNLDSIK IYTSDDEDENQLEYEDEFA KSSSEVVIDTTPEDNDSINQ EEVVGDPSDQGLEHPFPLG KFPEKEETPDVRRKDS	Nilsson, J. F., Baroudi, H., Gondelaud, F., Pesce, G., Bignon, C., Ptchelkine, D., Chamieh, J., Cottet, H., Kajava, A. V., & Longhi, S. (2022). Molecular Determinants of Fibrillation in a Viral Amyloidogenic Domain from Combined Biochemical and Biophysical Studies. International Journal of Molecular Sciences, 24(1). https://doi.org/10.3390/ijms24010399
HeV_PNT3_2	37	0	MSYYHHHHHHLESTSLYKK	Nilsson, J. F., Baroudi, H.,

00_310_WT			AGSTPTEEPPVIPEYYYGSG RRGDLSKSPPRGNVNLDSIK IYTSDDEDENQLEYEDEFA KSSSEVVIDTTPEDNDSINQ EEVVGDPSDQGLEHPFPLG KFPEKEETPDVRRKDS	Gondelaud, F., Pesce, G., Bignon, C., Ptchelkine, D., Chamieh, J., Cottet, H., Kajava, A. V., & Longhi, S. (2022). Molecular Determinants of Fibrillation in a Viral Amyloidogenic Domain from Combined Biochemical and Biophysical Studies. International Journal of Molecular Sciences, 24(1). https://doi.org/10.3390/ijms2401039 9
NiV_PNT3_20 0_314_WT	37	0	MSYYHHHHHHLESTSLYKK AGFDPAKDSPVIAEHYYGL GVKEQNVGPQTSRNVNLD SIKLYTSDDEEADQLEFED EFAGSSSEVIVGISPEDEEPS SVGGKPNESIGRTIEGQSIR DNLQAKDNKSTDVPGAGP KDS	Nilsson, J. F., Baroudi, H., Gondelaud, F., Pesce, G., Bignon, C., Ptchelkine, D., Chamieh, J., Cottet, H., Kajava, A. V., & Longhi, S. (2022). Molecular Determinants of Fibrillation in a Viral Amyloidogenic Domain from Combined Biochemical and Biophysical Studies. International Journal of Molecular Sciences, 24(1). https://doi.org/10.3390/ijms2401039 9
red1_288_345	25	0	GAMGISLPLLKQDDWLSSS KPFGSSTPNVVIEFDSDDD GDDFSNSKIEQSNLEKPPSN SENGGSHHHHHH	TBD
p150L_342_47 5	41	0	MAERLGKQLKLRAEREEK EKLKEEAKRAKEEAKKKK EEEKELKEKERREKREKD EKEKAEKQRLKEERRKER	Gopinathan Nair, A., Rabas, N., Lejon, S., Homiski, C., Osborne, M. J., Cyr, N., Sverzhinsky, A., Melendy, T., Pascal, J. M., Laue, E. D., Borden,

			QEALEAKLEEKRKKEEEK RLREEEKRIKAEKAEITRFF QKPKTPQAPKTLAGSCGKF APFEIKELEHHHHHH	K. L. B., Omichinski, J. G., & Verreault, A. (2022). Unorthodox PCNA Binding by Chromatin Assembly Factor 1. International Journal of Molecular Sciences, 23(19). https://doi.org/10.3390/ijms231911099
E1A_2022	36	0	GSMSHFEPPTLHELYDLDV TAPEDPNEEAVSQIFPDSV MLAVQEGIDLLTFPPAPGSP EPPHLSRQPEQPEQRALGP VSMPNLVPEVIDLYCYEQL NPPSDDEDEEGEEFVLDY	González-Foutel, N. S., Glavina, J., Borcherds, W. M., Safranchik, M., Barrera-Vilarmau, S., Sagar, A., Estaña, A., Barozet, A., Garrone, N. A., Fernandez-Ballester, G., Blanes-Mira, C., Sánchez, I. E., de Prat-Gay, G., Cortés, J., Bernadó, P., Pappu, R. V., Holehouse, A. S., Daughdrill, G. W., & Chemes, L. B. (2022). Conformational buffering underlies functional selection in intrinsically disordered protein regions. Nature Structural & Molecular Biology, 29(8), 781–790.
RelA_TAD	27	0	MGSVPKPAPQPYTFPASLST INFDEFSPMLLPSGQISNQA LALAPSSAPVLAQTMVPSSA MVPLAQPPAPAPVLTPGPP QSLSAPVPKSTQAGEGTLSE ALLHLQFDADEDLGALLG NSTDPGVFTDLASVDNSEF QQLLNQGVSMSHSTAEPML MEYPEAITRLVTGSQRPPDP APTPLGTSGLPNGLSGDED	Baughman, H. E. R., Narang, D., Chen, W., Villagrán Suárez, A. C., Lee, J., Bachochin, M. J., Gunther, T. R., Wolynes, P. G., & Komives, E. A. (2022). An intrinsically disordered transcription activation domain increases the DNA binding affinity and reduces the specificity of NFκB p50/RelA. The Journal of Biological

			FSSIADMDFSALLSQISSLEHHHHHH	Chemistry, 298(9), 102349.
EIF_450_1_249	52	0	GSMTDETAHPTQSASKQESAALKQTGDDQQESQQQRGYTNYNNGSNYTQKKPYNSNRPHQQRGGKFGPNRYNNRGNYNGGGSFRGGHMGANSSNVPWTGYYNNYPVYYQPQQMAAAGSAPANPIPVEEKSPVPTKIEITTKSGEHLDLKEQHKAKLQSQERSTVSPQPESKLKETSDSTSTSTPTPTTPSTNDSKASSEENISEAEKTRRNFIEQVKLRKAALEKKRKEQLEGSSGNNNIPMKTTPENVEEK	Chaves-Arquero, B., Martínez-Lumbreras, S., Sibille, N., Camero, S., Bernadó, P., Jiménez, M. Á., Zorrilla, S., & Pérez-Cañadillas, J. M. (2022). eIF4G1 N-terminal intrinsically disordered domain is a multi-docking station for RNA, Pab1, Pub1, and self-assembly. Frontiers in Molecular Biosciences, 9, 986121.
TIF2_624_774	37	0	ERADGQSRLHDSKGQTKLLQLLTTKSDQMEPSPLASSLSDTNKDSTGSLPGSGSTHGTSLKEKHKILHRLLQDSSSPVDLAKLTAEATGKDLSQESSSTAPGSEVTIKQEPVSPKKKENALLRYLLDKDDTKDIGLPEITPKLERLDSKT	Senicourt, L., le Maire, A., Allemand, F., Carvalho, J. E., Guee, L., Germain, P., Schubert, M., Bernadó, P., Bourguet, W., & Sibille, N. (2021). Structural insights into the interaction of the intrinsically disordered co-activator TIF2 with retinoic acid receptor heterodimer (RXR/RAR). Journal of Molecular Biology, 433(9), 166899.
IR_CTD	38	0	GPRRNQPAEQTTTTTTHTVVQQQTGGNTPAQGGTDATRAEDASLNRRDSQGSVAS	TBD

			THWSDSSSEVVNPYAEVGG ARNSLSAHQPEEHIYDEVA ADPGYSVIQNFSGSGPVTG RLIGTPGQGIQSTYALLANS GGLRLGMGGLTSGGESAVS SVNAAPTPGPVRFVWSHPQ FEK	
Tau_ht35_2022	46	0	EPPKSGDRSGYSSPGSPGTP GSRSRTPSLPTPPTREPKKV AVVRTPPKSPSSAKSRLQTA PVPMPDLKNVKSKIGSTEN LKHQPGGGKVQIINKKLD LSNVQSKCGSKDNIKHVPG GGSVQIVYKPVDLSKVTSK CGSLGNIHHKPGGGQVEV KSEKLDFKDRVQSKIGSLD NITHVPGGGNKKIETHKLT FRENAKAKTDHGAEIVYKS PVVSGDTSPRHLSNVSSTGSI DMVDSPQLATLADEVSASL AKQGL	Lyu, C., Da Vela, S., Al-Hilaly, Y., Marshall, K. E., Thorogate, R., Svergun, D., Serpell, L. C., Pastore, A., & Hanger, D. P. (2021). The Disease Associated Tau35 Fragment has an Increased Propensity to Aggregate Compared to Full-Length Tau. Frontiers in Molecular Biosciences, 8, 779240.
Tau_ht410_2N3R	63	0	MAEPRQEFEVMEDHAGTY GLGDRKDQGGYTMHQDQ EGDTDAGLKESPLQTPTED GSEEPGSETSDAKSTPTAED VTAPLVDEGAPGKQAAAQ PHTEIPEGTTAEEAGIGDTP SLEDEAAGHVTQARMVSKS KDGTGSDDKKAKGADGK TKIATPRGAAPPGQKGQAN ATRIPAKTPPAPKTPPSSGEP	Lyu, C., Da Vela, S., Al-Hilaly, Y., Marshall, K. E., Thorogate, R., Svergun, D., Serpell, L. C., Pastore, A., & Hanger, D. P. (2021). The Disease Associated Tau35 Fragment has an Increased Propensity to Aggregate Compared to Full-Length Tau. Frontiers in Molecular Biosciences, 8, 779240.

			PKSGDRSGYSSPGSPGTPGS RSRTPSLPTPPTREPKKVAV VRTPPKSPSSAKSRLQTAPV PMPDLKNVKSKIGSTENLK HQPGGGKVQIVYKPVDLS KVTSKCGSLGNIHHKPGG GQVEVKSEKLDFKDRVQS KIGSLDNITHVPGGGNKKI ETHKLTFRENAKAKTDHG AEIVYKSPVVSGDTSPRHLS NVSSTGSIDMVDSPQLATLA DEVSASLAKQGL	
Tau_ht410_2N4R	67	0	MAEPRQEFEVMEDHAGTY GLGDRKDQGGYTMHQDQ EGDTDAGLKESPLQTPTED GSEEPGSETSDAKSTPTAED VTAPLVDEGAPGKQAAAQ PHTEIPEGTTAEEAGIGDTP SLEDEAAGHVTQARMVSKS KDGTGSDDKKAKGADGK TKIATPRGAAPPGQKGQAN ATRIPAKTPPAPKTPPSSGEP PKSGDRSGYSSPGSPGTPGS RSRTPSLPTPPTREPKKVAV VRTPPKSPSSAKSRLQTAPV PMPDLKNVKSKIGSTENLK HQPGGGKVQIINKKLDLSN VQSKCGSKDNIKHVPGGGS VQIVYKPVDLSKVTSKCGSL GNIHHKPGGGQVEVKSEK LDFKDRVQSKIGSLDNITH VPGGGNKKIETHKLTFRE	Lyu, C., Da Vela, S., Al-Hilaly, Y., Marshall, K. E., Thorogate, R., Svergun, D., Serpell, L. C., Pastore, A., & Hanger, D. P. (2021). The Disease Associated Tau35 Fragment has an Increased Propensity to Aggregate Compared to Full-Length Tau. Frontiers in Molecular Biosciences, 8, 779240.

			NAKAKTDHGAEIVYKSPVVSGDTSPRHLSNVSSTGSIDMVDSPQLATLADEVSASLAKQGL	
SMAD_linker	29	0	GPLPPVLVPRHTEILTELPPLDDYTHSIPENTNFPAGIEPQSNYIPETPPPGYISEDGETSDQQLNQSMDTGSPAELSPTTLSPVNHSLD	Gomes, T., Martin-Malpartida, P., Ruiz, L., Aragón, E., Cordeiro, T. N., & Macias, M. J. (2021). Conformational landscape of multidomain SMAD proteins. Computational and Structural Biotechnology Journal, 19, 5210–5224.
MenV_LBD	25	0	TTIKIMDPGVGDGATAAKSKRLFKEAPVVVSGPVIGDNPIVDADTIQLDELARPSLPKTKSQ	Webby, M. N., Herr, N., Bulloch, E. M. M., Schmitz, M., Keown, J. R., Goldstone, D. C., & Kingston, R. L. (2021). Structural Analysis of the Menangle Virus P Protein Reveals a Soft Boundary between Ordered and Disordered Regions. Viruses, 13(9). https://doi.org/10.3390/v13091737
syndecan3_ED	65	0	MGSSHHHHHHSSGLVPRGSMAQRWRSENFERPVDLEGSGDDDSFPDDELDDLYSGSGSGYFEQESGIETAMETRFSPDVALAVSTTPAVLPTTNIQPVGTPFEELPSERPTLEPATSPLVVTEVPEEPSQRATTVSTTMETATTAATSTGDPTVATVPATVATATPSTPAAPPFTA	Gondelaud, F., Bouakil, M., Le Fèvre, A., Miele, A. E., Chirot, F., Duclos, B., Liwo, A., & Ricard-Blum, S. (2021). Extended disorder at the cell surface: The conformational landscape of the ectodomains of syndecans. Matrix Biology Plus, 12, 100081.

			TTAVIRTTGVRRLLPLPLTT VATARATTPEAPSPPTTAAV LDTEAPTPRLVSTATSRPRA LPRPATTQEPDIPERSTLPL GTTAPGPTEVAQTPTPETF LTTIRDEPEVPVSGGPSGDF ELPEEETTQPDTANEVVAV GGAAAKASSPPGTLPKGAR PGPGLLDNAIDSGSSAAQLP QKSILERKEVLVDYKDDD DK	
syndecan4	42	0	GSSHHHHHHSSGLVPRGSH MESIRETEVIDPQDLLEGRY FSGALPDDEDVVGPGQES DDFELSGSGDLDDLEDSMI GPEVVHPLVPLDNHIPERA GSGSQVPTEPKKLEENEVI PKRISPVEESEDVSNKVSMS STVQGSNIFERTEVLAGCPE HDYKDDDDK	Gondelaud, F., Bouakil, M., Le Fèvre, A., Miele, A. E., Chirot, F., Duclos, B., Liwo, A., & Ricard-Blum, S. (2021). Extended disorder at the cell surface: The conformational landscape of the ectodomains of syndecans. Matrix Biology Plus, 12, 100081.
N_FATZ_1	35	0	MAHHHHHHVDDDDKIMP LSGTPAPNKKRKSSKLIMEL TGGGQESSGLNLGKKISVP RDVMLEELSLLTNRGSKMF KLRQMRVEKFIYENHPDVF SDSSMDHFQKFLPTVGGQL GTAGQGFSYSKSNGRGGS QAGGSGSAGQYGSDQQHH LGSGSGAGGTGGPAGQAG RGGAAGTAGVGETGSGDQ	Sponga, A., Arolas, J. L., Schwarz, T. C., Jeffries, C. M., Rodriguez Chamorro, A., Kostan, J., Ghisleni, A., Drepper, F., Polyansky, A., De Almeida Ribeiro, E., Pedron, M., Zawadzka-Kazimierczuk, A., Mlynek, G., Peterbauer, T., Doto, P., Schreiner, C., Hollerl, E., Mateos, B., Geist, L., … Djinović-Carugo, K. (2021). Order from disorder in the sarcomere: FATZ forms a fuzzy but

			AGGEAE	tight complex and phase-separated condensates with α-actinin. Science Advances, 7(22). https://doi.org/10.1126/sciadv.abg7653
DeltaN_FATZ_1	39	0	GPTVGGQLGTAGQGFSYSKSNGRGGSQAGGSGSAGQYGSDQQHHLGSGSGAGGTGGPAGQAGRGGAAGTAGVGETGSGDQAGGEGKHITVFKTYISPWERAMGVDPQQKMELGIDLLAYGAKAELPKYKSFNRTAMPYGGYEKASKRMTFQMPKFDLGPLLSEPLVLYNQNLSNRPSFNRTPIPWLSSGEPVDYNVDIGIPLDGETEEL	Sponga, A., Arolas, J. L., Schwarz, T. C., Jeffries, C. M., Rodriguez Chamorro, A., Kostan, J., Ghisleni, A., Drepper, F., Polyansky, A., De Almeida Ribeiro, E., Pedron, M., Zawadzka-Kazimierczuk, A., Mlynek, G., Peterbauer, T., Doto, P., Schreiner, C., Hollerl, E., Mateos, B., Geist, L., … Djinović-Carugo, K. (2021). Order from disorder in the sarcomere: FATZ forms a fuzzy but tight complex and phase-separated condensates with α-actinin. Science Advances, 7(22). https://doi.org/10.1126/sciadv.abg7653
histatin_2021	15	0	DSHAKRHHGYKRKFHEKHHSHRGY	Sagar, A., Jeffries, C. M., Petoukhov, M. V., Svergun, D. I., & Bernadó, P. (2021). Comment on the Optimal Parameters to Derive Intrinsically Disordered Protein Conformational Ensembles from Small-Angle X-ray Scattering Data Using the Ensemble Optimization Method. Journal of Chemical Theory and Computation,

				17(4), 2014–2021.
synthELP	66	0	GGVPGAIPGGVPGGVFYPG AGLGALGGGALGPGGKPL KPVPGGLAGAGLGAGLGA FPAVTFPGALVPGGVADAA AAYKAAKAGAGLGGVPGV GGLGVSAGAVVPQPGAGV KPGKVPGVGLPGVYPGGV LPGARFPGVGVLPGVPTGA GVKPKAPGVGGAFAGIPG VGPFGGPQPGVPLGYPIKA PKLPGGYGLPYTTGKLPYG YGPGGVAGAAGKAGYPTG TGVGPQAAAAAAAKAAAK FGAGAAGVLPGVGGAGVP GVPGAIPGIGGIAGVGTPA AAAAAAAAAKAAKYGAAA GLVPGGPGFGPGVVGVPG AGVPGVGVPGAGIPVVPGA GIPGAAVPGVVSPEAAAKA AAKAAKYGARPGVGVGGI PTYGVGAGGFPGFGVGVG GIPGVAGVPGVGGVPGVG GVPGVGISPEAQAAAAAKA AKYGVGTPAAAAAKAAAK AAQFGLVPGVGVAPGVGV APGVGVAPGVGLAPGVGV APGVGVAPGVGVAPGIGP GGVAAAAKSAAKVAAKAQ LRAAAGLGAGIPGLGVGVG VPGLGVGAGVPGLGVGAG VPGFGAVPGALAAAKAAK	Lockhart-Cairns, M. P., Newandee, H., Thomson, J., Weiss, A. S., Baldock, C., & Tarakanova, A. (2020). Transglutaminase-mediated cross-linking of tropoelastin to fibrillin stabilises the elastin precursor prior to elastic fibre assembly. Journal of Molecular Biology, 432(21), 5736–5751.

			YGAAVPGVLGGLGALGGV GIPGGVVGAGPAAAAAAA KAAAKAAQFGLVGAAGLG GLGVGGLGVPGVGGLGGI PPAAAAKAAKYGAAGLGG VLGGAGQFPLGGVAARPG FGLSPIFPGGACLGKACGR KRK	
UL11	24	0	MGLSFSGTRPCCCRNNVLIT DDGEVVSLTAHDFDVVDIE SEEEGNFYVPPDMRGVTRA PGRQRLRSSDPPSRHTHRRT PGGACPATQFPPPMSDSEW SHPQFEK	Metrick, C. M., Koenigsberg, A. L., & Heldwein, E. E. (2020). Conserved Outer Tegument Component UL11 from Herpes Simplex Virus 1 Is an Intrinsically Disordered, RNA-Binding Protein. mBio, 11(3). https://doi.org/10.1128/mBio.00810-20
GON7_NTD	31	0	MGHHHHHHENLYFQGELL GEYVGQEGKPQKLRVSCE APGDGDPFQGLLSGVAQM KDMVTELFDPLVQGEVQH RVAAAPDEDLDGDDEDDA EDENNIDNRTNFDGPSAKR PKTPS	Arrondel, C., Missoury, S., Snoek, R., Patat, J., Menara, G., Collinet, B., Liger, D., Durand, D., Gribouval, O., Boyer, O., Buscara, L., Martin, G., Machuca, E., Nevo, F., Lescop, E., Braun, D. A., Boschat, A.-C., Sanquer, S., Guerrera, I. C., … Mollet, G. (2019). Defects in t6A tRNA modification due to GON7 and YRDC mutations lead to Galloway-Mowat syndrome. Nature Communications, 10(1), 3967.
Bmal1_CTD_P	28	0	GPDASSPGGKKILNGGTPD	Garg, A., Orru, R., Ye, W., Distler, U.,

624A			IPSTGLLPGQAQETPGYPYS DSSSILGENPHIGIDMIDND QGSSSPSNDEAAMAVIMSLL EADAGLGGPVDFSDLPWA L	Chojnacki, J. E., Köhn, M., Tenzer, S., Sönnichsen, C., & Wolf, E. (2019). Structural and mechanistic insights into the interaction of the circadian transcription factor BMAL1 with the KIX domain of the CREB-binding protein. The Journal of Biological Chemistry, 294(45), 16604–16619.
NID_2059_2325	47	0	GPHMQVPRTHRLITLADHI CQIITQDFARNQVPSQASTS TFQTSPSALSSTPVRTKTSSR YSPESQSQTVLHPRPGPRVS PENLVDKSRGSRPGKSPERS HIPSEPYEPISPPQGPAVHE KQDSMLLLSQRGVDPAEQR SDSRSPGSISYLPSFFTKLEST SPMVKSKKQEIFRKLNSSG GGDSDMAAAQPGTEIFNLP AVTTSGAVSSRSHSFADPAS NLGLEDIIRKALMGSFDDK VEDHGVVMSHPVGIMPGS ASTSVVTSSEARRDE	Cordeiro, T. N., Sibille, N., Germain, P., Barthe, P., Boulahtouf, A., Allemand, F., Bailly, R., Vivat, V., Ebel, C., Barducci, A., Bourguet, W., le Maire, A., & Bernadó, P. (2019). Interplay of protein disorder in retinoic acid receptor heterodimer and its corepressor regulates gene expression. Structure , 27(8), 1270–1285.e6.
MAP2c	67	0	MADERKDEGKAPHWTSAS LTEAAAHPHSPEMKDQGG SGEGLSRSANGFPYREEEE GAFGEHGSQGTYSDTKEN GINGELTSADRETAEEVSA RIVQVVTAEAVAVLKGEQE KEAQHKDQPAALPLAAEE TVNLPPSPPPSPASEQTAAL EEATSGESAQAPSAFKQAK	Melková, K., Zapletal, V., Jansen, S., Nomilner, E., Zachrdla, M., Hritz, J., Nováček, J., Zweckstetter, M., Jensen, M. R., Blackledge, M., & Žídek, L. (2018). Functionally specific binding regions of microtubule-associated protein 2c exhibit distinct conformations and dynamics. The Journal of Biological Chemistry,

			DKVTDGITKSPEKRSSLPRP SSILPPRRGVSGDREENSFSL NSSISSARRTTRSEPIRRAGK SGTSTPTTPGSTAITPGTPPS YSSRTPGTPGTPSYPRTPGT PKSGILVPSEKKVAIIRTPPK SPATPKQLRLINQPLPDLKN VKSKIGSTDNIKYQPKGGQ VQIVTKKIDLSHVTSKCGSL KNIRHRPGGGRVKIESVKL DFKEKAQAKVGSLDNAHH VPGGGNVKIDSQKLNFRE HAKARVDHGAEIITQSPSRS SVASPRRLSNVSSSGSINLLES PQLATLAEDVTAALAKQGL	293(34), 13297–13309.
TRF2_BR	17	0	GPPGSMAGGGGSSDGSGR AAGRRASRSSGRARRGRHE PGLGGPAERGAG	Necasová, I., Janoušková, E., Klumpler, T., & Hofr, C. (2017). Basic domain of telomere guardian TRF2 reduces D-loop unwinding whereas Rap1 restores it. Nucleic Acids Research, 45(21), 12170–12180.

Chapter 5: The SARS-CoV-2 Nucleocapsid Protein Is Dynamic, Disordered, and Phase Separates With RNA

This chapter was published in the journal Nature Communications as:
Cubuk J, Alston JJ, Incicco JJ, Singh S, Stuchell-Brereton MD, Ward MD, Zimmerman MI, Vithani N, Griffith D, Wagoner JA, Bowman GR, Hall KB, Soranno A, Holehouse AS. The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA. *Nat Commun.* 2021 Mar 29;12(1):1936. doi: 10.1038/s41467-021-21953-3. PMID: 33782395; PMCID: PMC8007728.

Author Contributions: J.C. designed, expressed, and purified the constructs, performed the single-molecule spectroscopy and oligomerization experiments, analyzed the corresponding data, and wrote the manuscript. J.J.A. performed coarse-grained simulations and wrote the manuscript. J.J.I. performed turbidity experiments and wrote the manuscript. M.D.S.B. designed the constructs for single-molecule spectroscopy experiments, supervised protein expression and purification and oligomerization experiments, and wrote the manuscript. S.S., M.D.W., M.I.Z. and N.V. set up, curated, analyzed, and managed molecular dynamics simulations on both local resources and the Folding@Home supercomputer. D.G. performed bioinformatic analysis. J.A.W. performed theoretical analysis. G.R.B. acquired funding. K.B.H. wrote the manuscript. A.S. conceived of the study, analyzed data, wrote the manuscript, and acquired funding. A.S.H. conceived of the study, analyzed data, performed and analyzed all-atom Monte Carlo simulations and coarse-grained simulations, wrote the manuscript, and acquired funding. G.R.B., K.B.H., A.S. and A.S.H. jointly supervised the work.

5.1 Abstract

The SARS-CoV-2 nucleocapsid (N) protein is an abundant RNA binding protein critical for viral genome packaging, yet the molecular details that underlie this process are poorly understood. Here we combine single-molecule spectroscopy with all-atom simulations to uncover the molecular details that contribute to N protein function. N protein contains three dynamic disordered regions that house putative transiently-helical binding motifs. The two folded domains interact minimally such that full-length N protein is a flexible and multivalent RNA binding protein. N protein also undergoes liquid-liquid phase separation when mixed with RNA, and polymer theory predicts that the same multivalent interactions that drive phase separation also engender RNA compaction. We offer a simple symmetry-breaking model that provides a plausible route through which single-genome condensation preferentially occurs over phase separation, suggesting that phase separation offers a convenient macroscopic readout of a key nanoscopic interaction.

5.2 Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is an enveloped, positive-strand RNA virus that causes the disease COVID-19 (Coronavirus Disease-2019) [622]. While coronaviruses typically cause relatively mild respiratory diseases, as of February 2021 COVID-19 is on course to kill 2.5 million people since its emergence in late 2019 [622–624]. While recent progress in vaccine development has been remarkable, the emergence of novel coronaviruses in human populations represents a continuing threat[625]. As a result, therapeutic approaches that address fundamental and general viral mechanisms will offer a key route for first-line intervention against future pandemics.

A challenge in identifying candidate drugs is our relatively sparse understanding of the molecular details that underlie the function of SARS-CoV-2 proteins. As a result, there is a surge of biochemical and biophysical exploration of these proteins, with the ultimate goal of identifying proteins that are suitable targets for disruption, ideally with insight into the molecular details of how disruption could be achieved [626,627].

While much attention has been focused on the Spike (S) protein, many other SARS-CoV-2 proteins play equally critical roles in viral physiology, yet we know relatively little about their structural or biophysical properties [628–631]. Here we performed a high-resolution structural and biophysical characterization of the SARS-CoV-2 nucleocapsid (N) protein, the protein responsible for genome packaging [632,633]. A large fraction of N protein is predicted to be intrinsically disordered, which constitutes a major barrier to conventional structural characterization [48]. To overcome these limitations, we combined single-molecule spectroscopy with all-atom simulations to build a residue-by-residue description of all three disordered regions in the context of their folded domains. The combination of single-molecule spectroscopy and simulations to reconstruct structural ensembles has been applied extensively to uncover key molecular details underlying disordered protein regions [56,90,140,219,237,324]. Our goal here is to provide biophysical and structural insights into the physical basis of N protein function.

In exploring the molecular properties of N protein, we discovered it undergoes phase separation with RNA, as was also reported recently [391,392,442,464,634–637.] Given N protein underlies viral packaging, we reasoned phase separation may in fact be an unavoidable epiphenomenon that reflects physical properties necessary to drive the compaction of long genomic RNA molecules. To explore this

principle further, we developed a simple physical model, which suggested symmetry breaking through a small number of high-affinity binding sites can organize anisotropic multivalent interactions to drive single-polymer compaction, as opposed to multi-polymer phase separation. Irrespective of its physiological role, our results suggest that phase separation provides a macroscopic readout (visible droplets) of a nanoscopic process (protein:RNA and protein:protein interaction). In the context of SARS-CoV-2, those interactions are expected to be key for viral packaging, such that assays which monitor phase separation of N protein with RNA may offer a convenient route to identify compounds that will also attenuate viral assembly.

5.3 Results

Coronavirus nucleocapsid proteins are multi-domain RNA binding proteins that play a critical role in many aspects of the viral life cycle [633,638]. The SARS-CoV-2 N protein shares substantial sequence conservation with other coronavirus nucleocapsid proteins (**Fig. S1-5**). Work on N protein from a range of model coronaviruses has shown that N protein undergoes both self-association, interaction with other proteins, and interaction with RNA, all in a highly multivalent manner.

The SARS-CoV-2 N protein can be divided into five domains; a predicted intrinsically disordered N-terminal domain (NTD), an RNA binding domain (RBD), a predicted disordered central linker (LINK), a dimerization domain, and a predicted disordered C-terminal domain (CTD) (**Fig. 1A-C**). While SARS-CoV-2 is a novel coronavirus, decades of work on model coronaviruses (including SARS coronavirus) have revealed a number of features expected to hold true in the SARS-CoV-2 N protein. Notably, all five domains are predicted to bind RNA [639–645], and while the dimerization domain facilitates the formation of well-defined stoichiometric dimers, RNA-independent higher-

order oligomerization is also expected to occur [644,646–648]. Importantly, protein-protein and protein-RNA interaction sites have been mapped to all three disordered regions.

Despite recent structures of the RBD (**Fig. 1B**) and dimerization domains (**Fig. 1C**) from SARS-CoV-2, the solution-state conformational behavior of the full-length protein remains elusive [649–651]. Understanding N protein function necessitates a mechanistic understanding of the flexible predicted disordered regions and their interplay with the folded domains. A recent small-angle X-ray study shows good agreement with previous work on SARS, suggesting the LINK is relatively extended, but neither the structural basis for this extension nor the underlying dynamics are known [639,652].

Here, we address these questions by probing three full-length constructs of the N protein with fluorescent labels (Alexa 488 and 594) flanking the NTD, the LINK, and the CTD (see **Fig. 1A** and **Table S1**). These constructs allow us to quantify conformations and dynamics of the disordered regions in the context of the full-length protein using single-molecule Förster Resonance Energy Transfer (FRET) and Fluorescence Correlation Spectroscopy (FCS) (see SI for details). We also investigated the stability of the RBD and truncated variants of the protein to test the role of long range interactions on the disordered regions (see **SI** and **Table S2**). In parallel to the experiments, we performed all-atom Monte Carlo simulations of each of the three IDRs in isolation and in context with their adjacent folded domains.

5.4 The NTD is disordered, flexible, and transiently interacts with the RBD

We started our analysis by investigating the NTD conformations. Under native conditions, single-molecule FRET measurements revealed the occurrence of a single population with a mean transfer

efficiency of 0.65 ± 0.03 (**Fig. 2A** and **Fig. S6**). To assess whether this transfer efficiency reports on a rigid distance (e.g., structure formation or persistent interaction with the RBD) or is a dynamic average across multiple conformations, we first compare the lifetime of the fluorophores with transfer efficiency. Under native conditions, the donor and acceptor lifetimes for the NTD construct lie on the line that represents fast conformational dynamics (**Fig. S7A**). To properly quantify the timescale associated with these fast structural rearrangements, we leveraged nanoseconds FCS. As expected for a dynamic population [168,234], the cross-correlation of acceptor-donor photons for the NTD is anticorrelated (**Fig. 2B and S12**). A global fit of the donor-donor, acceptor-acceptor, and acceptor-donor correlations yields a reconfiguration time τ_r = 170 ± 30 ns. This is longer than reconfiguration times observed for other proteins with a similar persistence length and charge content [36,168,653,654], hinting at a large contribution from internal friction due to rapid intramolecular contacts (formed either within the NTD or with the RBD) or transient formation of short structural motifs [185]. A conversion from transfer efficiency to chain dimensions can be obtained by assuming the distribution of distances computed from polymer models. Assuming a Gaussian chain distribution yields a root mean square distance between the fluorophores $r_{1\text{-}68}$ of 48 ± 2 Å. When using the recently proposed self-avoiding walk (SAW) model (Zheng et al., 2018) (see **SI**), we compute a value of $r_{1\text{-}68}$ 47 ± 2 Å. This corresponds to values of persistence length (see SI) equal to 4.5 ± 0.4 Å and 4.3 ± 0.4 Å for the Gaussian and SAW distribution, respectively, which are similar to values reported for another unfolded protein under native conditions [36,168,199,653]. Overall, these results confirm the NTD is disordered, as predicted by sequence analysis.

We next examined the interaction of the NTD with other domains in the protein. We studied a truncated N protein variant that contains only the NTD and RBD domains (NTD-RBD) and

samples identical labeling positions. The root-mean-square distance $r_{1\text{-}68}$ is 46 ± 2 Å for both the Gaussian and SAW models, within errors from the NTD-FL values, suggesting no or limited interaction between the NTD and the LINKER, DIMER, and CTD domains (see **Fig. S8** and **Table S2**). We then assessed the role of the folded RBD and its influence on the conformations of the NTD by studying the effect of a chemical denaturant on the protein. The titration with guanidinium chloride (GdmCl) reveals a decrease of transfer efficiencies when moving from native buffer conditions to 1 M GdmCl, followed by a plateau of the transfer efficiencies at concentrations between 1 M and 2 M and a subsequent further decrease at higher concentrations (**Fig. S6** and **S8**). This behavior can be understood assuming that the plateau between 1 M and 2 M GdmCl represents the average of transfer efficiencies between two populations in equilibrium that have very close transfer efficiency and are not completely resolved because of shot noise. We interpret these two populations as the contribution of the folding and unfolding fraction of the RBD domain on the distances probed by the NTD-FL construct, which includes a labeling position within the folded RBD. Indeed, this interpretation is supported by a broadening in the transfer efficiency peak between 1 M and 2 M GdmCl. Besides the effect of the unfolding of the RBD, the dimensions of the NTD FL are also modulated by a change in the solvent quality when adding denaturant **(Fig. 2C** and **Fig. S6, S8)** and this contribution to the expansion of the chain can be described using an empirical binding model [35,89,233,271,655]. A fit of the interdye root-mean-square distances to this model and the inferred stability of the RBD domain (midpoint: 1.3 ± 0.2 M; ΔG_0 = (5 ± 1) kcal mol^{-1}) are presented in **Fig. 2C.** A comparative fit of the histograms assuming two overlapping populations yields a consistent result in terms of RBD stability and protein conformations (**Fig. S9**). To confirm the inferred RBD stability results, we directly interrogated the RBD domain by measuring a full-length construct with labels in position 68 and 172, which flanks the folded RBD structure (see

section *RBD folding* in **SI**). Though the denaturation of the RBD reveals coexistence of up to three populations, which we identify as an unfolded, an intermediate, and a folded state, the range of the folding transition is compatible with the estimates made using the NTD constructs (midpoint: 1.68 ± 0.02 M, see **Fig. S9** and **Table S6**).

To better understand the sequence-dependent conformational behavior of the NTD we turned to all-atom simulations of an NTD-RBD construct. We used a novel sequential sampling approach that integrates long timescale MD simulations performed using the Folding@home distributed computing platform with all-atom Monte Carlo simulation performed with the ABSINTH forcefield to generate an ensemble of almost 400,000 distinct conformations (see methods) [104,283,284]. We also performed simulations of the NTD in isolation.

We observed good agreement between simulation and experiment for the equivalent inter-residue distance (**Fig. 2D**). The peaks on the left side of the histogram reflect specific simulations where the NTD engages more extensively with the RBD through a fuzzy interaction, leading to local kinetic traps [656]. We also identified several regions in the NTD where transient helices form, and using normalized distance maps found regions of transient attractive and repulsive interaction between the NTD and the RBD (**Fig. 2E**). In particular, the basic beta-strand extension from the RBD (**Fig. 1B**) repels the arginine-rich C-terminal region of the NTD, while a phenylalanine residue (F17) in the NTD engages with a hydrophobic face on the RBD (**Fig. 2G**). Finally, we noticed the arginine-rich C-terminal residues (residues 31 - 38) form a transient alpha helix projecting three of the four arginines in the same direction (**Fig. 2F, 2H**). These features provide molecular insight into previously reported functional observations (see *Discussion*).

5.5 The linker is highly dynamic and there is minimal interaction between the RBD and the dimerization domain

We next turned to the linker (LINK FL) construct to investigate how the disordered region modulates the interaction and dynamics between the two folded domains. Under aqueous buffer conditions, single-molecule FRET reveals the coexistence of two overlapping populations with mean transfer efficiencies of 0.55 ± 0.03 and 0.75 ± 0.03, respectively (**Fig. 3A**). A small change in ionic strength of the solution is sufficient to alter the equilibrium between these two populations and favor the low transfer efficiency state (see **inset Fig. 3C**). Comparison of the fluorescence lifetimes and transfer efficiencies indicates that, like the NTD, the transfer efficiencies represent dynamic conformational ensembles sampled by the LINK (**Fig. S7A**). ns-FCS confirms fast dynamics across the measured distribution of transfer efficiencies, with a characteristic reconfiguration time τ_r of 120 ± 20 ns (**Fig. 3B** and **S12**). This reconfiguration time is compatible with high internal friction effects, as observed for other unstructured proteins [168,653], but may also account for the drag of the surrounding domains. The root-mean-square interdye distance corresponding to the low transfer efficiency population $r_{172\text{-}245}$ is equal to 55 ± 2 Å (l_p= 5.4 ± 0.4 Å) when assuming a Gaussian chain distribution and 54 ± 2 Å (l_p= 5.2 ± 0.4 Å) when using a SAW model (see SI). The one corresponding to the high transfer efficiency population is equal to 42 ± 2 Å when assuming a Gaussian Chain distribution or 45 ± 2 Å using the SAW model (with a corresponding l_p = 3.2 ± 0.3 Å and l_p = 3.6 ± 0.3 Å, respectively) (see **SI**).

Next, we addressed whether the LINK segment populates elements of persistent secondary structure or forms stable interaction with the RBD or dimerization domains. The addition of

denaturant leads to the rapid loss of the high transfer efficiency population and a continuous shift of the remaining population toward lower transfer efficiencies (**Fig. S6,S8**). These results correspond to an almost linear expansion of the chain in response to denaturant (see **Fig. 3C**).

To better understand the nature of the two populations and explain the weak dependence of the linker expansion on denaturant, we investigated the same labeling positions in the absence of the DIMER and CTD domains (LINK ΔDIMER) (**Table S2**). smFRET measurements of this truncated version revealed a single population that undergoes a strong compaction with decreasing GdmCl concentration (**Fig. S6, S8**). Interestingly the transfer efficiency measured in aqueous buffer is equivalent to the one reported by the high transfer efficiency population of the LINK FL construct. The electrostatic nature of this compaction is clearly revealed by titrating a polar non ionic denaturant (Urea) and observing that the chain remains largely compact and recovers the same dimensions measured in GdmCl only when adding salt to the solution (**Fig. S10**). Overall, the LINK ΔDIMER observations lead us to speculate that the LINK domain can either self-interact or interact with the RBD domain, whereas addition of the DIMER and CTD domains restricts these configurations and largely favor more expanded states with the exceptions of very low ionic strength conditions. To further explore the configurations of the LINK, we turned again to Monte Carlo simulations.

As with the NTD, all-atom Monte Carlo simulations provide atomistic insight that can be compared with our spectroscopic results. Given the size of the system, an alternative sampling strategy to the NTD-RBD construct was pursued here that did not include MD simulations of the folded domains,

but we instead ran simulations of a construct that included the RBD, LINK and dimerization domain. In addition, we also performed simulations of the LINK in isolation.

We again found good agreement between simulations and experiment (**Fig. 3D**). The root mean square inter-residue distance for the low transfer efficiency population (between simulated positions 172 and 245) is 59.1 Å, which is within the experimental error of the single-molecule observations. Normalized distance map shows a number of regions of repulsion, notably that the RBD repels the N-terminal part of the LINK and the dimerization domain repels the C-terminal part of the LINK (**Fig. 3E**). We tentatively suggest this may reflect sequence properties chosen to prevent aberrant interactions between the LINK and the two folded domains. In the LINK-only simulations we identified two regions that form transient helices at low populations (20-25%), although these are less prominent in the context of the full-length protein (**Fig. 3F)**. These two helices encompass a serine-arginine (SR) rich region known to mediate both protein-protein and protein-RNA interaction. Helix H3 formation leads to the alignment of three arginine residues along one face of the helix. The second helix (H4) is a leucine/alanine-rich hydrophobic helix which may contribute to oligomerization, or act as a helical recognition motif for other protein interactions (notably as a nuclear export signal for Crm1, see *Discussion*).

5.6 The CTD engages in transient but non-negligible interactions with the dimerization domain

Finally, we again applied single-molecule FRET (**Fig. 4A**) and nsFCS (**Fig. 4B**) to understand the conformational behavior of the CTD FL construct. Single-molecule FRET experiments again reveal a single population with a mean transfer efficiency of 0.59 ± 0.03 (**Fig. 4A**) and the denaturant

dependence follows the expected trend for a disordered region, with a shift of the transfer efficiency toward lower values (**Fig. 4C**, and **Fig. S6** and **S8**), from 0.59 to 0.35. Interestingly, when studying the denaturant dependence of the protein, we noticed that the width of the distribution increases while moving toward aqueous buffer conditions. This suggests that the protein may form transient contacts or adopt local structure. Comparison with a truncated variant that contains only the CTD (**Fig. S8**) reveals a very similar distribution, with almost identical mean transfer efficiency but a narrower width (**Fig. S6)**, suggesting that part of the broadening is due to interactions with the neighboring domains.

To further investigate putative interaction between the CTD and neighboring domains, we turned to the investigation of protein dynamics. Though the comparison of the fluorophore lifetimes against transfer efficiency (**Fig. S7A**) appears to support a dynamic nature underlying the CTD FL population, nsFCS reveals a flat acceptor-donor cross-correlation on the nanosecond timescale (**Fig. 4B**). However, inspection of the donor-donor and acceptor-acceptor autocorrelations reveal a correlated decay. This is different from that expected for a completely static system such as polyprolines [170], where the donor-donor and acceptor-acceptor autocorrelation are also flat. An increase in the autocorrelations can be observed for static quenching of the dyes with aromatic residues. Interestingly, donor dye quenching can also contribute to a positive amplitude in the donor-acceptor correlation [204,657]. Therefore, a plausible interpretation of the flat cross-correlation data is that we are observing two populations in equilibrium whose correlations (one anticorrelated, reflecting conformational dynamics, and one correlated, reflecting quenching due contact formation) compensate each other.

To further investigate the possibility of two coexisting populations, we performed ns-FCS at increasing GdmCl concentrations. These experiments revealed a progressive increase of the anticorrelated amplitude in the cross-correlation, consistent with an increase of the dynamic population. Moreover, we also observed a simultaneous decrease in the overall donor-donor auto-correlation amplitude, consistent with a decrease in the quenched population (**Fig. S12**). Taken together, these results support our hypothesis that there are at least two distinct species existing in equilibrium. By analyzing the dynamic species between 0.16 and 0.6 M GdmCl, we quantified an average reconfiguration time (τ_r) of 64 ± 7 ns for the dynamic population in the CTD. Under the assumption that the mean transfer efficiency still originates (at least partially) from a dynamic distribution, the estimate of the inter-residue root-mean-square distance is $r_{363\text{-}419}$ = 51 ± 2 Å (l_p= 6.1 ± 0.5 Å) for a Gaussian chain distribution and $r_{363\text{-}419}$ = 48 ± 1 Å (l_p= 5.4 ± 0.4 Å) for the SAW model (see SI). However, some caution should be used when interpreting these numbers since we know there is some contribution from fluorophore static quenching, which may in turn contribute to an underestimate of the effective transfer efficiency [195].

We again obtained good agreement between all-atom Monte Carlo simulations and experiments (**Fig. 4D**). Scaling maps reveal extensive intramolecular interaction by the residues that make up H6, both in terms of local intra-IDR interactions and interaction with the dimerization domain (**Fig. 4E**). We identified two transient helices, one (H5) is minimally populated but the second (H6) is more highly populated in the IDR-only simulation and still present at ~20% in the folded state simulations (**Fig. 4F**). The difference reflects the fact that several of the helix-forming residues interact with the dimerization domain, leading to a competition between helix formation and intramolecular interaction. Mapping normalized distances onto the folded structure reveals that

interactions occur primarily with the N-terminal portion of the dimerization domain (**Fig. 4G**). As with the LINK and the NTD, a positively charged set of residues immediately adjacent to the folded domain in the CTD drive repulsion between this region and the dimerization domain. H6 is the most robust helix observed across all three IDRs, and is a perfect amphipathic helix with a hydrophobic surface on one side and charged/polar residues on the other (**Fig. 4H**). The cluster of hydrophobic residues in H6 engage in intramolecular contacts and offer a likely physical explanation for the complex nsFCS data in aqueous buffer.

5.7 N protein undergoes phase separation with RNA

Over the last decade, biomolecular condensates formed through phase separation have emerged as a new mode of cellular organization [352,357,438,658]. Many of the proteins that have been shown to drive phase separation *in vitro* are RNA binding proteins with intrinsically disordered regions [357,659]. Moreover, multivalency is the key molecular feature that determines if a biomolecule can undergo higher-order assembly [359]. Having characterized N protein to reveal three IDRs with distinct binding sites for both protein-protein and protein-RNA interactions it became clear that N protein poses all of the features consistent with a protein that may undergo phase separation. With these results in hand, we anticipated that N protein would undergo phase separation with RNA [80,473,660].

In line with this expectation, we observed robust droplet formation with homopolymeric RNA (**Fig. 5A-B**) under aqueous buffer conditions, both at 50 mM Tris and at a higher salt concentration of 50 mM NaCl. Turbidity assays at different concentrations of protein and poly(rU) (200-250 nucleotides) demonstrate the classical reentrant phase behavior expected for a system undergoing heterotypic interaction (**Fig. 5C-D**). It is to be noted that turbidity experiments do not exhaustively cover all the conditions for phase separation and are only indicative of the low-boundary concentration regime

explored in the current experiments. In particular, turbidity experiments do not provide a measurement of tie-lines, though they are inherently a reflection of the free energy and chemical potential of the solution mixture [661]. Interestingly, phase separation occurs at relatively low concentrations, in the low μM range, which are compatible with physiological concentration of the protein and nucleic acids. Though increasing salt concentration results in an upshift of the phase boundaries, one has to consider that in a cellular environment this effect might be counteracted by cellular crowding.

One peculiar characteristic of our measured phase-diagram is the narrow regime of conditions in which we observe phase separation of nonspecific RNA at a fixed concentration of protein. This leads us to hypothesize that the protein may have evolved to maintain tight control of concentrations at which phase separation can (or cannot) occur. Interestingly, when rescaling the turbidity curves as a ratio between protein and RNA, we find all the curve maxima aligning at a similar stoichiometry, approximately 20 nucleotides per protein in absence of added salt and 30 nucleotides when adding 50 mM NaCl (**Fig. S13**). These ratios are in line with the charge neutralization criterion proposed by Banerjee *et al.*, since the estimated net charge of the protein at pH 7.4 is +24 [421]. Finally, given we observed phase separation with poly(rU), the behavior we are observing is likely driven by relatively nonspecific protein:RNA interactions. In agreement, work from a number of other groups has also established this phenomenon across a range of solution conditions and RNA types [391,392,442,464,634–637].

Having established phase separation through a number of assays, we wondered what -if any- physiological relevance this may have for the normal biology of SARS-CoV-2.

5.8 A simple polymer model shows symmetry-breaking can facilitate multiple metastable single-polymer condensates instead of a single multi-polymer condensate

Why might phase separation of N protein with RNA be advantageous to SARS-CoV-2? One possible model is that large, micron-sized cytoplasmic condensates of N protein and RNA form through phase separation and facilitate genome packaging. These condensates may act as molecular factories that help concentrate the components for pre-capsid assembly (where we define a pre-capsid here simply as a species that contains a single copy of the genome with multiple copies of the associated N protein), a model that has been proposed in other viruses [390].

However, given that phase separation is unavoidable when high concentrations of multivalent species are combined, we propose that an alternative interpretation of our data is that in this context, phase separation is simply an inevitable epiphenomenon that reflects the inherent multi-valency of the N protein for itself and for RNA. This poses questions about the origin of specificity for viral genomic RNA (gRNA), and, of focus in our study, how phase separation might relate to a single genome packaging through RNA compaction.

Given the expectation of a single genome per virion, we reasoned SARS-CoV-2 might have evolved a mechanism to limit phase separation with gRNA (i.e., to avoid multi-genome condensates), with a preference instead for single-genome packaging (single-genome condensates). This mechanism may exist in competition with the intrinsic phase separation of the N protein with other nonspecific RNAs (nsRNA).

One possible way to limit phase separation between two components (e.g., gRNA/nsRNA and N protein) is to ensure the levels of these components are held at a sufficiently low total concentration such that the phase boundary is never crossed. While possible, such a regulatory mechanism is at the mercy of extrinsic factors that may substantially modulate the saturation concentration [662–664]. Furthermore, not only must phase separation be prevented, but gRNA compaction should also be promoted through the binding of N protein. In this scenario, the affinity between gRNA and N protein plays a central role in determining the required concentration for condensation of the macromolecule (gRNA) by the ligand (N protein).

Given a system composed of components with defined valencies, phase boundaries are encoded by the strength of interaction between the interacting domains in the components. Considering a long polymer (e.g., gRNA) with proteins adsorbed onto that polymer as adhesive points (stickers), the physics of associative polymers predicts that the same interactions that cause phase separation will also control the condensation of individual long polymers [80,359,398,408,409,665]. With this in mind, we hypothesized that phase separation is reporting on the physical interactions that underlie genome compaction.

To explore this hypothesis, we developed a simple computational model where the interplay between compaction and phase separation could be explored. Our setup consists of two types of species: long multivalent polymers and short multivalent binders (**Fig. 6A**). All interactions are isotropic, and each bead is inherently multivalent as a result. In the simplest instantiation of this model, favorable polymer:binder and binder:binder interactions are encoded, mimicking the scenario in which a binder (e.g., a protein) can engage in nonspecific polymer (RNA) interaction as well as

binder-binder (protein-protein) interaction. As expected for simulations of binders with homopolymer polymers we observed phase separation in a concentration-dependent manner (**Fig. 6B-E**). Phase separation gives rise to a single large spherical cluster with multiple polymers and binders (**Fig. 6D, 6H-L**).

Given our homopolymers undergo robust phase separation, we wondered if a break in the symmetry between intra- and inter-molecular interactions would be enough to promote single-polymer condensation in the same concentration regime over which we had previously observed phase separation. Symmetry breaking in our model is achieved through a single high-affinity binding site (**Fig. 6A**). We choose this particular mode of symmetry-breaking to mimic the presence of a packaging signal -a region of the genome that is essential for efficient viral packaging- an established feature in many viruses (including coronaviruses) although we emphasize this is a general model, as opposed to trying to directly model gRNA with a packaging signal [666–668].

We performed identical simulations to those in **Fig. 6C-D** using the same system with polymers that now possess a single high-affinity binding site (**Fig. 6E**). Under these conditions we did not observe large phase separated droplets (**Fig. 6F**). Instead, each individual polymer undergoes collapse to form a single-polymer condensate (**Fig. 6E**). Collapse is driven by the recruitment of binders to the high-affinity site, where they coat the chain, forming a local cluster of binders on the polymer. This cluster is then able to interact with the remaining regions of the polymer through weak nonspecific interactions, the same interactions that drove phase separation in **Fig. 6 B-D**. Symmetry breaking is achieved because the local concentration of binder around the site is high, such that intramolecular interactions are favored over intermolecular interaction. This high local concentration also drives

compaction at low binder concentrations. As a result, instead of a single multi-polymer condensate, we observe multiple single-polymers condensates, where the absolute number matches the number of polymers in the system (**Fig. 6G**).

The high affinity binding site polarizes the single-polymer condensate, such that they are organized, recalcitrant to fusion, and kinetically metastable. To illustrate this metastable nature, extended simulations using an approximate kinetic Monte Carlo scheme demonstrated that a high-affinity binding site dramatically slows assembly of multichain assemblies, but that ultimately these are the thermodynamically optimal configuration (**Fig. S18**). A convenient physical analogy is that of a micelle, which are non-stoichiometric stable assemblies. Even for micelles that are far from their optimal size, fusion is slow because it requires substantial molecular reorganization and the breaking of stable interactions [669,670].

Finally, we ran simulations under conditions in which binder:polymer interactions were reduced, mimicking the scenario in which non-specific protein:RNA interactions are inhibited (**Fig. 6L**). Under these conditions no phase separation occurs for polymers that lack a high-affinity binding site, while for polymers with a high-affinity binding site no chain compaction occurs (in contrast to when binder:polymer interactions are present, see **Fig. 6J**). This result illustrates how phase separation offers a convenient readout for molecular interactions that might otherwise be challenging to measure.

We emphasize that our conclusions from these coarse-grained simulations are subject to the parameters in our model. We present these results to demonstrate an example of how this single-

genome packaging could be achieved, offering a class of mechanism that may be in play. This is in contrast to the much stronger statement that this is how it is achieved, a statement that would require much more evidence to make. Recent elegant work by Ranganathan and Shakhnovich identified kinetically arrested microclusters, where slow kinetics result from the saturation of stickers within those clusters [432]. This is completely analogous to our results (albeit with homotypic interactions, rather than heterotypic interactions), giving us confidence that the physical principles uncovered are robust and, we tentatively suggest, quite general. Future simulations are required to systematically explore the details of the relevant parameter space in our system. However, regardless of those parameters, our model does establish that if weak multivalent interactions underlie the formation of large multi-polymer droplets, those same interactions cannot *also* drive polymer compaction inside the droplet.

5.9 Discussion

The nucleocapsid (N) protein from SARS-CoV-2 is a multivalent RNA binding protein critical for viral replication and genome packaging [632,633]. To better understand how the various folded and disordered domains interact with one another, we applied single-molecule spectroscopy and all-atom simulations to perform a detailed biophysical dissection of the protein, uncovering several putative interaction motifs. Furthermore, based on both sequence analysis and our single-molecule experiments, we anticipated that N protein would undergo phase separation with RNA. In agreement with this prediction, and in line with work from the Gladfelter and Yildiz groups working independently from us, we find that N protein robustly undergoes phase separation *in vitro* with model RNA under a range of different salt conditions. Using simple polymer models, we propose that the same interactions that drive phase separation may also drive genome packaging into a

dynamic, single-genome condensate. The formation of single-genome condensates (as opposed to multi-genome droplets) is influenced by the presence of one (or more) symmetry-breaking interaction sites, which we tentatively suggest could reflect packaging signals in viral genomes.

All three IDRs are highly dynamic

Our single-molecule experiments and all-atom simulations are in good agreement with one another and reveal that all three IDRs are extended and, depending on solution condition, highly dynamic. Simulations suggest the NTD may interact transiently with the RBD, which offers an explanation for the slightly slowed reconfiguration time measured by nanosecond FCS. The LINK shows rapid rearrangement, demonstrating the RBD and dimerization domain are not interacting. Finally, we see a pronounced interaction between the CTD and the dimerization domain, although these interactions are still highly transient.

Single-molecule experiments and all-atom simulations were performed on monomeric versions of the protein, yet N protein has previously been shown to undergo dimerization and form higher-order oligomers in the absence of RNA [646]. To assess the formation of oligomeric species, we use a combination of NativePAGE, crosslinking and FCS experiments (see **Fig. S14** and SI). These experiments also verified that under the conditions used for single-molecule experiments the protein exists only as a monomer.

5.10 Simulations identify multiple transient helices

We identified a number of transient helical motifs that provide structural insight into previously characterized molecular interactions. Transient helices are ubiquitous in viral disordered regions and

have been shown to underlie molecular interactions in a range of systems [390,671–673]. While the application of molecular simulations to identify transient helices in disordered regions can suffer from forcefield inaccuracies, it is worth noting that in prior work we have found good agreement between experimental and simulated secondary structure analysis across a range of systems explored in an analogous manner [80,93,674,675].

Transient helix H2 (in the NTD) and H3 (in the LINK) flank the RBD and organize a set of arginine residues to face the same direction (**Fig. 2H and 3F**). Both the NTD and LINK have been shown to drive RNA binding, such that we propose these helical arginine-rich motifs (ARMs) may engage in both nonspecific binding and may also contribute to RNA specificity, as has been proposed previously [639,676,677]. The serine-arginine SR-region (which includes H3) has been previously identified as engaging in interaction with a structured acidic helix in Nsp3 in the model coronavirus MHV, consistent with an electrostatic helical interaction [678,679]. Recent NMR data also shows excellent agreement with our results, identifying a transient helix that shows 1:1 overlap with H3 [392].The SR-region is necessary for recruitment to replication-transcription centers (RTCs) in MHV, and also undergoes phosphorylation, setting the stage for a complex regulatory system awaiting exploration [680,681].

Transient helix H4 (in the LINK, **Fig. 3F**) was previously predicted bioinformatically and identified as a conserved feature across different coronaviruses, in agreement with our own secondary structure predictions (**Fig. S19**) [639]. Furthermore, the equivalent region was identified in SARS coronavirus as a nuclear export signal (NES), such that we suspect this too is a classical Crm1-binding leucine-rich NES [682]. Jack *et al.* identified helix H4 as enriched for homotypic cross-links in

the context of droplets, supporting a model in which this region promotes protein:protein interactions, an interpretation corroborated by hydrogen-deuterium exchange mass spectrometry on RBD-LINK in the dilute phase [634,636].

Concerning the CTD, two transient helices are identified, helix H5 and H6. While transient helix H5 is weakly populated, the positive charge associated with this region may make it critical for protein:RNA interaction, a result strongly supported by the observation that deletion of this region ablates protein:RNA phase separation [634]. Transient helix H6 is an amphipathic helix with a highly hydrophobic face (**Fig. 4H**). Recent hydrogen-deuterium exchange mass spectrometry also identified H6 [651]. Residues in this region have previously been identified as mediating M-protein binding in other coronaviruses, such that we propose H6 underlies that interaction [391,683–685]. Recent work has also identified amphipathic transient helices in disordered proteins as interacting directly with membranes, such that an additional (albeit entirely speculative) role could involve direct membrane interaction, as has been observed in other viral phosphoproteins [686,687].

As a final note, while these helices are conserved between SARS, SARS-CoV-2, and in many bat-coronaviruses, they are less well conserved in MHV and MERS, suggesting these regions are malleable over evolution (Fig.S1/3/5).

5.11 The physiological relevance of nucleocapsid protein phase separation in SARS-CoV-2 physiology

Our work has revealed that SARS-CoV-2 N protein undergoes phase separation with RNA when reconstituted *in vitro*. The solution environment and types of RNA used in our experiments are very

different from the cytoplasm and viral RNA. However, similar results have been obtained in published and unpublished work by several other groups under a variety of conditions, including via *in cell* experiments [391,392,442,464,634–637]. Taken together, these results demonstrate that N protein *can* undergo *bona fide* phase separation, and that N protein condensates *can* form in cells. Nevertheless, the complexity introduced by multidimensional linkage effects *in vivo* could substantially influence the phase behavior and composition of condensates observed in the cell [408,664,688]. Of note, the regime we have identified in which phase separation occurs (**Fig. 5**) is remarkably relatively narrow, consistent with a model in which single-genome condensates for virion assembly are favored over larger multi-genome droplets.

Does phase separation play a physiological role in SARS-CoV-2 biology? Phase separation has been invoked or suggested in a number of viral contexts to date [364,369,371,375,689–691]. In SARS-CoV-2, one possible model suggests phase separation may drive recruitment of components to viral replication sites, although how this dovetails with the fact that replication occurs in double-membrane bound vesicles (DMVs) remains to be explored [392,692]. An alternative (and non-mutually exclusive) model is one in which phase separation catalyzes nucleocapsid polymerization, as has been proposed in elegant work on measles virus [390]. Here, the process of phase separation is decoupled from genome packaging, where gRNA condensation occurs through association with a helical nucleocapsid. If applied to SARS-CoV-2, such a model would suggest that (1) initially N protein and RNA phase separate in the cytosol, (2) some discrete pre-capsid state forms within condensates and, (3) upon maturation, the pre-capsid is released from the condensate and undergoes subsequent virion assembly by interacting with the membrane-bound M, E, and S structural proteins at the ER-Golgi intermediate compartment (ERGIC). While this model is attractive it places a number of constraints

on the physical properties of this pre-capsid, not least that the ability to escape the parent condensate dictates that the assembled pre-capsid must interact less strongly with the condensate components than in the unassembled state. This requirement introduces some thermodynamic complexities: how is a pre-capsid state driven to assemble if it is necessarily less stable than the unassembled pre-capsid, and how is incomplete or abortive pre-capsid formation avoided if – as assembly occurs – the pre-capsid becomes progressively less stable?

A phase separation and assembly model raises additional questions, such as the origins of specificity for recruitment of viral proteins and viral RNA, the kinetics of pre-capsid-assembly within a large condensate, and preferential packaging of gRNA over sub-genomic RNA. None of these questions are unanswerable, nor do they invalidate this model, but they should be addressed if the physiological relevance of large cytoplasmic condensates is to be further explored in the context of virion assembly.

Our preferred interpretation is that N protein has evolved to drive genome compaction for packaging (**Fig. 7**). In this model, a single-genome condensate forms through N protein gRNA interaction, driven by a small number of high-affinity sites. This (meta)-stable single-genome condensate undergoes subsequent maturation, leading to virion assembly. In this model, condensate-associated N proteins are in exchange with a bulk pool of soluble N protein, such that the interactions that drive compaction are heterogeneous and dynamic. Our model provides a physical mechanism in good empirical agreement with data for N protein oligomerization and assembly [693–695]. Furthermore, the resulting condensate is then in effect a multivalent binder for M protein, which interacts with N directly, and may drive membrane curvature and budding in a manner similar to

that proposed by Bergeron-Sandoval and Michnick (though with a different directionality of the force) and in line with recent observations from cryo-electron tomography (cryoET) [692,696–698]

An open question pertains to specificity of packaging gRNA while excluding other RNAs. One possibility is for two high-affinity N-protein binding sites to flank the 5' and 3' ends of the genome, whereby only RNA molecules with both sites are competent for compaction. A recent map of N protein binding to gRNA has revealed high-affinity binding regions at the 5' and 3' ends of the gRNA, in good agreement with this qualitative prediction [442]. Alternatively, only gRNA condensates may possess the requisite valency for N protein binding to drive virion assembly through interaction with M protein at the cytoplasmic side of the ERGIC, offering a physical selection mechanism for budding.

Genome compaction through dynamic multivalent interactions would be especially relevant for coronaviruses, which have extremely large single-stranded RNA genomes. This is evolutionarily appealing, in that as the genome grows larger, compaction becomes increasingly efficient, as the effective valence of the genome is increased [359,398]. The ability of multivalent disordered proteins to drive RNA compaction has been observed previously in various contexts [324,699]. Furthermore, genome compaction by RNA binding protein has been proposed and observed in other viruses [695,700,701], and the SARS coronavirus N protein has previously been shown to act as an RNA chaperone, an expected consequence of compaction to a dynamic single-RNA condensate that accommodates multiple N proteins with a single RNA [324,702]. Furthermore, previous work exploring the ultrastructure of phase separated condensates of G3BP1 and RNA through simulations and

cryoET revealed a beads-on-a-string type architecture, mirroring recent results for obtained from cryo-electron tomography of SARS-CoV-2 virions [473,692].

N protein has been shown to interact directly with a number of proteins studied in the context of biological phase separation which may influence assembly *in vivo* [80,626,635,663,703]. In particular, G3BP1 – an essential stress-granule protein that undergoes phase separation – was recently shown to co-localize with overexpressed N protein [392,473,663,704,705]. G3BP1 interaction may be part of the innate immune response, leading to stress-granule formation, or alternatively N protein may attenuates the stress response by sequestering G3BP1, depleting the cytosolic pool, and preventing stress granule formation, as has been shown for HIV-1 and very recently proposed explicitly for SARS-CoV-2 [690,705].

Our model is also in good empirical agreement with recent observations made for other viruses[706]. Taken together, we speculate that viral packaging may -in general- involve an initial genome compaction through multivalent protein:RNA and protein:protein interactions, followed by a liquid-to-solid transition in cases where well-defined crystalline capsid structures emerge. Liquid-to-solid transitions are well established in the context of neurodegeneration with respect to disease progression [707–709]. Here we suggest nature is leveraging those same principles as an evolved mechanism for monodisperse particle assembly.

Regardless of if phase separated condensates form inside cells, all available evidence suggests phase separation is reporting on a physiologically important interaction that underlies genome compaction (**Fig. 6L**). With this in mind, from a biotechnology standpoint, phase separation may be a convenient readout for *in vitro* assays to interrogate protein:RNA interaction. Regardless of which

model is correct, N protein:RNA interaction is key for viral replication. As such, phase separation provides a macroscopic reporter on a nanoscopic phenomenon, in line with previous work [80,398,404,710]. In this sense, we propose the therapeutic implications of understanding and modulating phase separation here (and elsewhere in biology) are conveniently decoupled from the physiological relevance of actual, large phase separated liquid droplets, but instead offer a window into the underlying physical interactions that lead to condensate formation [634].

5.12 The physics of single polymer condensates

Depending on the molecular details, single-polymer condensates may be kinetically stable (but thermodynamically unstable, as in our model simulations) or thermodynamically stable. Delineation between these two scenarios will depend on the nature, strength, valency and anisotropy of the interactions. It is worth noting that from the perspective of functional biology, kinetic stability may be essentially indistinguishable from thermodynamic stability, depending on the lifetime of a metastable species.

It is also important to emphasize that at higher concentrations of N protein and/or after a sufficiently long time period we expect robust phase separation with viral RNA, regardless of the presence of a symmetry-breaking site. Symmetry breaking is achieved when the apparent local concentration of N protein (from the perspective of gRNA) is substantially higher than the actual global concentration. As effective local and global concentrations approach one another, the entropic cost of intra-molecular interaction is outweighed by the availability of inter-molecular partners. On a practical note, if the readout in question is the presence/absence of liquid droplets, a high-affinity site may be observed as a shift in the saturation concentration which, confusingly, could

either suppress or enhance phase separation. Further, if single-genome condensates are kinetically stable and driven through electrostatic interactions, we would expect a complex temperature dependence, in which larger droplets are observed at higher temperature (up to some threshold). Recent work is showing a strong temperature-dependence of phase separation is consistent with these predictions [442].

Finally, we note no reason to assume single-RNA condensates should be exclusively the purview of viruses. RNAs in eukaryotic cells may also be processed in these types of assemblies, as opposed to in large multi-RNA RNPs. The role of RNA:RNA interactions both here and in other systems is also of particular interest and not an aspect explored in our current work, but we anticipate may play a key role in the relevant biology.

5.13 Methods

All-atom simulations

All-atom Monte Carlo simulations were performed with the ABSINTH implicit solvent model (abs_3.2_opls.prm) and CAMPARI simulation engine (V2) (http://campari.sourceforge.net/) 57,137 with the solution ion parameters of Mao et al.138. Simulations were performed using movesets and Hamiltonian parameters as reported previously 72,139. All simulations were performed in sufficiently large box sizes to prevent finite size effects (where box size varies from system to system). For simulations with IDRs in isolation all degrees of freedom available in CAMPARI are sampled. For simulations with folded domains with IDRs, the backbone dihedral angles in folded domains are not sampled, such that folded domains remain structurally fixed

(although sidechains are fully sampled). The IDR has backbone and sidechain degrees of freedom sampled. Simulation sequences used are defined in SI Table S7.

All-atom molecular dynamics simulations were performed using GROMACS (GROMACS 2019 locally, version 5.0.4 on Folding@Home), using the FAST algorithm in conjunction with the Folding@home platform 58,140,141. Post-simulation analysis was performed with Enspara 142. For additional simulation details see the supplementary information.

Coarse-grained polymer simulations

Coarse-grained Monte Carlo simulations were performed using the PIMMS simulation engine 143. All simulations were performed in a 70 x 70 x 70 lattice-site box. The results averaged over the final 20% of the simulation to give average values at equivalent states. The polymer species is represented as a 61-residue polymer with either a central high-affinity binding site or not. The binder is a 2-bead species. All simulations shown in Fig. 6 were run for 20 x 109 Monte Carlo steps, with four independent replicas. Bead interaction strengths were defined as shown in Fig. 6A. For additional simulation details see SI.

Protein Expression, purification, and labeling

SARS-CoV-2 Nucleocapsid protein (NCBI Reference Sequence: YP_009724397.2) including an N term extension containing His9-HRV 3C protease site was cloned into the BamHI EcoRI sites in the MCS of pGEX-6P-1 vector (GE Healthcare). Site-directed mutagenesis was performed on the His9-SARS-CoV-2 Nucleocapsid pGEX vector to create the N protein constructs (SI Table S1) and sequences were verified using Sanger sequencing. All variants were expressed recombinantly in BL21

Codon-plus pRIL cells (Agilent) or Gold BL21(DE3) cells (Agilent) and purified using a FF HisTrap column. The GST-His9-N tag was then cleaved using HRV 3C protease and further purified to remove the cleaved tag. Finally, purified N protein variants were analyzed using SDS-PAGE and verified by electrospray ionization mass spectrometry (LC-MS). Activity of the protein was assessed by testing whether the protein is able to bind and condense nucleic acids (see phase-separation experiments) as well as to form dimers (see oligomerization in SI).

All Nucleocapsid variants were labeled with Alexa Fluor 488 maleimide and Alexa Fluor 594 maleimide (Molecular Probes) under denaturing conditions following a two-step sequential labeling procedure (see SI).

Single-molecule fluorescence spectroscopy

Single-molecule fluorescence measurements were performed with a Picoquant MT200 instrument (Picoquant, Germany). FRET experiments were performed by exciting the donor dye with a laser power of 100 μW (measured at the back aperture of the objective). For pulsed interleaved excitation of donor and acceptor, the power used for exciting the acceptor dye was adjusted to match the acceptor emission intensity to that of the donor (between 50 and 70 mW). Single-molecule FRET efficiency histograms were acquired from samples with protein concentrations between 50 pM and 100 pM and the population with stoichiometry corresponding to 1:1 donor:acceptor labeling was selected. Trigger times for excitation pulses (repetition rate 20 MHz) and photon detection events were stored with 16 ps resolution. For FRET-FCS, samples of double-labeled protein with a concentration of 100 pM were excited by either the diode laser or the supercontinuum laser at the powers indicated above.

All samples were prepared in 50 mM Tris pH 7.32, 143 mM β-mercaptoethanol (for photoprotection), 0.001% Tween 20 (for limiting surface adhesion) and GdmCl at the reported concentrations. All measurements were performed in uncoated polymer coverslip cuvettes (Ibidi, Wisconsin, USA) and custom-made glass cuvette coated with PEG (see SI). Each sample was measured for at least 30 min at room temperature (295 ± 0.5 K).

5.14 Acknowledgements

We thank Amy Gladfelter, Christiane Iserman, Christine Roden, Ahmet Yildiz, Amanda Jack, Luke Ferro, Steve Michnick, Pascale Legault, and Jim Omichinski for sharing data and extensive discussion. We also thank Rohit Pappu for placing our groups in contact with one another. We thank the labs of John Cooper, Carl Frieden, and Silvia Jansen for providing some of the reagents we have used in this work. We thank Ben Schuler and Daniel Nettels for developing, maintaining, and sharing with us the software package used to analyze the single-molecule data. J.C. and J.J.A are supported by NIGMS R25 IMSD Training Grant GM103757. We are grateful to the citizen-scientists of Folding@home for donating their computing resources. G.R.B holds an NSF CAREER Award MCB-1552471, NIH R01GM12400701, a Career Award at the Scientific Interface from the Burroughs Wellcome Fund, and a Packard Fellowship for Science and Engineering from The David and Lucile Packard Foundation. A.S holds NIH grant R01AG062837. A.S.H. is supported by the Longer Life Foundation: An RGA/Washington University Collaboration.

5.15 Data availability

Data supporting the findings in this paper are available from the corresponding authors upon request. All-atom simulation data for Monte Carlo simulations and disorder prediction info are provided at https://github.com/holehouse-lab/supportingdata/tree/master/2021/cubuk_nucleocapsid_2021. Simulations and simulation analysis were performed with open source tools (http://campari.sourceforge.net/, https://camparitraj.readthedocs.io/, http://mdtraj.org/, https://www.gromacs.org/) and Folding@Home data are available for further analysis at https://covid.molssi.org//org-contributions/#folding--home.

5.16 Competing Interests

A.S.H. is a scientific consultant with Dewpoint Therapeutics. This affiliation in no way influenced the content of this study. All other authors declare no competing interests.

5.17 Figures

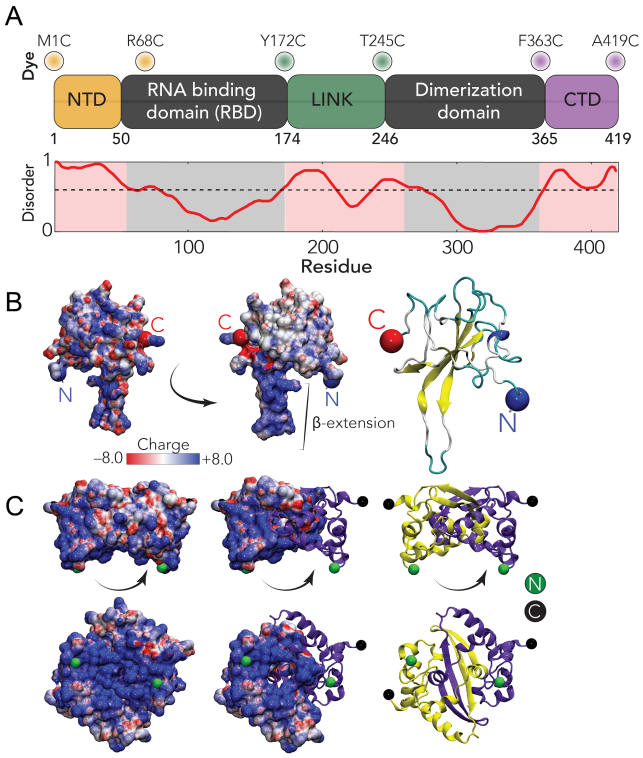


Figure 1. Sequence and structural summary of N protein

A. Domain architecture of the SARS-CoV-2 N protein with disorder prediction performed using IUPred2A [711]. Dye positions used in this study are annotated across the top, disorder prediction calculated across the bottom. The specific positions were selected such that fluorophores are sufficiently close to be in the dynamic range of FRET measurements. Labeling was achieved using cysteine mutations and thiol-maleimide chemistry. **B.** Structure of the SARS-CoV-2 RNA binding domain (RBD) (PDB: 6yi3). Center and left: coloured based on surface potential calculated with the Adaptive Poisson Boltzmann Method [712], revealing the highly basic surface of the RBD. Right: ribbon structure with N- and C-termini highlighted. **C.** Dimer structure of the SARS-CoV-2 dimerization domain (PDB: 6yun). Center and left: coloured based on surface potential, revealing the highly basic surface. Right: ribbon structure with N- and C-termini highlighted.

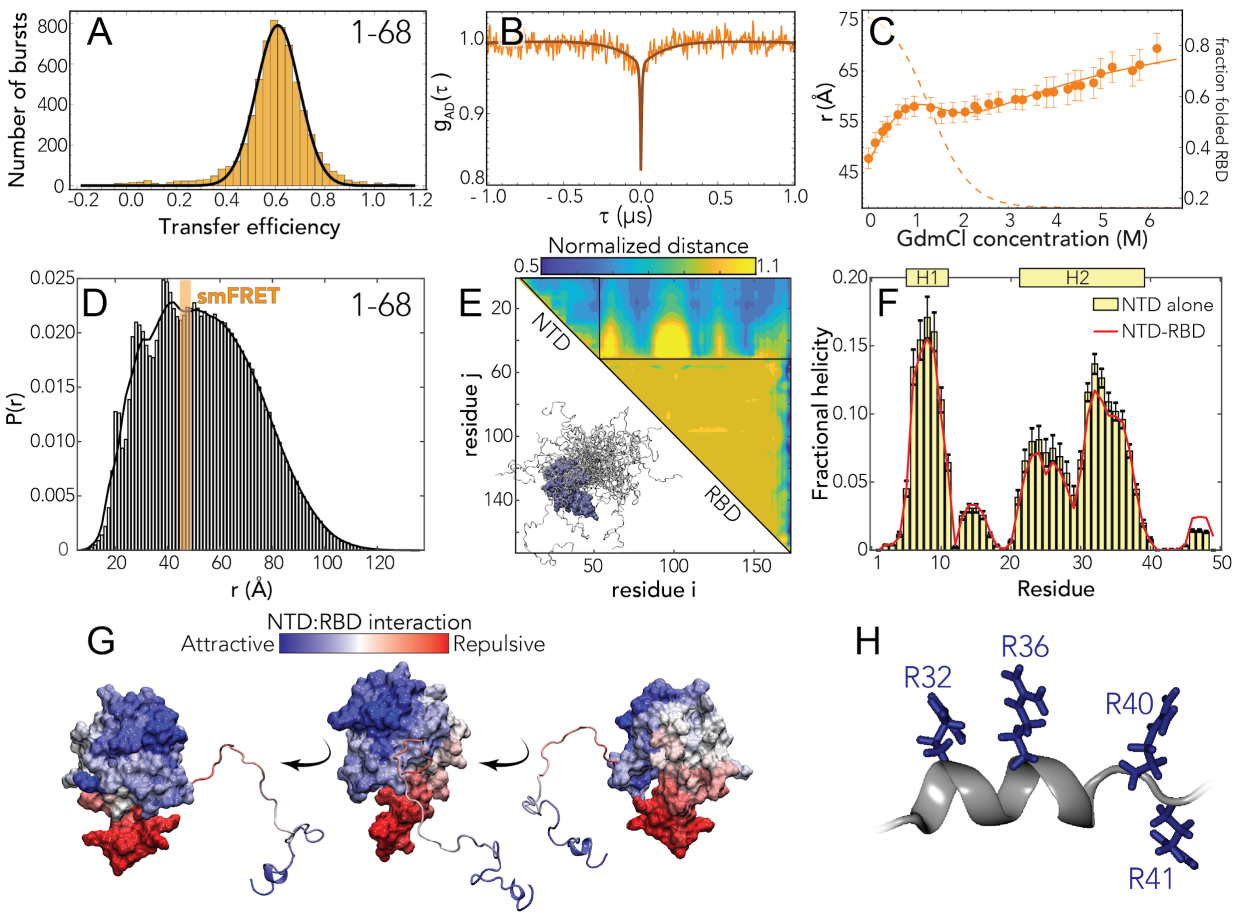


Figure 2. The N-terminal domain (NTD FL) is disordered with residual helical motifs

A. Histogram of the transfer efficiency distribution measured across the labeling positions 1 and 68 in the context of the full-length protein, under aqueous buffer conditions (50 mM Tris buffer). **B.** Donor-acceptor cross-correlation measured by ns-FCS (see **SI**). The observed anticorrelated rise is the characteristic signature of FRET dynamics and the timescale associated is directly related to the reconfiguration time of the probed segment. **C.** Root-mean-square interdye distance as extracted from single-molecule FRET experiments across different concentrations using a Gaussian chain distribution, examining residues 1-68 in the context of the full-length protein. The full line represents a fit to the model in **Eq. S7**, which accounts for denaturant binding (see **Table S2**) and

unfolding of the folded RBD. The dashed line represents the estimate of folded RBD across different denaturant concentrations based on **Eq. S8**. Error bars represent propagation ± 0.03 systematic error in measured transfer efficiencies (see **SI**). **D.** All-atom simulations of the NTD in the context of RBD reveal good agreement with smFRET-derived average distances. The peaks on the left shoulder of the histogram are due to persistent NTD-RBD interactions in a small subset of simulations. **E.** Normalized distance maps (scaling maps) quantify heterogeneous interaction between every pair of residues in terms of average distance normalized by distance expected for the same system if the IDR had no attractive interactions (the excluded volume limit [306]). Both repulsive (yellow) and attractive (blue) regions are observed for NTD-RBD interactions. **F.** Transient helicity (residues 5-11 and 21-39) in the NTD in isolation or in the context of the RBD. Perfect profile overlap suggests interaction between the NTD and the RBD does not lead to a loss of helicity. Error bars are standard error of the mean calculated from forty independent simulations. **G.** Projection of normalized distances onto the folded domain reveals repulsion is through electrostatic interaction (positively charged NTD is repelled by the positive face of the RBD, which is proposed to engage in RNA binding) while attractive interactions are between positive, aromatic, and polar residues in the NTD and a slightly negative and hydrophobic surface on the RBD (see **Fig. 1B**, center). **H.** The C-terminal half of transient helicity in H2 encodes an arginine-rich surface.

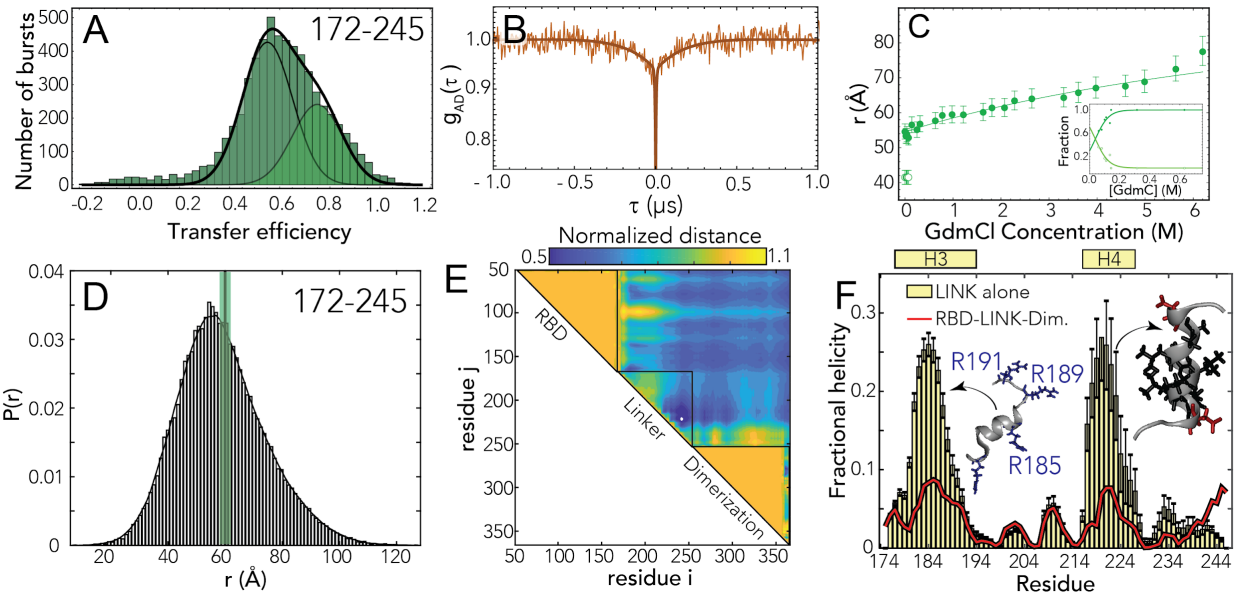


Figure 3. The RNA binding domain (RBD) and dimerization domains are interconnected by a flexible disordered linker (LINK)

A. Histogram of the transfer efficiency distribution measured across the labeling positions 172 and 245 in the context of the full-length protein, under aqueous buffer conditions. **B.** Donor-acceptor cross-correlation measured by ns-FCS (see SI). The observed anticorrelated rise is the characteristic signature of FRET dynamics and the timescale associated is directly related to the reconfiguration time of the probed segment. **C.** Interdye distance as extracted from single-molecule FRET experiments across different denaturant concentrations. The full line represents a fit to the model in **Eq. S6**, which accounts for denaturant binding. The inset provides an estimate of the fraction of each population in the low GdmCl concentration regime. Error bars are the propagation of ± 0.03 systematic error in measured transfer efficiencies (see SI). **D.** Inter-residue distance distributions calculated from simulations (histogram) show good agreement with distances inferred from single-molecule FRET measurements (green bar). **E.** Scaling maps reveal repulsive interactions between the N- and C-terminal regions of the LINK with the adjacent folded domains. We also observe

relatively extensive intra-LINK interactions around helix H4 (see **Fig. 3F**). **F.** Two transient helices are observed in the linker (residues 177-194 and 216-227). The N-terminal helix H3 overlaps with part of the SR-region and orientates three arginine residues in the same direction, analogous to behavior observed for H2 in the NTD. The C-terminal helix H4 overlaps with a Leu/Ala rich motif and may be a conserved nuclear export signal (see Discussion). Error bars are standard errors of the mean calculated from thirty independent simulations.

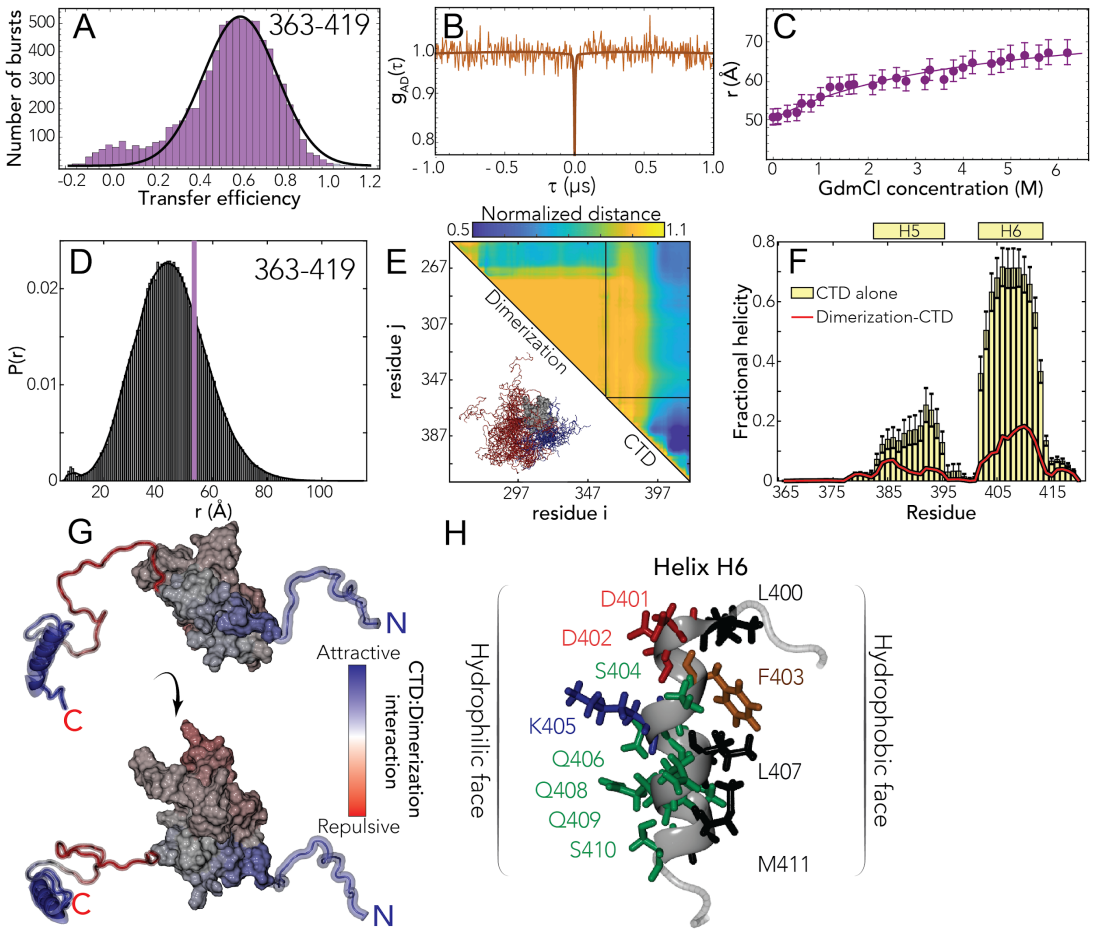


Figure 4. The C-terminal domain (CTD) is disordered, engages in transient interaction with the dimerization domain, and contains a putative helical binding motif

A. Histogram of the transfer efficiency distribution measured across the labeling positions 363 and 419 in the context of the full-length protein, under aqueous buffer conditions. **B.** Donor-acceptor cross-correlation measured by ns-FCS (see SI). The flat correlation indicates a lack of dynamics in the studied timescale or the coexistence of two populations in equilibrium whose correlations (one correlated and the other anticorrelated) compensate each other. **C.** Interdye distance as extracted from single-molecule FRET experiments across different denaturant concentrations. The full line represents a fit to the model in **Eq. S6**, which accounts for denaturant binding. Error bars are the propagation of ± 0.03 systematic error in measured transfer efficiencies (see SI). **D.** Inter-residue

distance distributions calculated from simulations (histogram) show good agreement with distances inferred from single-molecule FRET measurements (purple bar). **E.** Scaling maps describe the average inter-residue distance between each pair of residues, normalized by the distance expected if the CTD behaved as a self-avoiding random coil. H6 engages in extensive intra-CTD interactions and also interacts with the dimerization domain. We observe repulsion between the dimerization domain and the N-terminal region of the CTD. **F.** Two transient helices (H5 and H6) are observed in the CTD (residues 383-396 and 402-415). Both show a reduction in population in the presence of the dimerization domain at least in part because the same sets of residues engage in transient interactions with the dimerization domain. Error bars are standard error of the mean calculated from forty independent simulations. **G.** The normalized distances are projected onto the surface to map CTD-dimerization interaction. The helical region drives intra-molecular interaction, predominantly with the N-terminal side of the dimerization domain. **H.** Helix H6 is an amphipathic helix with a polar/charged surface (left) and a hydrophobic surface (right).

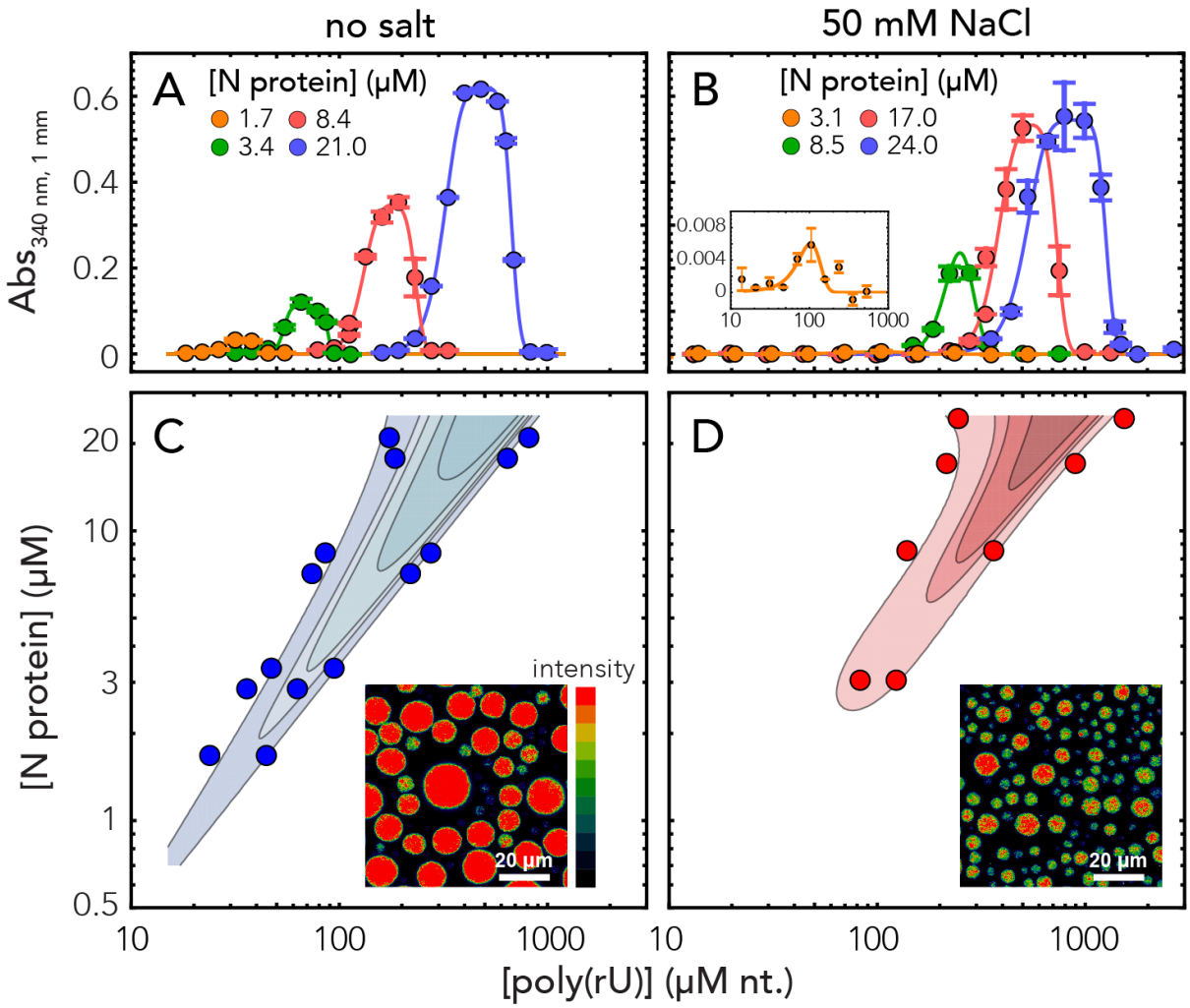


Figure 5. Nucleocapsid protein undergoes phase separation with RNA. A-B

Appearance of solution turbidity upon mixing was monitored to determine the concentration regime in which N protein and poly(rU) undergo phase separation. Representative turbidity titrations with poly(rU) in 50 mM Tris, pH 7.5 (HCl) at room temperature, in the absence of added salt (A) and in the presence of 50 mM NaCl (B), at the indicated concentrations of N protein. Points and error bars represent the mean and standard deviation of 2 (absorbance < 0.005) and 4 (absorbance ⩾ 0.005) consecutive measurements from the same sample. Solid lines are simulations of an empirical equation fitted individually to each titration curve (see SI). An inset is provided for the titration at

3.1 μM N protein in 50 mM NaCl to show the small yet detectable change in turbidity on a different scale. **C-D.** Projection of phase boundaries for poly(rU) and N protein mixtures highlights a re-entrant behavior, as expected for phase-separations induced by heterotypic interactions. Turbidity contour lines are computed from a global fit of all titration curves (see SI). **Insets**: confocal fluorescence images of droplets doped with fluorescently labeled N protein. Total concentrations are 22 μM N protein, 0.5 nM labeled N protein and 0.54 mM nt. poly(rU). At a higher salt concentration, a lower concentration of protein in the droplet is detected.

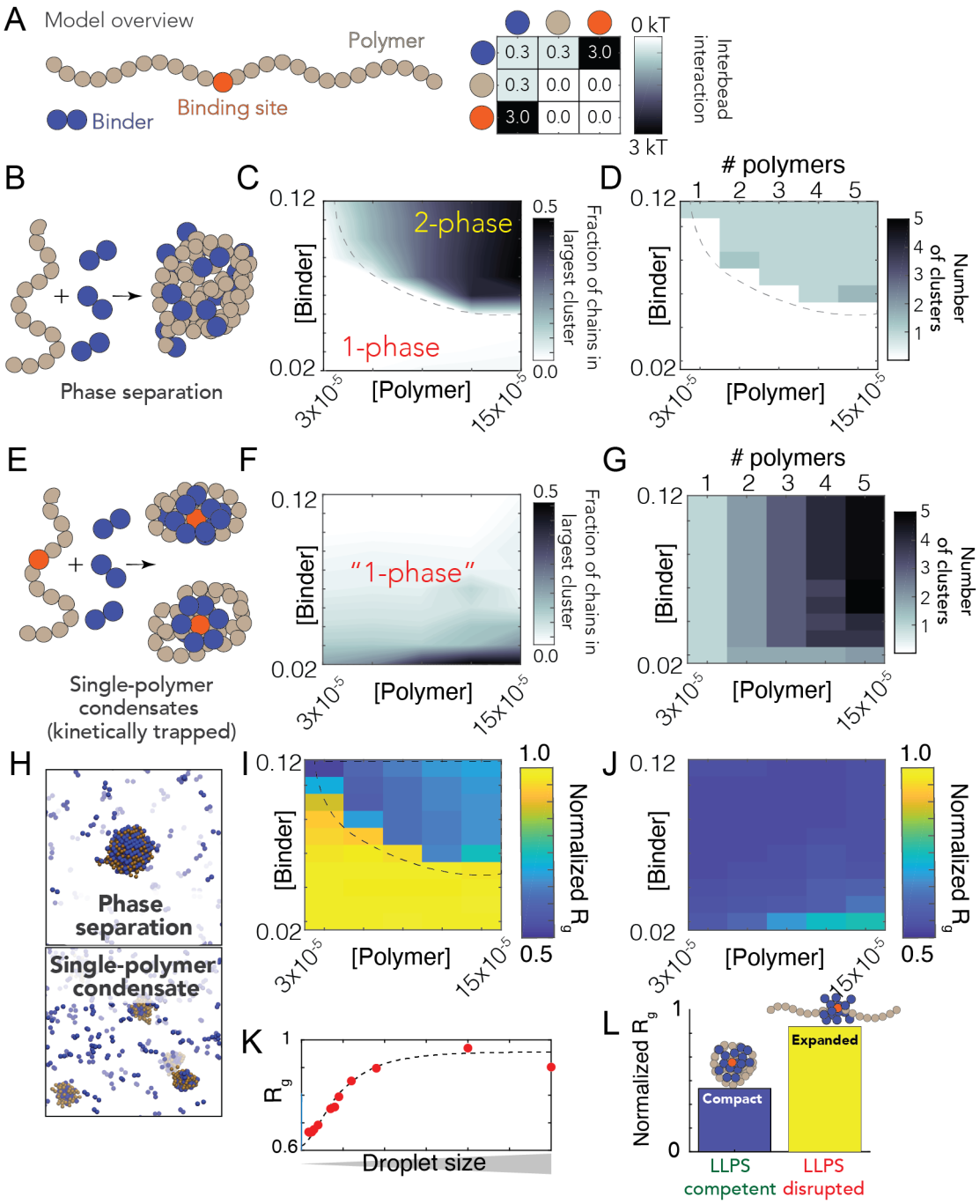

A
Model overview
Polymer
Binder
Binding site
0.3 0.3 3.0
0.3 0.0 0.0
3.0 0.0 0.0
0 kT
3 kT
Interbead interaction
B
Phase separation
C
2-phase
1-phase
[Binder]
[Polymer]
Fraction of chains in largest cluster
D
polymers
Number of clusters
E
Single-polymer condensates (kinetically trapped)
F
"1-phase"
G
H
Phase separation
Single-polymer condensate
I
J
Normalized R_g
K
R_g
Droplet size
L
Compact
Expanded
LLPS competent
LLPS disrupted

Figure 6. A simple polymer suggests symmetry breaking can promote single-polymer condensates over multi-polymer assemblies

A. Summary of our model setup, which involves long polymers (61 beads per molecules) or short binders (2 beads per molecules). Each bead is multivalent and can interact with every adjacent lattice site. The interaction matrix to the right defines the pairwise interaction energies associated with each of the bead types. **B.** Concentration dependent assembly behavior for polymers lacking a high-affinity binding site. Schematic showing polymer architecture (brown) with binder (blue). **C.** Phase diagram showing the concentration-dependent phase regime - dashed line represents the binodal (phase boundary) and is provided to guide the eye. **D.** Analysis in the same 2D space as panel C, assessing the number of droplets at a given concentration. When phase separation occurs, a single droplet appears in almost all cases. **E.** Concentration dependent assembly behavior for polymers with a high-affinity binding site (red bead). **F.** No large droplets are formed in any of the systems, although multiple polymer:binder complexes form. **G.** The number of clusters observed matches the number of polymers in the system - i.e., each polymer forms an individual cluster. **H.** Simulation snapshots from equivalent simulations for polymers with (top) or without (bottom) a single high-affinity binding site. **I.** Polymer dimensions in the dense and dilute phase (for the parameters in our model) for polymers with no high-affinity binding site. Note that compaction in the dense phase reflects finite-size effects, as addressed in panel K, and is an artefact of the relatively small droplets formed in our systems (relative to the size of the polymer). The droplets act as a bounding cage for the polymer, driving their compaction indirectly. **J.** Polymer dimensions across the same concentration space for polymers with a single high-affinity binding site. Across all concentrations, each individual polymer is highly compact. **K.** Compaction in the dense phase (panel I) is due to small droplets. When droplets are sufficiently large, we observe chain expansion, as expected from

standard theoretical descriptions. **L.** Simulations performed under conditions in which nonspecific interactions between binder and polymer are reduced (interaction strength = 0 kT). Under these conditions phase separation is suppressed. Equivalent simulations for polymers with a high-affinity site reveal these chains are no longer compact. As such, phase separation offers a readout that - in our model - maps to single-polymer compaction.

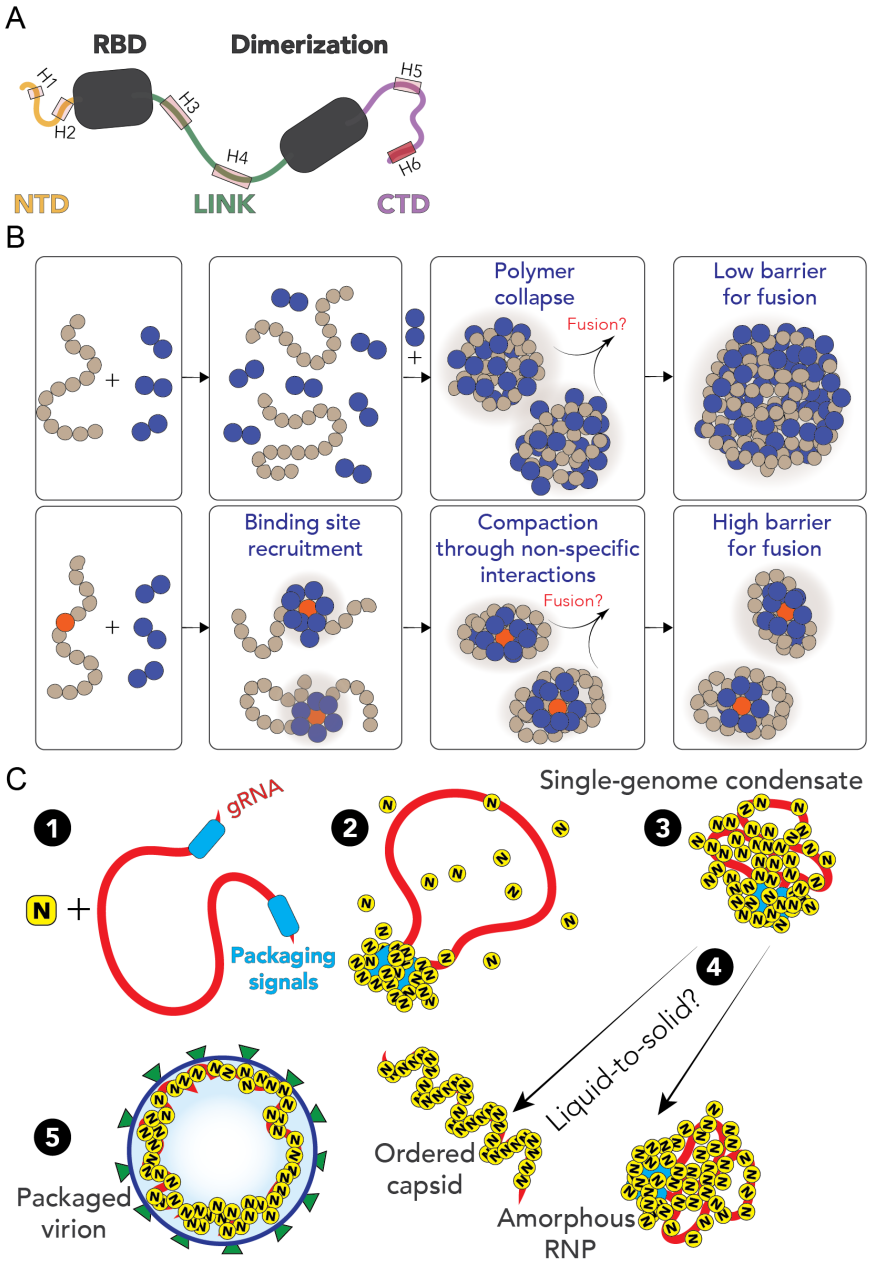


Figure 7. Summary and proposed model

A. Summary of results from single-molecule spectroscopy experiments and all-atom simulations. All three predicted IDRs are disordered, highly flexible, and house a number of putative helical binding regions which overlap with subregions identified previously to drive N protein function. **B.** Overview of general symmetry breaking model. For homopolymers, local collapse leads to single-polymer condensates with a small barrier to fusion, rapidly assembling into large multi-polymer condensates. When one (or a small number of) high-affinity sites are present, local clustering of binders at a lower concentration organize the polymer such that single-polymer condensates are kinetically stable. **C.** Proposed model for SARS-CoV-2 genome packaging. **(1)** Simplified model of SARS-CoV-2 genome with a pair of packaging region at the 5' and 3' end of the genome **(2)** N protein preferentially binds to packaging signal regions in the genome, leading to a local cluster of N protein at the packaging signal RNA. **(3)** The high local concentration of N protein drives condensation of distal regions of the genome, forming a stable single-genome condensate. **(4)** Single-genome condensates may undergo subsequent maturation through a liquid-to-solid (crystallization) transition to form an ordered crystalline capsid, or solidify into an amorphous ribonuclear particle (RNP), or some combination of the two. While in some viruses an ordered capsid clearly forms, we favor a model in which the SARS-CoV-2 capsid is an amorphous RNP. Compact single-genome condensates ultimately interact with E, S and M proteins at the membrane, whose concerted action leads to envelope formation around the viral RNA and final virion packaging

5.18 Supplementary Information

Sequence Analysis

Disorder prediction was performed using IUPred2.0, with additional analysis and sequence parsing done with localCIDER and protfasta, respectively [455,711,713].

Amino acid sequence of the N protein used in simulations. Highlighted regions delineate folded domains. Underline bolded residues highlighted in red identify the sites of dyes for single-molecule fluorescence experiments.

1 **M**SDNGPQNQR NAPRITFGGP SDSTGSNQNG ERSGARSKQR RPQGLPNNTA
51 SWFTALTQHG KEDLKFP**R**GQ GVPINTNSSP DDQIGYYRRA TRRIRGGDGK
101 MKDLSPRWYF YYLGTGPEAG LPYGANKDGI IWVATEGALN TPKDHIGTRN
151 PANNAAIVLQ LPQGTTLPKG F**Y**AEGSRGGS QASSRSSSRS RNSSRNSTPG
201 SSRGTSPARM AGNGGDAALA LLLLDRLNQL ESKMSGKGQQ QQGQ**T**VTKKS
251 AAEASKKPRQ KRTATKAYNV TQAFGRRGPE QTQGNFGDQE LIRQGTDYKH
301 WPQIAQFAPS ASAFFGMSRI GMEVTPSGTW LTYTGAIKLD DKDPNFKDQV
351 ILLNKHIDAY KT**F**PPTEPKK DKKKKADETQ ALPQRQKKQQ TVTLLPAADL
401 DDFSKQLQQS MSSADSTQ**A**

A complete list of constructs is presented in **Table S1**.

Simulation Methods

All-atom Monte Carlo Simulations.

All Monte Carlo simulations were performed using the CAMPARI simulation engine and ABSINTH implicit solvent model (abs_3.2_opls.prm) using the monovalent ion parameters derived by Mao et al.[714]. All simulations were performed at 330 K and at 15 mM NaCl, as have been used previously in various systems [80,91,602,603]. The base keyfile used for all Monte Carlo simulations can be found at https://github.com/holehouse-lab/supportingdata/.

Simulation analysis was performed with MDTraj and camparitraj (http://ctraj.com/) [715]. For IDR only simulations, all degrees of freedom were fully sampled (backbone and sidechain dihedral angles and rigid-body positions) as is standard in CAMPARI Monte Carlo simulations [104]. For simulations of IDRs in the context of folded domains, the backbone dihedral angles of the folded domains were held fixed, while all sidechains were fully sampled, as were backbone dihedral angles for the disordered regions, as applied previously [716]. The folded state starting structures were obtained from PDB structures obtained from molecular dynamics (MD) simulations (see below for more details).

For IDR-only simulations, 30-40 independent simulations were run generating final ensembles of 40-60 K conformations. For simulations of IDRs in the context of folded domains, the number of independent simulations and the length of the simulation varied. For the NTD-RBD simulations 400 independent simulations were run using an initial molecular dynamics based sampling approach to obtain starting states for the folded domain, with 2 independent simulations per starting seed from MD simulations (see methods below) leading to a final ensemble of ~400 K conformations (24 M steps per simulation). For the RBD-LINK-dimerization construct, thirty-five independent simulations were run for a final ensemble of 32 K conformers (66 M steps per simulation). For the dimerization-CTD construct 200 independent simulations were run providing a final ensemble of 40

K conformations (66 M steps per simulation). For a complete description of simulation details see **Table S5**, and **Table S7** for a list of sequences.

For both the NTD-RBD construct and the DIM-CTD construct, we used a sequential sampling approach in which long timescale MD simulations of the RBD in isolation performed on the Folding@home distributed computing platform were first used to generate hundreds of starting conformations [284]. Those RBD conformations were then used as starting structures for independent all-atom Monte Carlo simulations. Monte Carlo simulations were performed with the ABSINTH forcefield in which the RBD backbone dihedral angles are held fixed but the NTD is fully sampled, as are RBD sidechains. For simulations of the monomeric dimerization domain we discovered that as a monomer, the first 21 residues of the dimerization domain appear disordered, in agreement with sequence predictions (**Fig. 1A**) but in contrast to their behavior in the dimeric structure (**Fig. 1C**). As a result, we choose to also model these residues as fully disordered.

The RBD starting structure used was taken as the first chain extracted from the 6VYO PDB crystal structure, which is structurally almost identical to many of the 6YI3 NMR model shown in **Fig. 1A.** At the time that our work on this project began the 6VYO structure was the only available structure of the RBD. Irrespective, the extensive molecular dynamics (MD) simulation run prior to our Monte Carlo simulations are such that any small difference in starting structure are negated by many microseconds of simulation sampling.

To generate the monomeric starting structure of the dimerization domain, we first built a homology model of the SARS-CoV-2 dimerization dimer from the NMR structure of the SARS dimerization structure (PDB: 2JW8) using SWISS-MODEL [642,717]. We chose this strategy because at the time, no dimerization structure existed, a situation that has since resolved itself [650,651]. Nevertheless, the SARS

and SARS-CoV-2 dimerization domains are essentially identical, such that this is a minor detail. As with the RBD, the application of extensive MD simulations prior to Monte Carlo simulations negates any differences in starting structure.

For RBD-link-dimerization domain simulations (316 residue systems), we opted to use a single starting seed structure for the folded domains based on the NMR and crystal-structure conformations for the RBD and dimerization domains, respectively. During these simulations, a subset of the trajectories became stuck due to long-lived interactions between the RBD and the dimerization domain, an effect likely that rose from exposed hydrophobic residues in the dimerization domain being exposed as 'folded' residues. To mitigate the impact of these unphysiological sub-ensembles, we identified trajectories in which we found contiguous simulation frames in which 25% or more of the total simulation ensemble showed unvarying interdye distance. This diagnostic identified 3 of the 31 independent replicas as being problematic, and these were discarded from our analysis. The remaining ensemble consists of 29 independent trajectories.

Excluded volume (EV) simulations were performed using the same setup, but with a modified Hamiltonian under which solvation, attractive Lennard-Jones, and polar (charge) interactions are scaled to zero, as described previously [306].

Molecular Dynamics Simulations

All molecular dynamics simulations of SARS-CoV-2 nucleoprotein were performed with Gromacs 2019 using the AMBER03 force field with explicit TIP3P solvent [718–720]. Simulations were prepared by placing the starting structure in a dodecahedron box that extends 1.0 Å beyond the protein in any dimension. The system was then solvated, and energy minimized with a steepest descents algorithm

until the maximum force fell below 100 kJ/mol/nm using a step size of 0.01 nm and a cutoff distance of 1.2 nm for the neighbor list, Coulomb interactions, and van der Waals interactions. For production runs, all bonds were constrained with the LINCS algorithm and virtual sites were used to allow a 4 fs time step [721,722]. Cutoffs of 1.1 nm were used for the neighbor list with 0.9 for Coulomb and van der Waals interactions. The Verlet cutoff scheme was used for the neighbor list. The stochastic velocity rescaling (v-rescale) thermostat was used to hold the temperature at 300 K [723]. Conformations were stored every 20 ps.

The FAST algorithm was used to enhance conformational sampling and quickly explore the dominant motions of nucleoprotein [724,725]. FAST-pocket simulations were run for 6 rounds, with 10 simulations per round, where each simulation was 40 ns in length (2.4 μs aggregate simulation). The FAST-pocket ranking function favored restarting simulations from states with large pocket openings. Additionally, a similarity penalty was added to the ranking to promote conformational diversity in starting structures, as has been described previously [726]. The FAST dataset was clustered using a k-centers algorithm based on RMSD between frames using backbone heavy atoms (C, Cα, Cβ, N, O) to generate 1421 discrete states, which were then launched on the distributed computing platform Folding@home [284].

To generate large-scale ensembles of the folded domains, extensive simulations on the Folding@home platform were used. For the RBD, folding@home produced 500 μs of aggregate simulation data. For a monomeric version of dimerization domain, Folding@home produced 2.12 ms of aggregate simulation data. For each of these datasets, a final k-centers clustering was performed with the combined Folding@home and FAST data using Enspara (https://github.com/bowman-lab/enspara) [727]. This clustering was performed the same as described

above and generated 200 discrete states that capture maximal diversity in the conformational ensemble of the two folded domains. These states were then used as the starting seeds for the folded domain conformations in CAMPARI simulations.

Sequential Molecular Dynamics + Monte Carlo Sampling Approach

The NTD and RBD combined are 173 residues of folded and disordered protein, while the dimerization domain and CTD combined are almost exactly the same size at 172 residues. Systems of this size raises a significant challenge for all-atom sampling. To address this we leveraged a novel approach in which we first ran long all-atom molecular dynamics simulations of folded domains alone using the Folding@Home platform and the FAST approach for enhanced conformational sampling [284,725]. From each of the trajectories of the RBD or dimerization domain, we then identified 200 conformationally distinct states based on these simulations which we used as "seeds" for the starting structures of the folded domains in our Monte Carlo simulations. Using these seeds, we reconstructed the previously missing disordered regions (NTD and CTD, respectively) and ran all-atom Monte Carlo simulations in which the disordered regions are fully sampled, the folded domain sidechains are fully sampled, but the folded domains backbone dihedral angles are held fixed. For the NTD-RBD construct we ran two replicas of each starting conformation were run, with 400 independent simulations generating a total ensemble of ~400 K conformations. For the dimerization domain we did not run independent replicas from the same starting configuration, such that 200 independent simulations were run that generated an ensemble of 200 K conformations. In parallel, we also ran simulations of the NTD and CTD in isolation, enabling an assessment of the impact of the folded domain.

Coarse-Grained Polymer Simulations

Coarse-grained simulations were performed using the PIMMS software package [80,407]. PIMMS is a Monte Carlo lattice-based simulation engine in which each bead engages in anisotropic interactions with every adjacent lattice site. Moves used here were cluster translation/rotation moves and single-bead perturbation moves. Specifically, every simulation step, each bead in the system is sampled to move to adjacent sites in random order 50^3 of times multiplied by a factor that reflects the length of the chain. Every 100 moves (on average) a cluster of chains is randomly selected and translated or rotated, where a cluster reflects a collection of two or more chains in direct contact. This moveset provides changes to the system that reflect physical movements expected in a dynamical system, allowing us to - for equivalently sized systems - compare the apparent dynamics of assembly, as has been done previously [728–731]. We repeated the simulations presented using a range of different movesets and, while convergence varied from set-to-set, we always observed analogous results.

All simulations were performed in a 70 x 70 x 70 lattice-site box using period boundary conditions. The results reported are averaged over the final 20% of the simulation to give average values after equivalent numbers of MC steps. The "polymer" is represented as a 61-residue polymer with either a central high-affinity binding site or not. The binder is a 2-bead species. Every simulation was run for 20 x 10^9 Monte Carlo steps, with four independent replicas. Simulations were run with 1,2,3,4 or 5 polymers and 50, 75, 100, 125, 150, 175, 200, 250, 300, 400 binders.

To further explore the physical basis for single-chain polymer condensates we ran additional extended simulations for 60 x 10^9 Monte Carlo with a moveset that includes the ability for clusters to move. Simulations were run using the same conditions for other simulations, with ten independent simulations for condition (**Fig. S18**).

Extended Discussion on Coarse-Grained Simulations

For simulations of homopolymeric polymers as shown in **Fig. 6C,D** the balance of chain-compaction and phase separation is determined in part through chain length and binder K_d. In our system the polymer is largely unbound in the one-phase regime (suggesting the concentration of ligand in the one-phase space is below the K_d) but entirely coated in the two-phase regime, consistent with highly-cooperative binding behavior. In the limit of long, multivalent polymers with multivalent binders, the sharpness of the coil-to-globule transition is such that an effective two-state description of the chain emerges, in which the chain is *either* expanded (non-phase separation-competent) OR compact (coated with binders, phase separation competent).

An alternative framework for understanding our simulations of single-polymer condensates comes from the idea of two distinct concentration (phase) boundaries - one for binder:high affinity site interaction (c_1), and a second boundary for "nonspecific" binder:polymer interactions (c_2) at a higher concentration. c_2 reflects the boundary observed in **Fig. 6C** that delineated the one and two-phase regimes. At global concentrations below c_2, (but above c_1) the clustering of binders at a high affinity site raises the apparent local concentration of binders above c_2, from the perspective of other beads on the chain. In this way, a local high affinity binding site can drive "local" phase separation of a single polymer.

Protein expression, purification, and labeling

Plasmid Construct Design.

SARS-CoV2 Nucleocapsid protein (NCBI Reference Sequence: YP_009724397.2) including an N term extension containing **His_9**-*HRV 3C protease site* –

CATCATCACCATCATCATCATCACCAC*CTCGAAGTTCTGTTCCAAGGCCCG*ATGAGTG
ATAACGGTCCCCAGAATCAACGGAATGCGCCCAGAATCACGTTCGGCGGTCCAAGCG
ACAGTACAGGTTCGAATCAGAATGGTGAACGCTCTGGGGCCCGAAGCAAACAGCGT

CGTCCACAGGGTTTGCCGAACAATACGGCTAGCTGGTTCACTGCGCTGACGCAGCAC
GGAAAAGAAGACTTAAAATTTCCGCGAGGCCAGGGGGTCCCGATTAATACTAACTCC
TCCCCTGACGATCAAATTGGTTATTATCGTCGTGCAACCCGCCGTATCCGCGGCGGA
GACGGTAAAATGAAAGATCTGTCACCGCGCTGGTATTTTTACTACCTGGGAACAGGT
CCTGAAGCAGGCTTGCCGTATGGCGCTAACAAAGATGGCATTATCTGGGTGGCTACC
GAGGGTGCCCTTAATACGCCGAAAGATCATATTGGAACCCGTAACCCAGCCAATAAC
GCAGCAATCGTACTGCAGCTGCCGCAGGGGACAACCCTGCCGAAAGGCTTTTATGCG
GAAGGGAGTCGTGGCGGCAGCCAAGCCAGCTCCCGTAGCTCCTCGCGCTCTCGCAAC
TCCTCGCGGAATAGTACACCGGGTTCATCACGCGGCACCTCGCCGGCACGCATGGCT
GGCAACGGGGGGGATGCGGCTTTGGCGTTACTTTTACTGGATAGGCTTAACCAGTT
GGAAAGTAAAATGAGCGGTAAAGGCCAGCAGCAGCAGGGTCAGACTGTGACCAAAA
AGAGCGCGGCAGAGGCGTCGAAAAAACCTAGACAAAAGCGTACTGCGACCAAAGCC
TACAATGTTACGCAGGCATTCGGCCGGCGCGGTCCGGAACAAACCCAGGGCAACTTT
GGTGACCAGGAGCTGATTCGTCAGGGAACCGATTACAAACACTGGCCACAGATCGC
GCAATTTGCCCCCTCGGCGTCAGCCTTTTTTGGTATGTCTCGCATTGGGATGGAGGT
AACCCCGTCTGGCACGTGGCTGACGTACACGGGCGCTATAAAGCTGGATGATAAAGA
TCCGAACTTCAAAGACCAGGTGATCTTACTGAACAAACATATTGACGCCTATAAAACG
TTCCCCCCTACTGAACCTAAGAAAGATAAAAAAAAAAAGGCCGATGAAACCCAAGCG
CTACCACAACGCCAGAAAAAGCAGCAGACCGTCACCCTCCTGCCGGCAGCGGACCTC
GACGATTTTTCTAAGCAACTGCAACAAAGCATGTCAAGCGCCGATAGTACACAGGCG
TAA

- was cloned into the BamHI EcoRI sites in the MCS of pGEX-6P-1 vector (GE Healthcare) to express the protein product:

GST-
LEVLFQGPLGS**HHHHHHHHHH***LEVLFQGP*MSDNGPQNQRNAPRITFGGPSDSTGSNQ
NGERSGARSKQRRPQGLPNNTASWFTALTQHGKEDLKFPRGQGVPINTNSSPDDQIGYY
RRATRRIRGGDGKMKDLSPRWYFYYLGTGPEAGLPYGANKDGIIWVATEGALNTPKDH
IGTRNPANNAAIVLQLPQGTTLPKGFYAEGSRGGSQASSRSSSRSRNSSRNSTPGSSRGTSP
ARMAGNGGDAALALLLLDRLNQLESKMSGKGQQQQGQTVTKKSAAEASKKPRQKRTA
TKAYNVTQAFGRRGPEQTQGNFGDQELIRQGTDYKHWPQIAQFAPSASAFFGMSRIGM
EVTPSGTWLTYTGAIKLDDKDPNFKDQVILLNKHIDAYKTFPPTEPKKDKKKKADETQ
ALPQRQKKQQTVTLLPAADLDDFSKQLQQSMSSADSTQA

Site-directed mutagenesis was performed on the His_9-SARS-CoV2 Nucleocapsid pGEX vector to create the N protein constructs (**Table S1**). All cloning and site-directed mutagenesis steps were performed by Genewiz and sequences were verified using sanger sequencing.

Protein Expression and Purification

Both GST-His_9-SARS-CoV2 NTD FL and LINK FL Nucleocapsid variants were expressed recombinantly in BL21 Codon-plus pRIL cells (Agilent). 4L cultures were grown in LB medium containing carbenicillin (100 ug/mL) to OD_{600} ~ 0.6 and induced with 0.2 mM IPTG for 12 hours at 16ºC. Harvested cells were lysed with sonication at 4ºC in lysis buffer (50 mM Tris pH 8, 500 mM NaCl, 10% glycerol, 10 mg/mL lysozyme, 5 mM BME, cOmplete™ EDTA-free Protease Inhibitor Cocktail (Roche), DNAse I (NEB), RNAse H (NEB)). The supernatant was cleared by centrifugation (140,000 x g for 1 hr) and bound to an HisTrap FF column (GE Healthcare) in buffer A (50 mM Tris pH 8, 500 mM NaCl, 10% glycerol, 20 mM imidazole, 5 mM BME). GST-His_9-N protein fusion was eluted with buffer B (buffer A + 500 mM imidazole) and dialyzed into cleavage buffer (50 mM Tris pH 8, 50 mM NaCl, 10% glycerol, 1 mM DTT) with HRV 3C protease, thus cleaving the GST-His_9-N fusion yielding FL N protein with two additional N-term residues (GlyPro). FL N protein was then bound to an SP sepharose FF column (GE Healthcare) and eluted using a gradient of 0-100% buffer B (buffer A: 50 mM Tris pH 8, 50 mM NaCl, 10% glycerol, 5 mM BME, buffer B: buffer A + 1 M NaCl) over 100 min. Purified N protein variants were analyzed using SDS-PAGE and verified by electrospray ionization mass spectrometry (LC-MS). Concentrations were determined spectroscopically in 50 mM Tris (pH 8.0), 500 mM NaCl, 10% (v/v) glycerol using an extinction coefficient = 42530 M^{-1} cm^{-1}

GST-His_9-SARS-CoV2 wild-type, RBD-FL, LINK-ΔDimer, NTD-RBD, and CTD-FL Nucleocapsid variants were expressed recombinantly in Gold BL21(DE3) cells (Agilent). 4 L cultures were grown in LB medium with carbenicillin (100 ug/mL) to OD_{600} ~ 0.6 and induced with 0.2 mM IPTG for 3 hours at 37ºC. Harvested cells were lysed with sonication at 4ºC in lysis buffer (listed above). The supernatant was cleared by centrifugation (140,000 x g for 1 hr) and the pellet was resuspended in 50 mM Tris pH 8, 500 mM NaCl, 10% glycerol, 6 M Urea, 5 mM BME and incubated at 4ºC for one hour. The resuspension was cleared by centrifugation (140,000 x g for 1 hr) and the GST-His_9-N protein in the supernatant was bound to a FF HisTrap column (GE Healthcare) in buffer A (50 mM Tris pH 8, 500 mM NaCl, 10% glycerol, 20 mM imidazole, 5 mM BME) containing 6 M Urea. The column was then washed with buffer A allowing the protein to refold on the column. The GST-His_9-N protein fusion was then eluted with buffer B (buffer A containing 500 mM imidazole) and dialyzed into cleavage buffer (50 mM Tris pH8, 50 mM NaCl, 10% glycerol, 1 mM DTT) containing HRV 3C protease. FL N protein was then bound to an SP sepharose FF column (GE Healthcare) and eluted using a gradient of 0-100% buffer B (buffer A: 50 mM Tris pH 8, 50 mM NaCl, 10% glycerol, 5 mM BME, buffer B: buffer A + 1 M NaCl) over 100 min. Purified N protein variants were analyzed using SDS-PAGE and/or verified by electrospray ionization mass spectrometry (LC-MS). Protein concentrations of stock solutions were determined spectroscopically in 50 mM Tris (pH 8.0), 200-500 mMNaCl, 10% (v/v) glycerol using extinction coefficients of 42530 M^{-1} cm^{-1}(FL) , 26400M^{-1} cm^{-1} (LINK-ΔDimer), and 25200M^{-1} cm^{-1} (NTD-RBD).

GST-His9-SARS-CoV2 CTD Nucleocapsid was expressed recombinantly in Gold BL21(DE3) cells (Agilent). 4 L cultures were grown in LB medium with carbenicillin (100 ug/mL) to OD600 ~ 0.6

and induced with 0.2 mM IPTG for 3 hours at 37ºC. Harvested cells were lysed with sonication at 4ºC in lysis buffer (50 mM MES pH 6, 500 mM NaCl, 10% glycerol, 5 mM BME, 10mg/mL lysozyme). The supernatant was cleared by centrifugation (140,000 x g for 1 hr) and the GST-His9-N protein in the supernatant was bound to a FF HisTrap column (GE Healthcare) in buffer A (50 mM MES pH 6, 500 mM NaCl, 20 mM imidazole, 10% glycerol, 5 mM BME). The GST-His9-N protein fusion was then eluted with buffer B (buffer A containing 500 mM imidazole) and dialyzed into cleavage buffer (A. 50 mM MES pH 6, 50 mM NaCl, 10% glycerol, 1 mM DTT) containing HRV 3C protease. FL N protein was then bound to an SP sepharose FF column (GE Healthcare) and eluted using a gradient of 0-100% buffer B (buffer A: 50 mM MES pH 6, 50 mM NaCl, 10% glycerol, 5 mM BME, buffer B: buffer A + 1 M NaCl) over 100 min. Purified N protein was analyzed using SDS-PAGE. Protein concentrations of stock solutions were determined spectroscopically in 50 mM MES (pH 6.0), 300 mM NaCl, 10% (v/v) glycerol using an extinction coefficient = 120M^{-1} cm^{-1}

Choice of labeling positions

The choice of the labeling positions has been obtained as a compromise between flanking the regions of interest and a series of different criteria that regards the biophysics of disordered proteins, the structural properties of the protein, and the physicochemical properties of the fluorophores. In particular, we have attempted to obtain an optimal spacing of the fluorophores to ensure we could make use of the whole FRET dynamic range. A separation between 60 to 70 amino acids is expected to provide a transfer efficiency of about 0.5 for a disordered region with scaling exponent close to 0.5 and 0.8-0.9 for a folded or collapsed state with a scaling exponent of 0.33 [35]. We have attempted to avoid altering amino acids that are clearly involved in structurally relevant interactions based on

inspection of known structures of the folded domains. When looking for labeling positions in a folded domain, we have aimed for surface exposed residues to maximize the accessibility of the cysteine residues during labeling. We have avoided placing fluorophores adjacent to charged residues to avoid possible interactions with the charges of the fluorophores. Finally, we have attempted to limit the effects of quenching between fluorophores and aromatic residues [195,657]. Regarding this point, tryptophan residues have been identified as major quenchers of Alexa 488 and 594 and a spacing of twenty or more residues would be optimal [195,657]. Following these criteria, we have preferred not to label the NTD construct in position 50 due to the close proximity with a tryptophan residue and opted for a residue within the structured RBD. Similarly, we have opted to insert the labels within the LINKER such that mutations were not altering the net charge of the LINK sequence. Finally, for the CTD we have opted for spacing the labeling position far apart from the tryptophan residue within the folded dimerization domain, though this may not have been sufficient based on the ns-FCS observations of the CTD-FL.

Fluorescent Dye Labeling

All Nucleocapsid variants were labeled with Alexa Fluor 488 maleimide (Molecular Probes) under denaturing conditions in buffer A (50 mM Tris pH8, 50 mM NaCl, 10% glycerol, 6M Urea, 1 mM DTT) at a dye/protein molar ratio of 0.7/1 for 2 hrs at room temperature. Single labeled protein was isolated via ion-exchange chromatography (Mono S 5/50 GL, GE Healthcare - protein bound in buffer A and eluted with 0-100% buffer B (buffer A + 1 M NaCl) gradient over 100 min) and UV-Vis spectroscopic analysis to identify fractions with 1:1 dye:protein labeling. Single labeled Alexa Fluor 488 maleimide labeled N protein was then subsequently labeled with Alexa Fluor 594 maleimide at a dye/protein molar ratio of 1.3/1 for 2 hrs at room temperature. Double labeled

(488:594) protein was then further purified via ion-exchange chromatography (Mono S 5/50 GL, GE Healthcare - see above).

Single Molecule Spectroscopy

Experimental Setup and Procedure

Single-molecule fluorescence measurements were performed with a Picoquant MT200 instrument (Picoquant, Germany). For single-molecule FRET measurements, a diode laser (LDH-D-C-485, PicoQuant, Germany) was synchronized with a supercontinuum laser (SuperK Extreme, NKT Photonics, Denmark), filtered by a z582/15 band pass filter (Chroma) and pulsed at 20 MHz for pulsed interleaved excitation (PIE) [215] of labeled molecules. Emitted photons were collected with a 60x1.2 UPlanSApo Superapochromat water immersion objective (Olympus, Japan), passed through a dichroic mirror (ZT568rpc, Chroma, USA), and filtered by a 100 μm pinhole (Thorlabs, USA). Photons are counted and accumulated by a HydraHarp 400 TCSPC module (Picoquant, Germany). For FRET-FCS measurements, the same diode laser was used in continuous-wave mode to excite the donor dye. Photons emitted from the sample were collected by the objective, and scattered light was suppressed by a filter (HQ500LP, Chroma Technology) before the emitted photons passed the confocal pinhole (100 mm diameter). The emitted photons were then distributed into four channels, first by a polarizing beam splitter and then by a dichroic mirror (585DCXR, Chroma) for each polarization. Donor and acceptor emission was filtered (ET525/50m or HQ642/80m, respectively, Chroma Technology) and then focused on SPAD detectors (Excelitas, USA). The arrival time of every detected photon was recorded with a HydraHarp 400 TCSPC module (PicoQuant, Germany).

FRET experiments were performed by exciting the donor dye with a laser power of 100 μW (measured at the back aperture of the objective). For pulsed interleaved excitation experiments, the power used for exciting the acceptor dye was adjusted to match a total emission intensity after acceptor excitation to the one observed upon donor excitation (between 50 and 70 mW). Single-molecule FRET efficiency histograms were acquired from samples with protein concentrations between 50 pM and 100 pM. Trigger times for excitation pulses (repetition rate 20 MHz) and photon detection events were stored with 16 ps resolution.

For fluorescence correlation spectroscopy (FCS) experiments, acceptor-donor labeled samples with a concentration of 100 pM were excited by either the 485 nm diode laser or the supercontinuum laser at the powers indicated above. However, in the experiments on protein oligomerization, due to an increase in the fluorescence background upon addition of unlabeled protein above 1 μM, only the correlations corresponding to direct acceptor excitation (582 nm) have been considered reliable for the analysis.

For nsFCS, FRET samples of acceptor-donor labeled protein with a concentration of approximately 100 pM were excited by the same diode laser but in continuum wavelength mode.

All measurements were performed in 50 mM Tris pH 7.32, 143 mM β-mercaptoethanol (for photoprotection), 0.001% Tween 20 (for surface passivation) and GdmCl at the reported concentrations. A residual concentration of 0.05-0.06 M GdmCl is present from dilution of the protein from the stock denatured sample. All measurements were performed in uncoated polymer coverslip cuvettes (Ibidi, Wisconsin, USA) and custom-made glass cuvette coated with PEG (see

PEGylation section below). Both materials outperform normal glass cuvette and contribute to reduced sticking of the protein to the surface. At low salt we observed improved protection from sticking when using the PEG coated cuvette.

Each sample was measured for at least 30 min at room temperature (295 ± 0.5 K).

PEGylation of Glass Surfaces

Glass cuvettes were assembled using 8 mm glass cloning cylinders (Hilgenberg) and 25mm circular coverslips (Deckglaser) glued together with optical adhesive 61 (Norland). Then, glass cuvettes were washed with 2% Contrad, rinsed with double distilled water, dried, and immediately filled with 100% methanol (Sigma-Aldrich). Methanol was replaced with an amino-modifying solution (methanol, acetic acid (Sigma-Aldrich), amino silane (UCT Specialties LLC)) and the solution was incubated for 10 min, followed by a one-minute sonication. After sonication, the solution was incubated for further 10 minutes and then rinsed with 100% methanol followed by a second wash with double distilled water and dried. Immediately after, the cuvettes were filled with a solution containing PEG (0.1M sodium bicarbonate(Santa Cruz Biotechnology, Inc.), mPEG-SVA (Laysan Bio)). Cuvettes were placed in a glass petri dish, covered, and stored in a dark humid environment at 4C overnight. The following morning the cuvettes were rinsed well with double distilled water, dried, vacuum sealed, and stored at -20C.

FRET Efficiency Histograms

Fluorescence bursts from individual molecules were identified by time-binning photons in bins of 1 ms and retaining the burst if the total number of photons detected after donor excitation was larger

than at least 20. The exact threshold was selected based on the background contribution identified in the photon counting histograms with 1 ms binning. Transfer efficiencies for each burst were calculated according to $E=nA/(nA+nD)$, where nD and nA are the numbers of donor and acceptor photons, respectively. Corrections for background, acceptor direct excitation, channel crosstalk, differences in detector efficiencies, and quantum yields of the dyes were applied [732]. The labeling stoichiometry ratio S was computed accordingly to $S = I_D/(\gamma_{PIE} I_A + I_D)$where $I_D$and $I_A$represents the total intensities observed after donor and acceptor excitation and γ_{PIE} provides a correction factor to account for differences in the detection efficiency and laser intensities. Bursts with stoichiometry corresponding to 1:1 donor:acceptor labeling (in contrast to donor and acceptor only populations) were selected and finally from the selected bursts a histogram of transfer efficiencies is constructed. Variations in the selection criteria for the stoichiometry ratio do not impact significantly the observed mean transfer efficiency (within experimental errors).

To estimate the mean transfer efficiency and deconvolve multiple populations (e.g for the NTD construct) from the transfer efficiency histograms, each population was approximated with a Gaussian peak function. For fitting more than one peak, the histogram was analyzed with a sum of Gaussian peak functions. Under these assumptions the mean transfer efficiency is computed as an average quantity across hundreds of independent molecules freely diffusing in the confocal volume. For the conversion of transfer efficiency to distances, we used the value of the Förster radius for Alexa488 and Alexa594 previously determined and reported in literature, $R_0 = 5.4$ nm [162]. We further correct the value accounting for the dependence of the Förster radius on the solution refractive index. To this end, we quantified the change in refractive index for each sample, which enables us to strongly reduce the source of error due to possible pipetting mistakes and properly determined

concentrations of denaturant and salt. The changes in refractive index caused by increasing concentrations of GdmCl or KCl were measured with an Abbe refractometer (Bausch & Lomb, USA).

We estimated a systematic error on transfer efficiency of ± 0.03, based on the variation of transfer efficiency of the same reference samples after different calibrations of the instrument over the last two years, a number in line with previously reported systematic errors for analogous instrumentation and calibration [657,733]. Standard deviation of the transfer efficiency for multiple replicates of the same experimental conditions commonly results in a standard deviation equal or less than ± 0.01. Since we aim for a comparison with simulations, here we consider the systematic error as the largest source of error and we propagate the corresponding effect on all the calculated distances.

Each point in the denaturant titration is obtained from independently prepared samples. Reproducibility of the mean transfer efficiency results have been confirmed by independent replicates of measurements in aqueous buffer and at various concentrations of the denaturation curve. For the NTD FL construct, we performed two independent sample preparation and measurements for 0, 0.3, 0.6, 0.8, 2.3, 4.5, and 6 M GdmCl as well as 0.5, 1, 1.5, 2 M Urea. The corresponding standard deviation for each of the measurements is equal to or smaller than 0.01. For the NTD-RBD, we have performed duplicates at 0 and 6 M GdmCl (with standard deviation equal or less than 0.01) and we have found a remarkable agreement of the measured transfer efficiencies across all denaturant concentrations with the NTD-FL. For the LINK-FL, reproducibility has been confirmed by 2 independent replicates at 0, 1, 2, 4 M GdmCl as well as 50 and 150 mM KCl. Standard deviation of independent replicates is less than 0.01. Measurements of coexistent

populations below 0.15 M GdmCl provides a further indication of the small deviations across independent measurements reporting about the same distance distribution. Reproducibility of experimental is further corroborated by overlapping of data points with the independent preparation measuring the LINK-ΔDimer construct in high denaturant where both constructs converge to equal transfer efficiencies. Regarding the LINK-ΔDimer construct, besides the overlap of transfer efficiencies in high denaturant, we additionally performed duplicates at 0, 0.5, 0.75 M GdmCl and at 1, 2, 3, 4 M Urea. For the CTD FL, we tested reproducibility by performing duplicates at 0.25, 0.5, 0.75, 1, 1.25, and 6 M GdmCl, as well as at 300 and 500 mM KCl. While all these measurements results in a standard deviation equal or smaller than 0.01, repeated measurements in aqueous buffer (4 measurements) and in 1 and 2 M Urea (2 measurements each) revealed larger standard deviations comparable or smaller than 0.03. We attribute these observations to the specificity of the CTD (and possibly DIMER domain) and its larger propensity to interact with the surface. This effect is not observed at higher GdmCl or salt concentrations that 0.15 M, but seems to persist in Urea, suggesting a possible contribution of electrostatic interactions. Finally, we confirmed reproducibility of the results for the CTD fragment by performing independent duplicates of 1, 1.5, 2, 2.75 M GdmCl as well as 4 independent measurements of the sample in aqueous buffer. Each set of measurements report a standard deviation less than 0.01, suggesting that the peculiarity of the CTD FL sample is connected to the presence of the DIMER domain. Reproducibility is further corroborated by the overlapping of data points with the measurement of the CTD FL. Overall, testing reproducibility of the samples across multiple experimental conditions revealed deviations not exceeding the systematic error that is intrinsic to the instrument calibrations.

Fluorescence Lifetimes and Anisotropies Analysis

A quantitative interpretation of this transfer efficiency in terms of distance distribution requires the investigation of protein dynamics. A first method to assess whether the transfer efficiency reports about a rigid distance (e.g. structure formation or persistent interaction with the RBD) or is the result of a dynamic average across multiple conformations is the comparison of transfer efficiency and fluorescence lifetime.The interdependence of these two factors is expected to be linear if the protein conformations are identical on both timescales (nanoseconds as detected by the fluorescence lifetime, milliseconds as computed from the number of photons in each burst). Alternatively, protein dynamics give rise to a departure from the linear relation and an analytical limit can be computed for configurations rearranging much faster than the burst duration (see SI). The dependence of the fluorescence lifetimes on transfer efficiencies determined for each burst was compared with the behavior expected for fixed distances and for a chain sampling a broad distribution of distances. For a fixed distance, *R*, the mean donor lifetime in the presence of acceptor is given by $t_D(R) = t_{D0}\,(1 - E(R))$, where t_D is the lifetime in the absence of acceptor, and $E(R) = 1/(1 + R^6/\ R_0^6)$. For a chain with a dye-to-dye distance distribution *P*(*R*), the donor lifetime is $t_D = \int tI(t)dt\ /\ \int I(t)dt$, where $I(t) = I_0\,P(R)\,Exp[-t/tD(R)]\,dR$ is the time-resolved fluorescence emission intensity following donor excitation. A similar calculation can be carried out for describing the acceptor lifetime delay given by $(t_A(R) - t_{A0})/t_{D0}$ [169]. Donor and acceptor lifetimes at different concentrations of GdmCl were analyzed by fitting subpopulation-specific time-correlated photon counting histograms after donor and acceptor excitation, respectively, using a tail fit. Errors associated with the tail fit are estimated by varying the "tail" region that undergoes the fitting procedure and computing mean and standard deviation of the fit results. In computing the average of multiple measurements, errors of the single dataset are propagated accordingly.

Multiparameter detection allows also excluding possible artifacts, such as insufficient rotational averaging of the fluorophores or quenching of the dyes. Subpopulation-specific anisotropies were determined for both donor and acceptor of all three constructs for NTD, LINK, and CTD, and values were found to vary between 0.1 and 0.2 for the donor and between 0.1 and 0.2 for the acceptor, sufficiently low to assume as a good approximation for the orientational factor $\kappa^2 = 2/3$.

Fluorescence Correlation Spectroscopy (FCS) Analysis

In order to determine changes in the hydrodynamic radius (R_h) of the protein, FCS correlations were analyzed assuming 3D diffusion of the molecule across a three-dimensional Gaussian profile of the confocal volume [734]. For 1 diffusing species, and in the absence of photophysical transitions in the time scale of the lag times analyzed, this formalism amounts to the following time autocorrelation function $g(\tau) = 1 + \frac{1}{N}(1+\frac{\tau}{\tau_D})^{-1}(1+\frac{\tau}{\alpha^2\tau_D})^{-1/2}$, where N is the average number of molecules in the confocal volume, τ_D is the diffusion time along the xy plane, a is the eccentricity of the three dimensional Gaussian observational volume. $\tau_D = \omega_{xy}^2 / 4D$, where D is the 3D translational diffusion coefficient and ω_{xy} is the radius from the center of the laser beam at which the light intensity decreases e^2 times from its maximum value at the center $a = \omega_z/\omega_{xy}$.

Additionally, in order to account for contributions of the photophysics of the fluorophore to the correlation observed in the μs timescale, we added two triplet terms multiplying the diffusion correlation term (see for example work by Krichevsky [735]). The overall equation that we fit to the FCS traces is then $g(\tau) = 1 + (g_{diff}(\tau) - 1)(1 + c_{T1}Exp[-\frac{\tau}{\tau_{T1}}])(1 + c_{T2}Exp[-\frac{\tau}{\tau_{T2}}])$ where τ_{T1}, τ_{T2}, c_{T1}, and c_{T1}, denotes the characteristic times and amplitudes of

the contributions of two triplet states to $g(\tau)$. Parameters τ_{diff}, τ_{T1}, τ_{T2}, c_{T1}, c_{T2} and N were fitted by least square nonlinear regression analysis for each concentration of unlabeled protein tested (**Fig. S14 A-B**), while a was fixed at a value of 6 determined independently from analysis of fluorescence intensity profiles of fluorescent nanobeads.

Making use of the definition of τ_{diff} and the Stokes-Einstein equation, we have, for each concentration of unlabeled protein (τ_{diff} / τ_{diff0}) = (R_h / R_{h0}), where τ_{diff0} and R_{h0} are the diffusion time and hydrodynamic radius in the absence of unlabeled protein, respectively. Error bars in **Fig. S14 B** are the standard errors of R_h / R_{h0} estimated from propagation of the standard errors across multiple measurements of the diffusion times obtained from the fit.

Nanosecond Fluorescence Correlation Spectroscopy

Autocorrelation curves of acceptor and donor channels and cross-correlation curves between acceptor and donor channels were calculated with the methods described previously [170,736]. All samples have been measured at a concentration of 100 pM and bursts with a transfer efficiency between 0.3 and 0.8 have been selected to eliminate the contribution of donor only to the correlation amplitude. Finally, the correlation was computed over a time window of 5 µs and characteristics timescales were extracted according to:

$$g_{ij}(\tau\ ) = 1 + \frac{1}{N}(1 - c_{AB}Exp[-(\tau - \tau_0)/\tau_{AB}])(1 + c_{CD}Exp[-(\tau - \tau_0)/\tau_{CD}])(1 + c_T Exp[-(\tau - \tau_0)/\tau_T]) \quad \text{(Eq. S1)}$$

where N is the mean number of molecules in the confocal volume and i and j indicate the type of signal (either from the Acceptor or Donor channels). The three multiplicative terms describe the contribution to amplitude and timescale of photon antibunching (AB), chain dynamics (CD), and triplet blinking of the dyes (T). $\tau_{CD}$is then converted in the reconfiguration time of the interdye distance $\tau_r$correcting for the filtering effect of FRET as described previously [197]. An additional multiplicative CD term has been added only for the donor-donor correlations to describe the fast decay observed at very short time. Such a decay is not found in the correlations of other disordered proteins measured on the instrument and we associate the fast decay with the rotational motion of the overall protein. A fit to this fast decay is about 2 ns. To test reproducibility, we perform multiple independent measurements: 3 for the NTD-FL, 4 for the LINK-FL, and 6 for the CTD-FL.

Polymer Models of Distance Distributions

Conversion of mean transfer efficiencies for fast rearranging ensembles requires the assumption of a distribution of distances. Here, we compared the results of two distinct polymer models: the Gaussian model and a Self-Avoiding Walk (SAW) model that accounts for changes in the excluded volume [184]. This second model has been shown to provide a better description of chain distribution and scaling exponent when compared to distance distributions from MD simulations [173]. Importantly, both models rely only on one single fitting parameter, the root mean square interdye distance $r = \langle R^2 \rangle^{1/2}$ for the Gaussian chain and the scaling exponent $\boldsymbol{\nu}$ for the SAW model.

Estimates of these parameters are obtained by numerically solving:

$$\langle E \rangle = \int_0^{l_c} P(R)\, E(R)\, dr \qquad \text{(Eq. S2)}$$

where *R* is the interdye distance, l_c is the contour length of the chain, $P(r)$represents the chosen distribution, and $E(R)$is the Förster equation for the dependence of transfer efficiency on distance R and Förster radius:

$$E(R) = \frac{R_0^6}{R_0^6 + R^6}. \quad \text{(Eq. S3)}$$

The Gaussian chain distribution is given by:

$$P_{FJC}(R, r) = 4\pi R^2 \left(\frac{3}{2\pi r^2}\right)^{3/2} exp\left(-\frac{3R^2}{2r^2}\right) \quad \text{(Eq. S4)}$$

The SAW model can be expressed as:

$$P_{SAW}(R, v) = A_1 \frac{4\pi}{b_0 N^v} \left(\frac{R}{b_0 N^v}\right)^{2+g} Exp\left[-A_2 \left(\frac{R}{b_0 N^v}\right)^{\delta}\right] \quad \text{(Eq. S5)}$$

where $A_1 = \frac{\delta}{4\pi} \frac{\Gamma[5+g/\delta]^{\frac{3+g}{2}}}{\Gamma[3+g/\delta]^{\frac{5+g}{2}}}$, $A_2 = \left(\frac{\Gamma[5+g/\delta]}{\Gamma[3+g/\delta]}\right)^{\frac{\delta}{2}}$, $g = \frac{(\gamma-1)}{v}$,$\delta = \frac{1}{(1-v)}$, $\gamma =$ 1.1615,$\Gamma$is the Euler Gamma Function, b_0 = 0.55 nm is an empirical prefactor [173] , *N* is the number of residues between the fluorophores, and $\boldsymbol{\nu}$ is the scaling exponent.

Finally, when converting the distance from transfer efficiencies, to account for the length of dye linkers and compare the experimental data with simulations, the root-mean-squared interdye distance *r* was rescaled according to $r_{m,n} = |m-n|^{0.5}/_{\text{dye}} \; |m-n+2/_{\text{dye}}|^{0.5}$ with $/_{\text{dye}}$ = 4.5 [233,733]. Finally, the persistence length is computed using the Gaussian conversion $r^2 = 2\, l_p l_c$ [398].

Binding of Denaturant and Folding.

As in previous works [168,185,231], we model the chain expansion with the denaturant in terms of a simple binding model:

$$r(c) = r_0\left(1 + \rho \frac{Kc}{1+Kc}\right) \quad \text{(Eq. S6)}$$

Where $r_0$is the mean square interdye distance at zero denaturant,$\rho$is a term the captures the extent of chain expansion with the denaturant compared to r_0, and the K is the binding constant, and c is the concentration of denaturant.

In presence of folded domains, we can imagine the folding/unfolding of the domains can affect the overall size of the chain because of an increase or decrease of excluded volume due to the surrounding folded domains (which screen part of the available conformations) or because of the folding or unfolding of elements in the region between the fluorophores. To account for this effect, as in the case of the NTD, we weighed the effect of denaturant on the chain for the fraction folded $f_f$and unfoldedf_u accordingly to:

$$r(c) = (r_{0f} f_f + r_{0u} f_u)\left(1 + \rho \frac{Kc}{1+Kc}\right) \quad \text{(Eq. S7)}$$

where $r_{0f}$and r_{0u} are the root mean square interdye distance in presence of folded or unfolded domains in native buffer,

$$f_f = \frac{Exp[-m(c-c_{1/2})/RT]}{1+Exp[-m(c-c_{1/2})/RT]} \quad \text{(Eq. S8)}$$

and $f_u = 1 - f_f$, where$c_{1/2}$is the midpoint concentration and m the denaturant m value, representing the dependence of free energy on denaturant concentration. The stability parameterΔG_0can be computed as $\Delta G_0 = m\, c_{1/2}$.

Folding of RBD Domain.

While characterizing the NTD denaturant dependence, we discovered a plateau at transfer efficiencies between 1 and 2 M GdmCl, which we interpret as the contribution of the coexistence of

folding and unfolding conformations (**Eq. S7)**. To test whether this corresponds to the actual range of the folding transition, we designed, expressed, and labeled a construct with dyes in position 68 and 172, which directly monitors the folding of this domain. Single-molecule FRET measurements reveal up to three distinct populations (**Fig. S6**). One is abundant at high GdmCl concentration and disappears at low GdmCl concentrations and therefore we assign it as an unfolded state. Another one is only transiently populated between 1 and 2 M GdmCl and we assign it as an intermediate folding state. A third one, with a higher transfer efficiency compatible with the distance expected from the known RBD structure, is stabilized below 2 M GdmCl and, therefore, is assigned as the folded configuration. In absence of evident differences in brightness between these three species, the relative area of each state represents the fraction of the corresponding population. We use a three-state model where the fraction of each state can be computed from the partition function of the system, leading to:

$$f_u=\frac{1}{1+K^{u-i}+K^{i-f}};\ f_i=\frac{1}{1+(K^{u-i})^{-1}+K^{i-f}};\quad f_f=\frac{1}{1+(K^{u-i})^{-1}(K^{i-f})^{-1}+(K^{i-f})^{-1}} \qquad \text{(Eq. S9)}$$

where K^{u-i} and K^{i-f} are

$$K^{u-i}=Exp[-m^{u-i}(c-c_{1/2}^{u-i})/RT];\ K^{i-f}=Exp[-m^{i-f}(c-c_{1/2}^{i-f})/RT] \qquad \text{(Eq. S10)}$$

Fitted values to the model are reported in **Table S2.** Importantly, the observed values confirm in large measure the inferred stability measured via the NTD. The small discrepancy in the overall stability observed (**Fig. S9**) can either be assigned to the complicated decoupling of folding and chain expansion when observing the transition from the perspective of the NTD or by the "local" nature of the RBD unfolding probed by the NTD.

Salt Dependence of NTD, LINK, and CTD Conformations

In addition to studying the conformations under native buffer conditions, we investigate how salt affects the conformations of the three disordered regions. We started by testing the effects of electrostatic interactions on the NTD conformational ensemble. Moving from buffer conditions and increasing concentration of KCl, we observed a small but noticeable shift toward lower transfer efficiencies, which represents an expansion of the NTD due to screening of electrostatic interactions. This can be rationalized in terms of the polyampholyte theory of Higgs and Joanny [231,737] (see **Table S3**), where the increasing concentration of ions screens the interaction between oppositely charged residues (see **Fig. S11**).

We then analyzed for comparison the LINK FL construct. Interestingly, we find a negligible effect of salt screening on the root mean square distance of the low transfer efficiency population as measured by FRET (see **Fig. S11**). Predictions of the Higgs & Joanny theory (see SI) for the content of negative and positive charges within the LINK construct indicates a variation of interdye distance dimension that is comparable with the measurement error. It has to be noted that in this case the excluded volume term in the Higgs and Joanny theory will empirically account not only for the excluded volume of the amino acids in the chain, but also for the excluded volume occupied by the two folded domains.

To better understand the weak dependence on salt (and denaturant) of the dimensions LINK FL and the occurrence of two populations at low salt screening, we further investigated a truncated version of the same protein, the LINK-ΔDimer construct. First of all, we observe a sharp collapse as a function of GdmCl (**Fig. S8**), which starkly contrasts with the weak change of the LINK-FL. This strongly implies an effect of the two domains in modulating the dimensions of the LINK. We

then asked whether such modulation in a low denaturant regime contains a strong electrostatic component. To separate the effect of structural destabilization and electrostatic attraction in disordered proteins, we chose to use Urea. When comparing the conformation in the two denaturants, we clearly observed that Urea maintains the LINK-ΔDimer in a more compact configuration and by addition of 0.5 M KCl we can recover the expansion observed in GdmCl (**Fig. S10**). For comparison no change is observed when studying the NTD FL under the same conditions (**Fig. S10**). These observations for the LINK-ΔDimer mimic what was previously observed in the case of the Cold Shock Protein from *Thermotoga Maritima* [231] and confirms a strong electrostatic contribution in controlling the dimensions of the LINK region in absence of DIMER and CTD domains. It is reasonable to assume that similar electrostatic interactions are at play also in the full-length protein and are at the origin of the coexistence of two populations in low ionic strength solutions.

Finally, we test if the addition of salt can provide similar effects than those obtained by GdmCl on the conformations of the CTD: interestingly, we do not observe any significant variation either in transfer efficiency (**Fig. S11**), suggesting that the broadening of the population observed for the CTD does not originate exclusively from electrostatic interactions. However, when comparing the denaturing effect of GdmCl and Urea on the CTD-FL we observe more compact conformations of the chain in GdmCl.

Polymer Model of Electrostatic Interactions

The disordered regions of the N protein are enriched in positive and negative charges. To provide a term of comparison in the interpretation of protein conformations as function of salt concentration,

we use the polymer theory for polyampholyte solutions developed by Higgs and Joanny [231,737], which has been shown previously to capture quantitatively the conformational changes of unstructured proteins. Briefly, the root mean square interdye distance is equal$r = N^{0.5} l_0\ \alpha$ where $N$is the number of monomers in the disordered region, $l_0$is the length of elementary segment (here 0.36 nm) and α is the ratio between l and l_0, with $l$being a rescaled segment that accounts for excluded volume and electrostatic interactions.

α is computed according to the equation proposed by Higgs and Joanny [231,737]:

$$\alpha^5 - \alpha^3 = \frac{4}{3}\left(\frac{3}{2\pi}\right)^{1.5} N^{0.5}\ v^* \qquad \text{(Eq. S13)}$$

where v^* is an effective excluded volume given by the sum of three terms:

$$v^* b^3 = v\ b^3 + \frac{4\pi l_B (f-g)^2}{k^2} - \frac{\pi l_B^2 (f-g)^2}{k} \qquad \text{(Eq. S14)}$$

Here, v is the excluded volume (accounting for physical excluded volume and positive and attractive interactions that are not due to electrostatics), f and $g$are the fraction of positive and negative residue respectively for considered segment of the protein, k is the Debye screening length, and $l_b$is the Bjerrum length.

Importantly, when accounting for the fraction of negative charges, we also account for the contribution of the -2 net charge of each dye at pH 7.3.

5.19 Additional Methods

Testing Protein Oligomerization

NativePAGE experiments were performed to verify that purified recombinantly expressed SARS-CoV-2 N protein is capable of forming dimers and oligomers, in analogy to SARS-CoV N protein, and as shown in more recent work for SARS-CoV-2 [639,642,652]. Indeed, NativePAGE experiments reveal the existence of multiple bands (**Fig. S14 C-D**). However, since the lowest band in the NativePAGE corresponds to an apparent molecular weight of ~70-80 kDa, we wanted to verify the oligomeric state of this band.

To test whether the apparent mass is due to a slow mobility of the protein because of its high positive charge, we performed crosslinking experiments. These experiments confirm the formation of dimers, tetramers, and high oligomeric species, as a function of protein concentration above 500 nM (**Fig. S14 E-F**). These oligomeric species are in equilibrium with the monomer, the smallest species on the denaturing SDS PAGE (which has the expected molecular weight of ~45 kDa). It has to be noted that, because of the slow reactivity of the crosslinking agent (see Methods below), the crosslinking experiments do not represent the population of monomeric and oligomeric species at equilibrium. However, the comparison between the NativePAGE and the crosslinking experiments suggests that the smallest band in the NativePAGE is indeed the monomer protein. This suggests that the labeled protein can form higher oligomeric species in a concentration regime comparable to the one observed in NativePAGE and SDS PAGE experiments. Caution must be used in the interpretation of the oligomeric bound species observed in FCS experiments, since labeling mutation may have affected the affinity of the dimerization domain and the overall dimer size. Future experiments will address the role of labeling mutations on dimerization.

We finally turned to Fluorescence Correlation Spectroscopy (FCS) to test whether labeled protein can form dimers. We measured the CTD construct that carries one labeling position at the end of the oligomerization domain. When increasing the concentration of unlabeled protein, we observe a systematic increase in the hydrodynamic radius when compared to the hydrodynamic radius under native conditions (**Fig S14 A-B**). This suggests that the labeled protein can form higher oligomeric species in a concentration regime comparable to the one observed in NativePAGE and SDS PAGE experiments and that at 100 pM (the concentration used in single-molecule experiments), no oligomer is formed. Caution must be used in the interpretation of the oligomeric bound species observed in FCS experiments, since labeling mutation may have affected the affinity of the dimerization domain. Future experiments will address the role of mutation on dimerization. Finally, all experiments have been performed at two different time points, after 1 hour and after 24 hours of incubation of the labeled sample with unlabeled protein to test any kinetic effect on the measured value. No significant difference has been observed.

Taken together, NativePAGE crosslinking experiments support the fact that the protein can oligomerize. Together with the observation of similar transfer efficiencies in full-length and truncated variants of the proteins, these results further suggest that single-molecule experiments are monitoring the behavior of the monomeric SARS-CoV-2 N protein.

Protein Crosslinking Methods

50 mM disuccinimidyl suberate (DSS) (Thermo Scientific) stock solution was prepared (10 mg into 540 uL of anhydrous DMSO (Sigma)). All protein samples were prepared in 20 mM NaPi pH 7.4 (with and without 200 mM NaCl) at the following concentrations: 0.1, 0.5, 1, 5, 10 and 20uM. DSS stock solution was added to each sample to a final concentration of 1.25 mM. Samples were

incubated for 1 hour at room temperature. Samples were then quenched to a final concentration of 200 mM Tris pH 7.4 and allowed to incubate for 15 minutes. Crosslinked proteins were then analyzed using SDS PAGE and Coomassie staining.

NativePAGE Methods

All protein samples were prepared in 20 mM NaPi pH 7.4 (with and without 200 mM NaCl) at the following concentrations: 0.05, 0.1, 0.5, 1, 5, 10 and 20 μM. Samples were subjected to NativePAGE (Invitrogen) and protein mobility was analyzed with Coomassie staining.

Turbidity Measurements.

Development of turbidity in solutions of N protein and poly(rU) was followed through measurements of absorbance at 340 nm in a microvolume spectrophotometer (NanoDrop, Thermo, USA). Mixtures were prepared in 500 μl plastic reaction tubes by adding 4 μl protein solution into 3 μl of poly(rU) and absorbance was recorded 45 s – 75 s after mixing. Working solutions were kept at room temperature during experiments. Reaction media was 50 mM Tris, pH 7.5 (HCl), 0.002 % v/v Tween20, and NaCl as indicated in *Results*.

poly(rU) (Midland Certified Reagent Company, TX, USA, lot number 011805) was reconstituted into this media from stocks dissolved in RNAse free water. According to the manufacturer, the size of poly(rU) molecules is mostly less than 250 nucleotides (nt.) and longer than 200 nt. Protein stocks (in 50 mM Tris pH 8.0, 500 mM NaCl, 10% v/v glycerol) were buffer exchanged into the desired buffer through size exclusion chromatography in Zeba Spin 7 k MWCO desalting columns (Thermo, USA). poly(rU) concentrations in working dilutions were assessed through the absorbance

at 260 nm employing an extinction coefficient of 9.4 mM^{-1} cm^{-1} [738]. Protein concentrations were assessed through the absorbance at 280 nm employing an extinction coefficient of 42.53 mM^{-1} cm^{-1}, computed according to the method proposed by Pace *et al.* [739].

The limiting concentrations of nucleic acid across which an increase in turbidity was detected were estimated through interpolation of the data. To this end, an empirical equation, describing the trends observed at all concentrations, was fitted to the data and then was solved to extract the poly(rU) concentrations at which turbidity reaches a limit value above the background signal. We used a limiting absorbance value of 0.005 units (340 nm, 1 mm path length). We found that an appropriate function for this end is an exponential of a Gaussian distribution function $F(x)$:

$$F(x) = A(1 - Exp[-\beta\gamma(x)]) \quad \text{(Eq. S11)}$$

where

$$\gamma(x) = \frac{1}{(2\pi)^{0.5}\sigma} Exp[-(x-\mu)^2/2\sigma^2] \quad \text{(Eq. S12)}$$

where x denotes poly(rU) concentration and A, β, σ and μ are parameters fitted through weighted minimum least squares for each protein concentration (solid lines in **Fig. 5 A-B** and limiting value points in panels **C-D**). To characterize the observed global trends of turbidity, as a function of both RNA and protein concentration, we determined approximate functional forms of the dependence on protein concentration of the individually fitted parameters ($A(p)$, $\beta(p)$, $\sigma(p)$ and $\mu(p)$, where p is protein concentration). The observed dependencies were increasing linearly for $\mu(p)$ and quadratic for $\beta(p)$ and $\sigma(p)$. A was the worst defined parameter and thus displayed the least clear trend. For the results in absence of added salt we employed an increasing power function with exponent as a fitting parameter (best fit value was < 1), whereas for the results in presence of 50 mM NaCl the trend of $A(p)$ was better described by a decreasing exponential function.

We thus used the functional forms $A(p)$, $\beta(p)$, $\sigma(p)$ and $\mu(p)$ to construct a global function dependent on both protein and RNA concentration. Global fitting of this equation to the whole set of turbidity titration curves provided the turbidity contour plots shown in **Fig. 5 C-D** (solid lines). Contour lines were computed at 1, 10, 20, 50 and 100 times the limiting value employed ($A_{340\ nm,1mm} = 0.005$).

5.20 Supplementary Figures

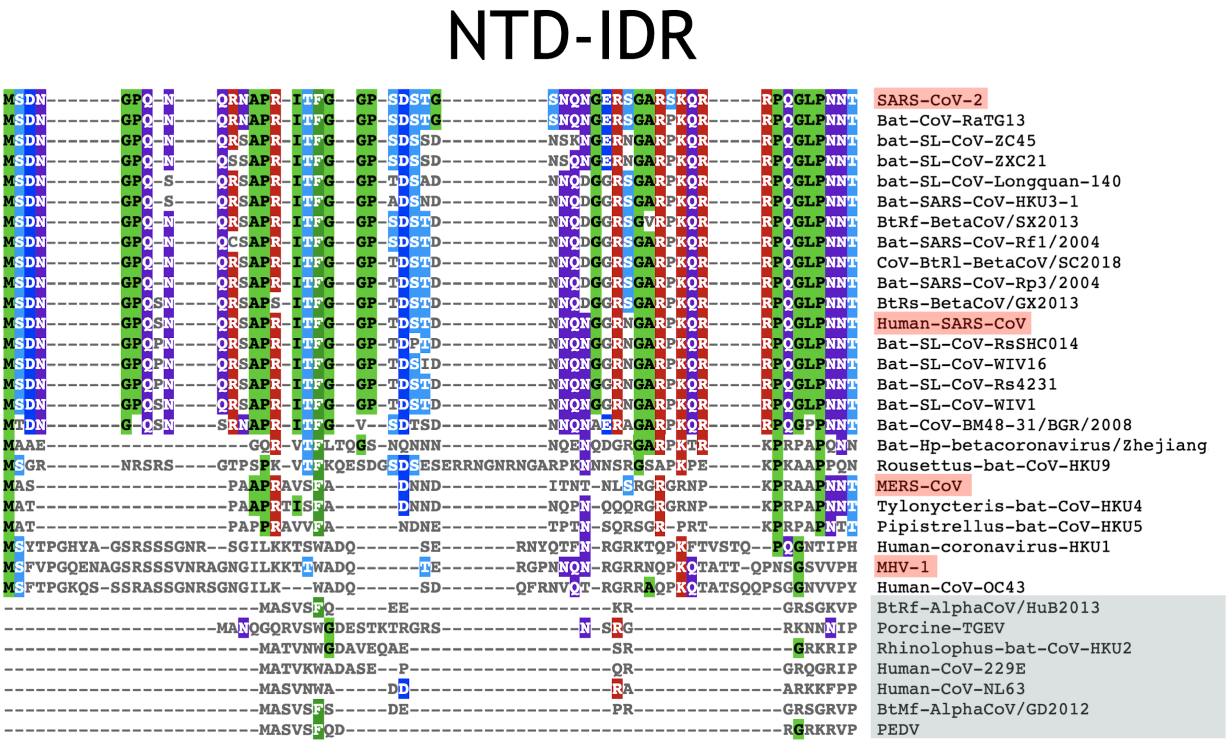


Fig. S1. Sequence alignment of the coronavirus N-terminal domain (NTD)

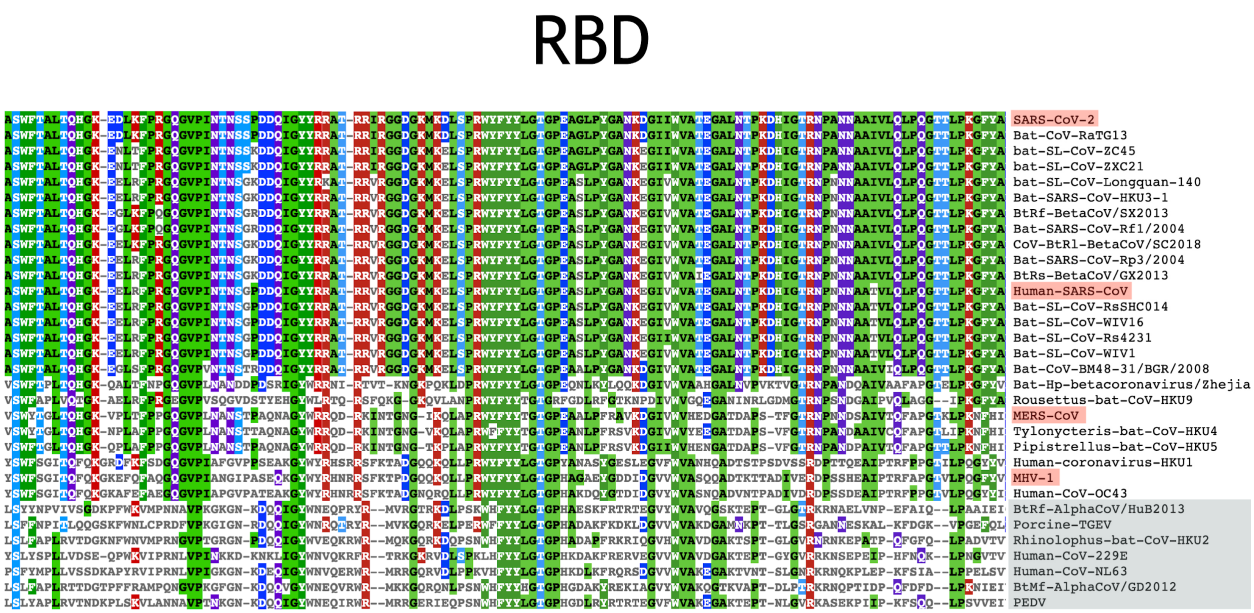


Fig. S2. Sequence alignment of the coronavirus RNA binding domain (RBD)

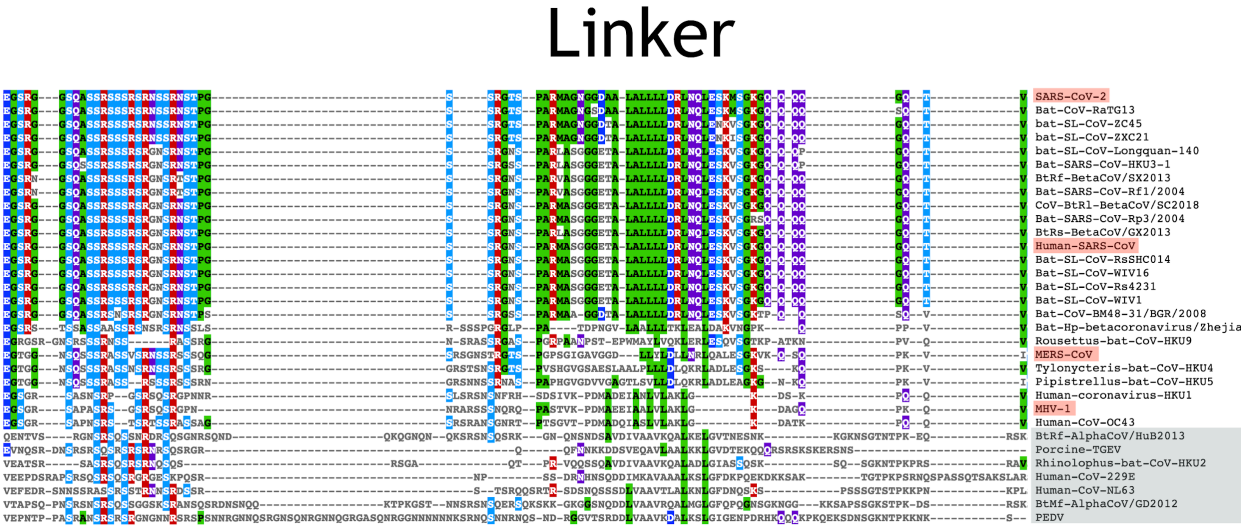


Fig. S3. Sequence alignment of the coronavirus linker (LINK)

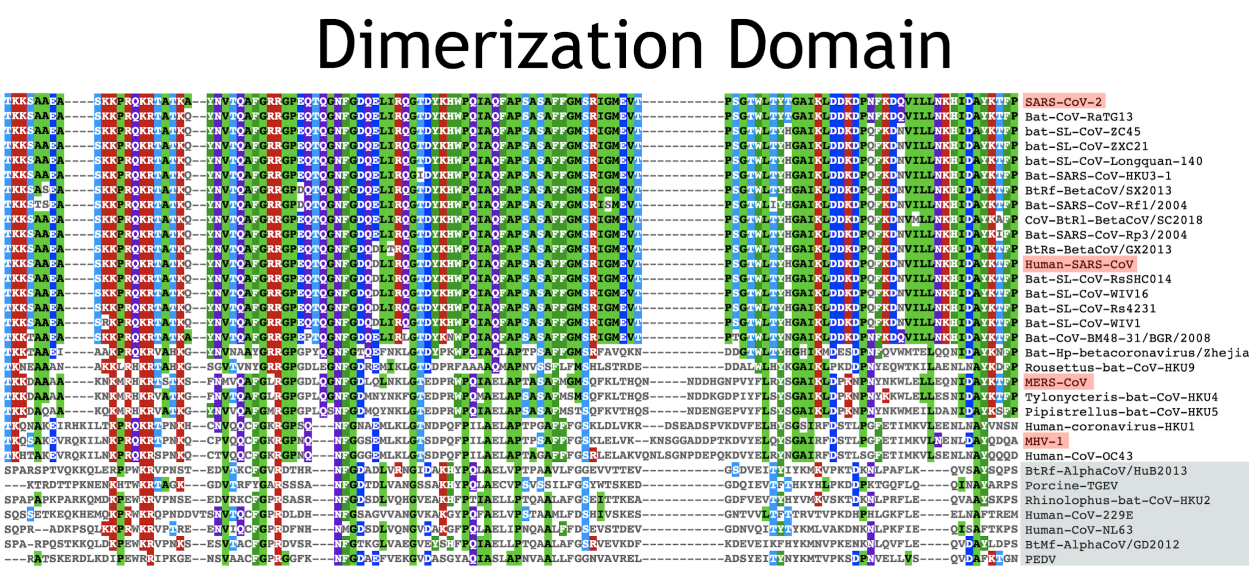


Fig. S4. Sequence alignment of the coronavirus dimerization domain

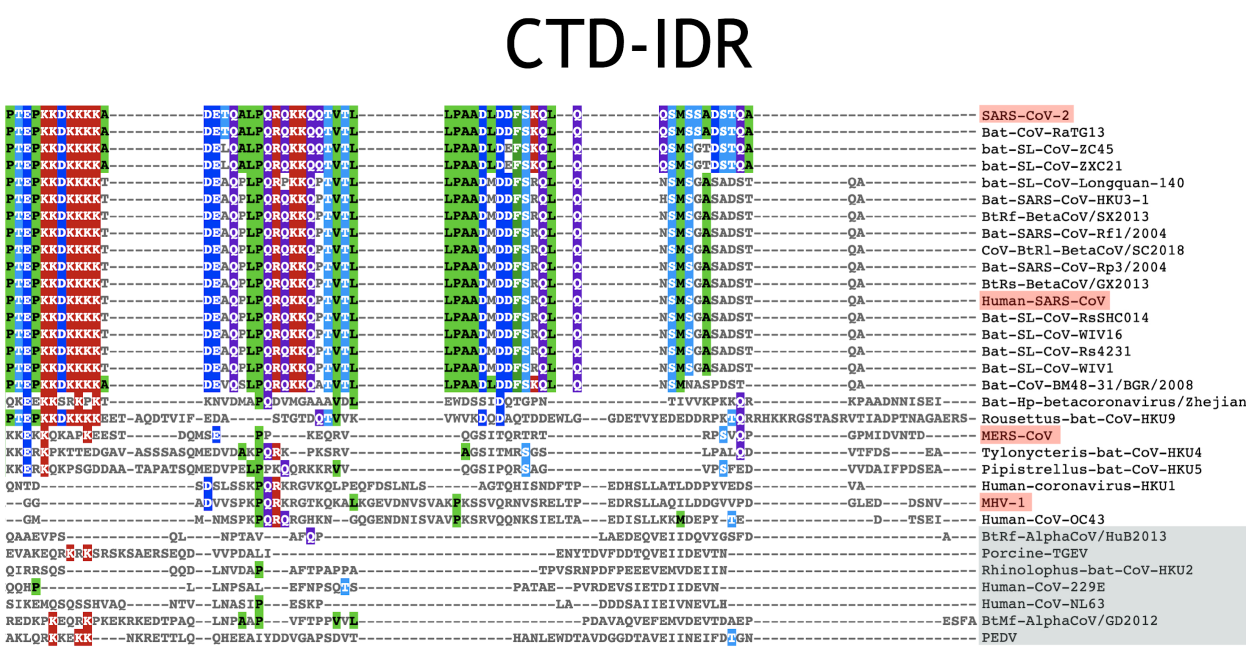


Fig. S5. Sequence alignment of the coronavirus C-terminal domain (CTD)

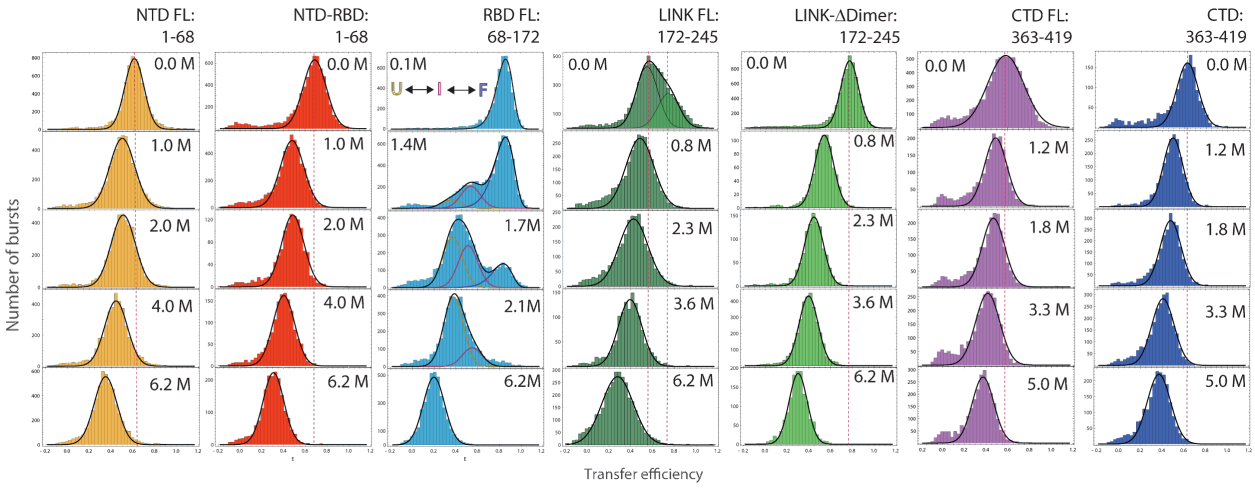


Fig. S6. Histograms of transfer efficiency distributions across GdmCl concentrations:

for NTD FL (orange), NTD-RBD (red), RBD FL (cyan), LINK FL (dark green), LINK-ΔDimer (green), CTD FL (purple) and CTD fragment (blue) constructs.

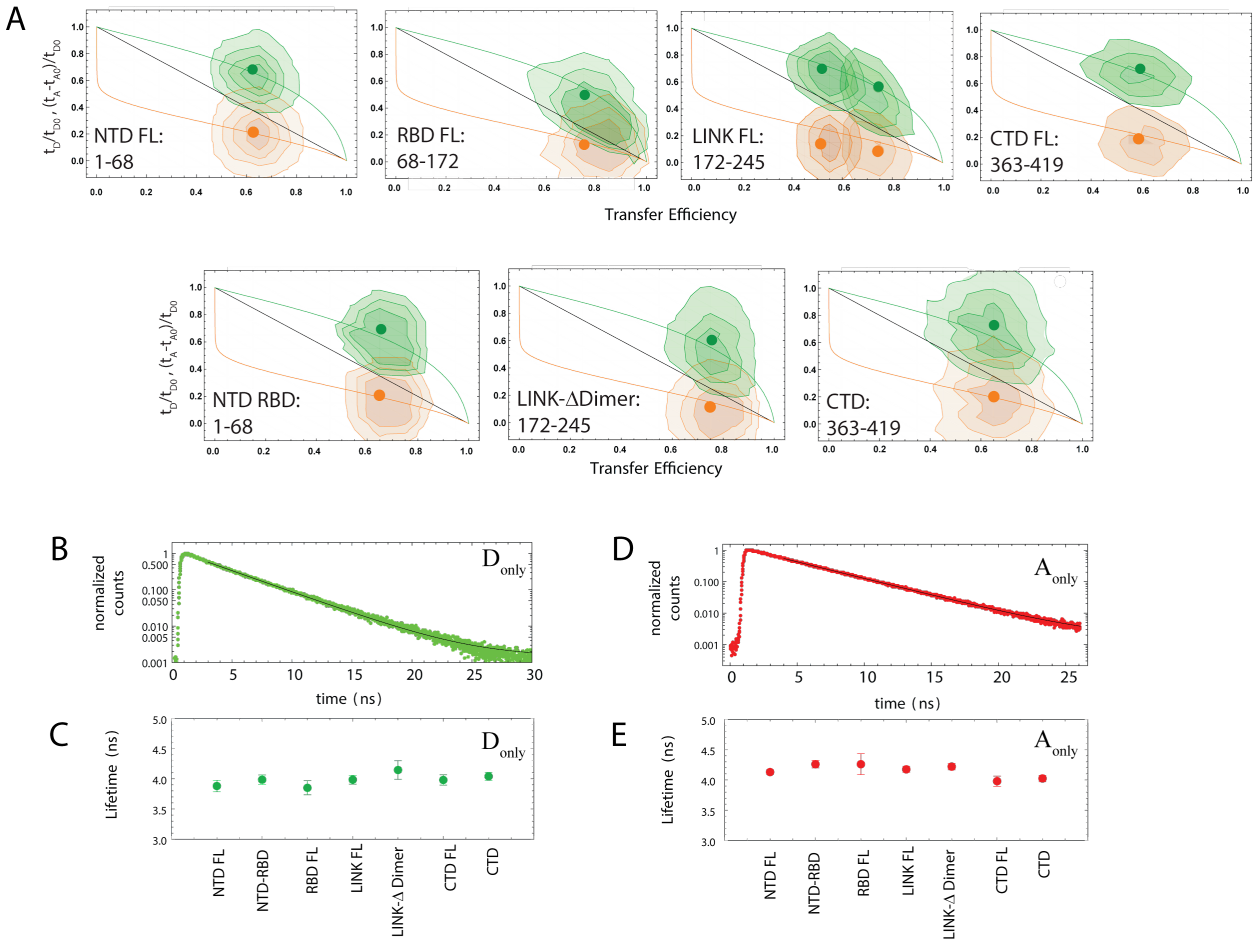


Fig. S7. Dependence of fluorescence lifetime on transfer efficiency

A. NTD FL, RBD FL, LINK FL, CTD FL, NTD RBD, LINK-ΔDimer, and CTD construct. Black line: linear dependence expected for a rigid molecule. Green line: the donor lifetime (normalized by the donor lifetime in absence of FRET: t_D/t_{D0}) in the limit of dynamics much faster than the burst duration but slower than the fluorophore lifetime. Orange line: the acceptor lifetime delay (normalized by the donor lifetime in absence of FRET: $(t_A-t_{A0})/t_{D0}$). The green and orange contour plots represent the corresponding distributions of donor lifetime and acceptor lifetime delay as observed in single-molecule experiments under native conditions (**Fig. 2A, 3A, 4A**). The green and orange dots represent the mean value of the measured distributions. **B.** Example of lifetime measurements extracted from the donor-only population and corresponding tail fit. **C**. Observed

lifetimes for each construct under aqueous buffer conditions as extracted from the tail fit. **D.** Example of acceptor lifetime measurement from the acceptor only population and corresponding tail fit. **E.** Corresponding acceptor lifetime in aqueous condition for each construct. No significant dynamic quenching is observed in both donor and acceptor. This does not exclude the possible occurrence of static quenching (see **Fig. S12**). Data in panels **C** and **E** are presented as mean ± standard deviation.

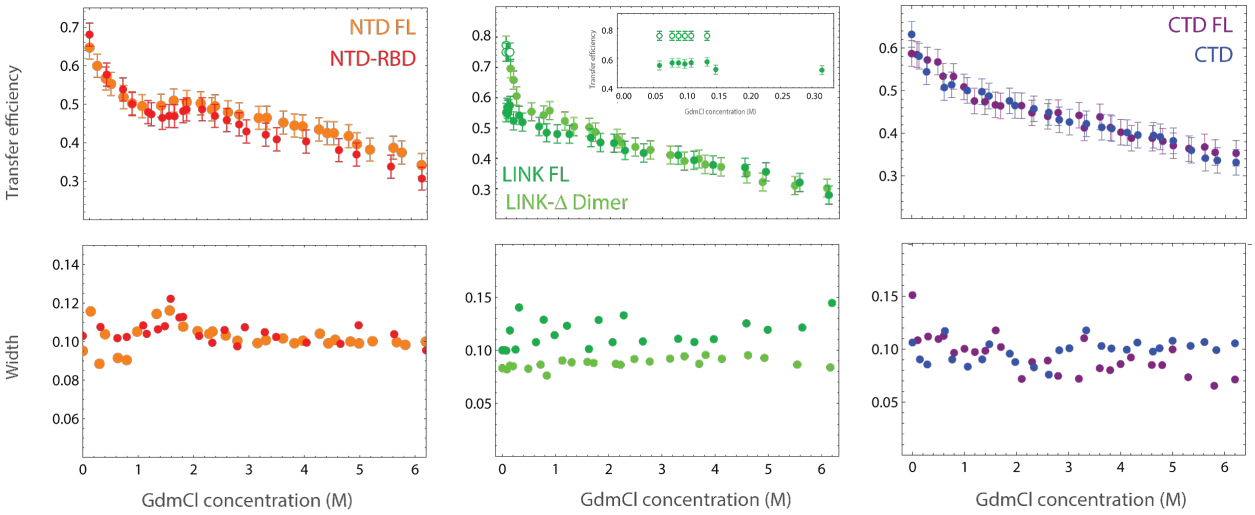


Fig. S8. Mean transfer efficiency and width of NTD FL vs NTD-RBD, LINK FL vs LINK-ΔDimer, CTD-FL vs CTD fragment across GdmCl concentration

The mean transfer efficiency of the NTD FL domain (orange) exhibits a plateau between 1 and 2 M GdmCl; at the same concentration we observe a small but systematic increase in the amplitude of the transfer efficiency distribution hinting to the coexistence of two populations in slow exchange with very similar transfer efficiencies. The same behavior is closely reproduced by the truncated variant NTD-RBD (red). The LINK FL (dark green) exhibits two distinct populations at very low GdmCl concentration (open and close circles), suggesting a strong contribution of electrostatics in favoring one of the two configurations. Inset shows coexistence of the two states between 0 M and 0.75 M GdmCl. The truncated variant LINK-ΔDimer (green) shows a continuous collapse that interpolates the two positions observed for the LINK FL, suggesting interaction of the LINK with itself or with the RBD domain in absence of the DIMER domain. Finally, the CTD FL (blue) and the CTD fragment (purple) exhibit similar conformations across denaturant concentrations. The small increase in the width of the transfer efficiency distribution that may reflect the formation of local structure under native conditions (e.g. the putative helical binding motif). Transfer efficiencies

data represent the mean value of the corresponding distribution ± 0.03 systematic error in measured transfer efficiencies due to instrument calibration (see **FRET histograms** section in **SI**).

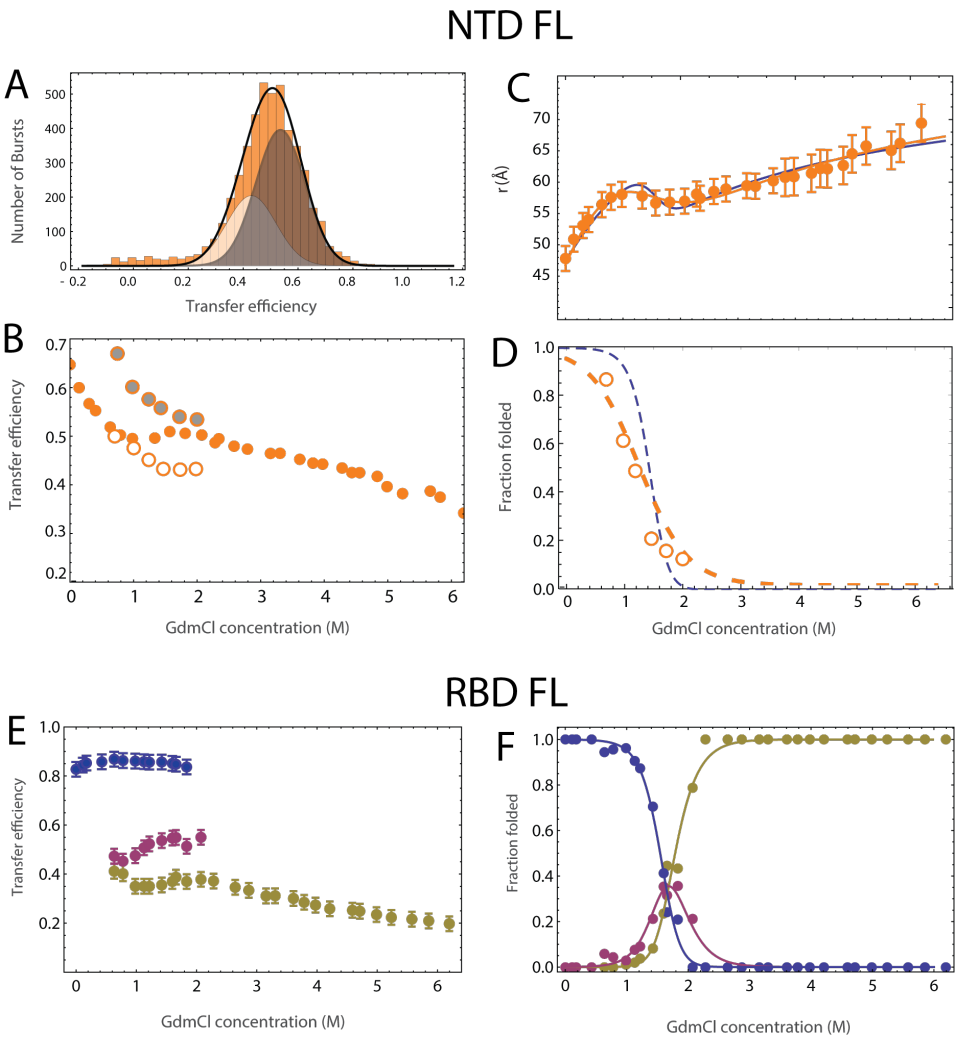


Fig. S9. Fit of NTD construct with two populations compared to folding of RBD domain

To address the change in amplitude that occurs from the NTD construct between 1 and 2 M GdmCl, we attempt a fit of the NTD FL data using two populations with a fixed width equal to average width below 1 M and above 2 M GdmCl (see for comparison **Fig. S8**) **A.** Fit of the transfer efficiency histogram at 1.5 M GdmCl. The white- and gray- shaded areas reflect fits to the "folded RBD" population and to the "unfolded RBD" population. *Central panel:* Comparison of transfer efficiencies with a single fit (solid orange circles, compare **Fig. S8**) and from the two populations: gray solid circles for the "unfolded RBD" population and unfilled circles for the "folded RBD"

population. *Lower panel*: Fraction folded estimated from the fit with **Eq. S7** compared to the fraction of "folded RBD" obtained from computing the ratio between the area under "folded RBD" species and the total histogram area. Transfer efficiencies in E are presented as the mean value of the corresponding distribution ± 0.03 systematic error in measured transfer efficiencies due to instrument calibration (see **FRET histograms** section in **SI**).

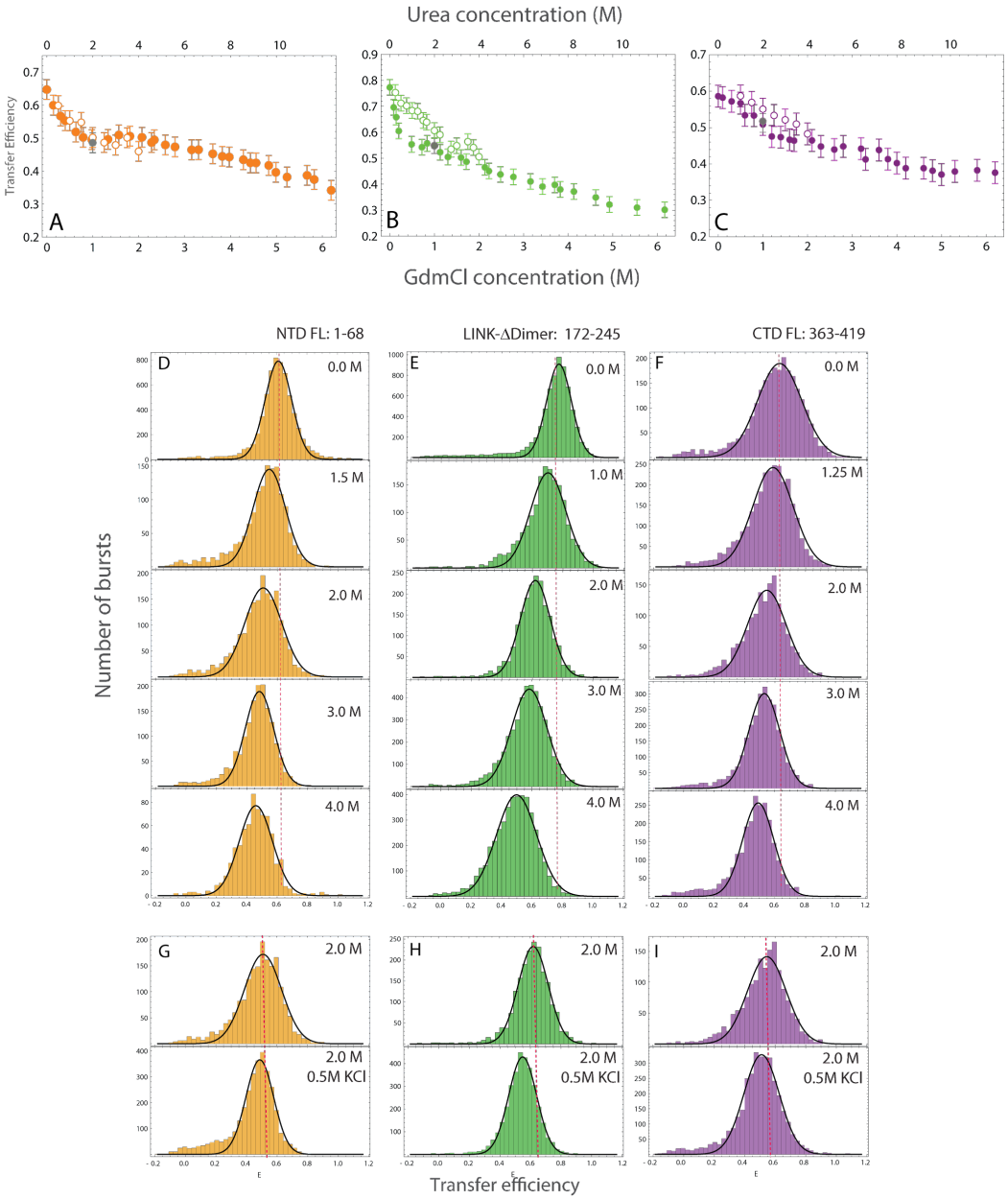


Fig. S10. Effects of Urea denaturation on NTD FL, LINK-ΔDimer, and CTD FL

A-C. Comparison of Urea (open circles) and GdmCl (close circles) effects on the transfer efficiencies of NTD FL (orange), LINK-ΔDimer (green), and CTD-FL (purple). The Urea range is rescaled by a factor of 2 compared to the GdmCl range to account for the different denaturing effect. Grey dots correspond to the same concentration of Urea with the addition of 0.5 M KCl.

Data represent the mean value of the distribution ± 0.03 systematic error in measured transfer efficiencies (see **FRET histograms** section in **SI**). **D-F.** Examples of transfer efficiencies distribution as function of Urea. **G-I.** Comparison between 2 M Urea histograms in presence and absence of 0.5 M KCl.

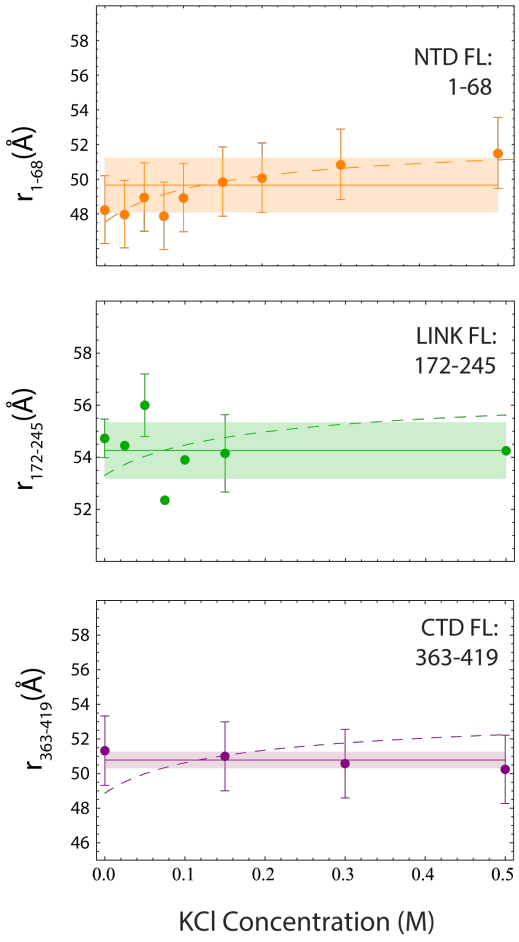


Fig. S11. Interdye distances of NTD, LINK, CTD in presence of salt (KCl)

Upper panel: root mean square interdye distance between position 1 and 68. Dashed line: fit according to the Higgs & Joanny model **(Eq. S11-12)** predicts a comparable change to the one observed. *Central panel:* root mean square interdye distance between position 172 and 245. Dashed line: fit according to the Higgs & Joanny model (**Eq. S11-12**) predicts a comparable change to the one observed. Solid line and shaded area: average value of the root-mean-square interdye distance across all salt conditions and corresponding standard deviation. The standard deviation is comparable to the measurement error. *Lower panel*: root mean square interdye distance between position 363 and 419. Dashed line: fit according to the Higgs & Joanny model (**Eq. S11-12**) does not capture the observed trend. This can be possibly explained considering the significant predicted population of

helical conformations in the CTD. Solid line and shaded area: average value of the root-mean-square interdye distance across all salt conditions and corresponding standard deviation. All measured root means square distances are presented as the value corresponding to the mean of the transfer efficiency distribution ± 0.03 systematic error in measured transfer efficiencies (see **FRET histograms** section in **SI**).

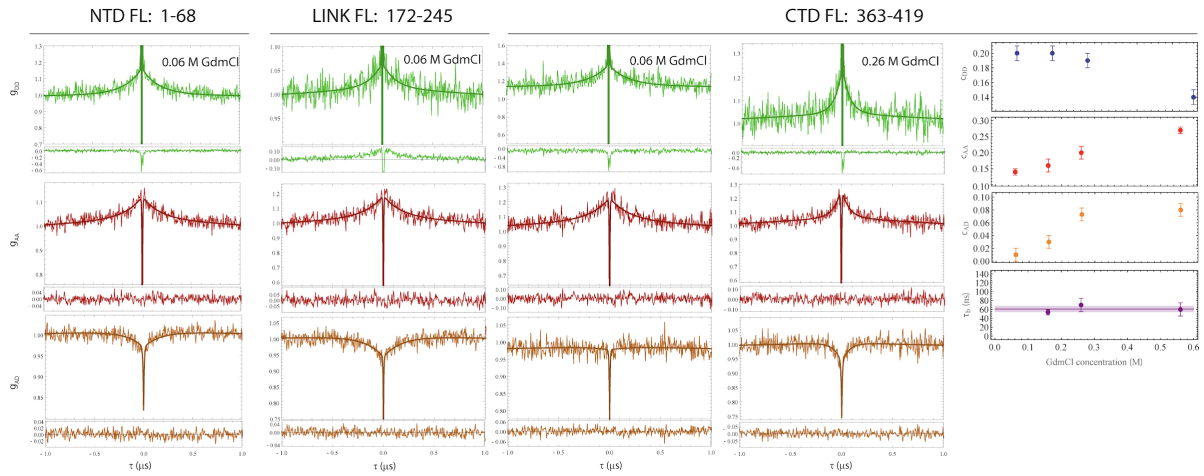


Fig. S12. Chain dynamics measured via ns-FCS

Nanosecond FCS measurements for the NTD, LINK, and CTD constructs provide a measure of the dynamics on the nanosecond timescale. All correlations are normalized to the value measured at 1 μs for highlighting the amplitude relative to the reconfiguration term. The donor-donor (green), acceptor-acceptor (red), and donor-acceptor (orange) correlation are fitted to a global model that accounts for antibunching, FRET dynamic populations, and triplet. The acceptor-donor correlation shows a clear anticorrelated change for NTD FL and LINK FL in the signal that reflects the anticorrelated nature of the donor-acceptor energy transfer as a function of distance: an increase in acceptor reflects a decrease in donor. The CTD FL cross-correlation exhibits a flat behavior, which is consistent with absence of dynamics or compensation between two populations, one anti-correlated (dynamic) and one correlated (static). [204 657]. Addition of GdmCl (e.g., 0.26 M) causes a decrease in the transfer efficiency distribution width **(Fig. S8**) and leads to the appearance of an anticorrelated increase in the cross-correlation of CTD. A plot of the corresponding change in amplitude for the donor-donor, acceptor-acceptor, and acceptor-donor correlation is shown for

comparison. We interpret the decrease in the donor-donor correlation and the increase in the acceptor-acceptor and acceptor-donor correlations as the result of destabilization of the quenched species in favor of the dynamic population. A decay correlation time can be globally fitted starting from 0.16 M GdmCl and appears to be constant across the measured values, up to 0.6 M GdmCl. The average decorrelation time t_{CD} is equal to 61 ± 7 ns. For comparison, the correlation decay in the donor-donor and acceptor-acceptor autocorrelations at 0 M GdmCl hold characteristic times of 80 ± 20 ns and 110 ± 20 ns respectively. Fitted amplitudes and times are presented as best fit values ± the error of the fit.

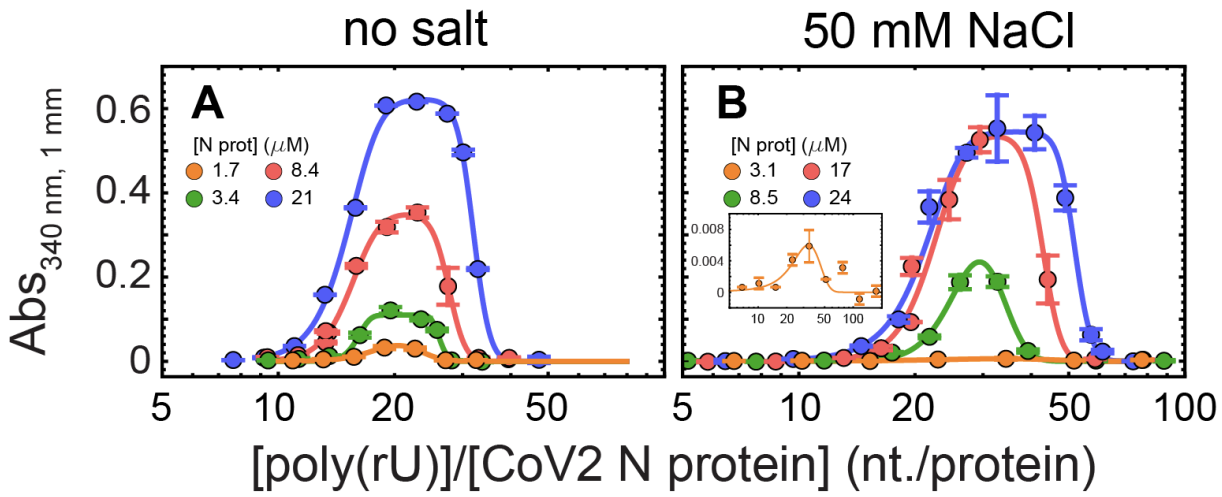


Fig. S13. Turbidity experiments plotted against RNA/protein ratio

Representative turbidity titrations with poly(rU) in 50 mM Tris, pH 7.5 (HCl) at room temperature, in absence of added salt **(A)** and in presence of 50 mM NaCl **(B),** at the indicated concentrations of N protein. On the x-axis, the concentration of poly(rU) is rescaled for the protein concentration. Points and error bars represent the mean and standard deviation of 2 (absorbance < 0.005) and 4 (absorbance ⩾ 0.005) consecutive measurements from the same sample. Solid lines are simulations of an empirical equation fitted individually to each titration curve. An inset is provided for the titration at 3.1 μM N protein in 50 mM NaCl to show the small yet detectable change in turbidity on a different scale. Interestingly, within the experimental error, we observe a clear alignment of the turbidity curves with a maximum at ~20 nucleotides per protein in the absence of added salt (A) and ~30 nucleotides per protein in the presence of 50 mM NaCl (B).

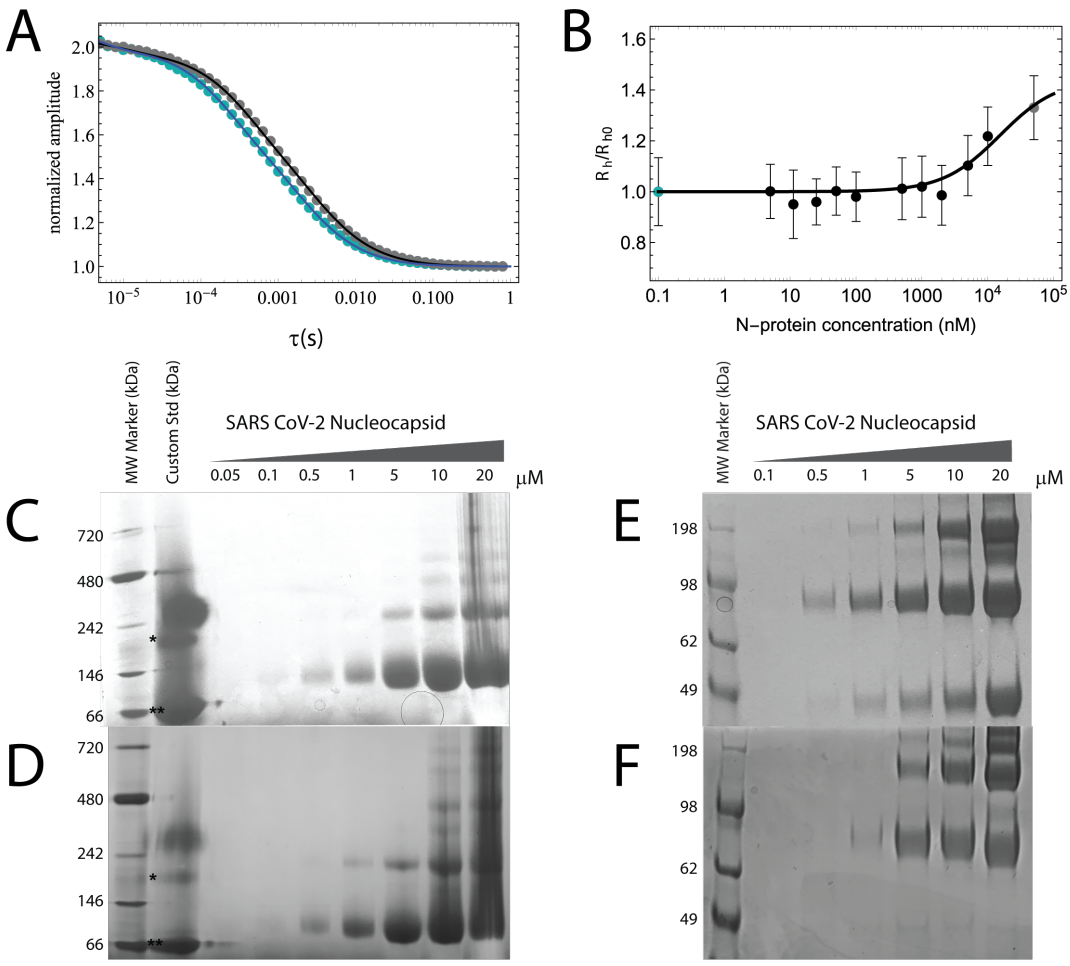


Fig. S14. Testing SARS-CoV-2 N protein oligomerization

(A-B) Fluorescence Correlation Spectroscopy (FCS) of full-length SARS-CoV-2 N protein as a function of protein concentration. **A.** FCS traces of 100pM Alexa 488/Alexa 594 N protein labeled at positions 363 and 419 in the absence (blue dots) and the presence (gray dots) of 50 μM unlabeled N protein. **B.** Hydrodynamic radius of SARS-CoV-2 N protein obtained from FCS (blue dot: 100 pM labeled N protein; gray dot: 100 pM labeled N protein + 50 μM unlabeled N protein) normalized to the protein dimensions determined in aqueous buffer conditions. Error bars represent propagation of errors (standard deviation) measured for the hydrodynamic radius at each N protein concentration. **C-D.** NativePAGE of full-length SARS-CoV-2 N protein in 20 mM NaPi pH 7.4 as

a function of protein concentration in the presence of 200 mM NaCl (C) and in the absence of added salt (D). 'Custom Std' lane contains Alcohol Dehydrogenase (* , 150 kDa) and Bovine Serum Albumin (** , 66 kDa). **E-F.**SDS PAGE of crosslinked full-length SARS-CoV-2 N protein in 20 mM NaPi pH 7.4, 1.25 mM DSS as a function of protein concentration in the presence of 200 mM NaCl (E) and in the absence of added salt (F). Each gel was repeated to confirm results.

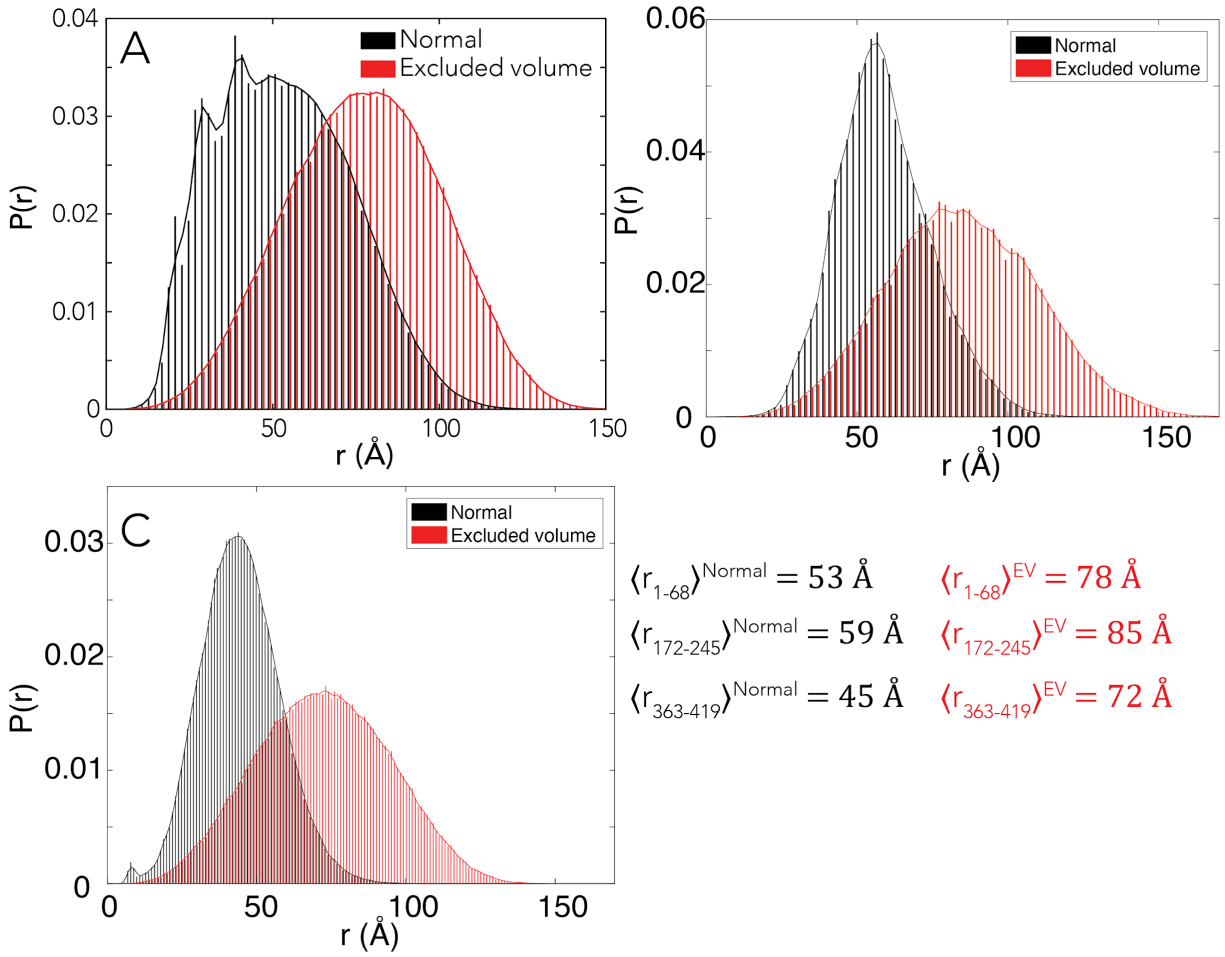


Fig. S15. Distributions of inter-residue distance from ABSINTH simulations (black) vs. excluded volume simulations (red)

Comparison of simulations with the full ABSINTH Hamiltonian (normal, black) against simulations performed in the excluded volume (EV, red) limit for **A.** NTD in the NTD-RBD context, **B.** LINK in the NTD-LINK-DIM context, and **C.** CTD in the DIM-CTD context. In all three cases, the EV simulations are performed in the analogous structural context, and report substantially larger average distances than the ABSINTH simulations, as expected given the absence of any attractive intramolecular interactions. The distances reported from the EV simulations are also slightly more

expanded than under fully denatured conditions, consistent with systems studied previously (see previous work [91,740]).

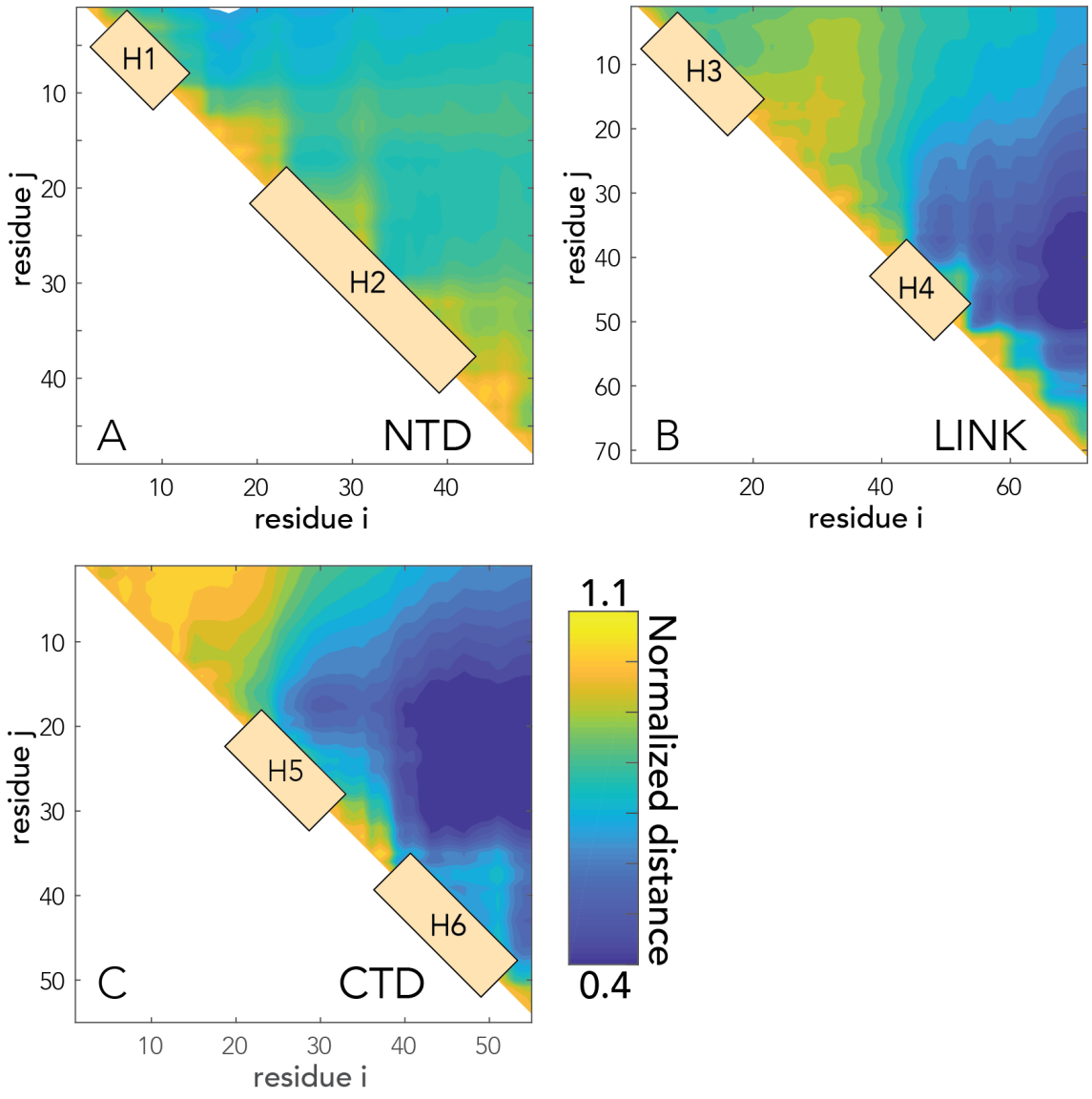


Fig. S16. Scaling maps for IDR-only simulations

Scaling maps report on the normalized distance between pairs of residues, where normalization is done by the distance expected if the IDRs behaved as self-avoiding chains in the excluded-volume

limit. Scaling maps for IDR-only simulations of the **A.** NTD, **B.** LINK and **C.** CTD. For each sequence, transient helices are annotated on the scaling maps. Note that in the LINK we observe interaction between the C-terminal region of the LINK and H4, while H3 does not interact with any parts of the sequence. Similarly, in CTD we see extensive intramolecular interactions between H5 and H6.

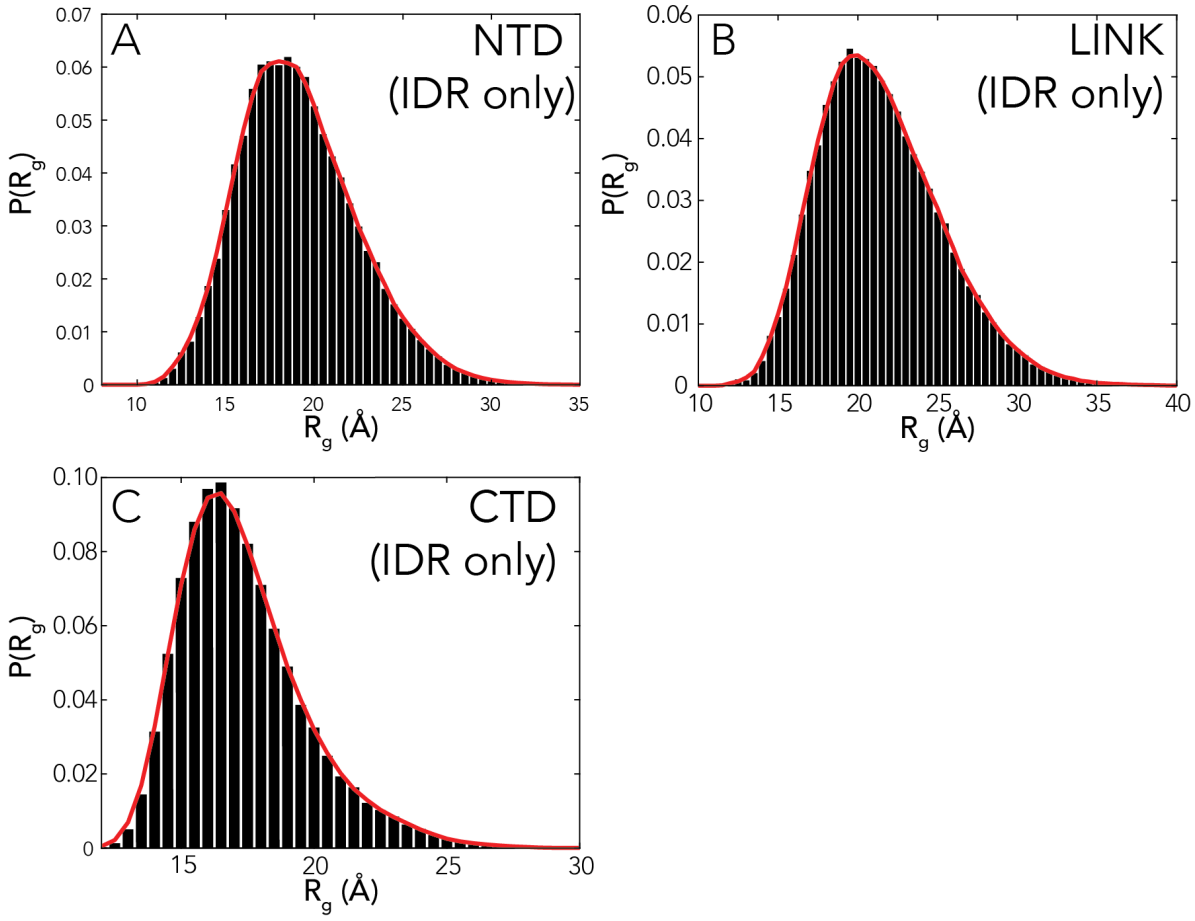


Fig. S17. Distributions for the radius of gyration (R_g) of for IDR-only simulations

R_g distributions for **A.** NTD, **B.** LINK and **C.** CTD. Average R_g for each IDR in isolation is 19.1 Å (NTD), 21.4 Å (LINK), and 17.1 Å (CTD).

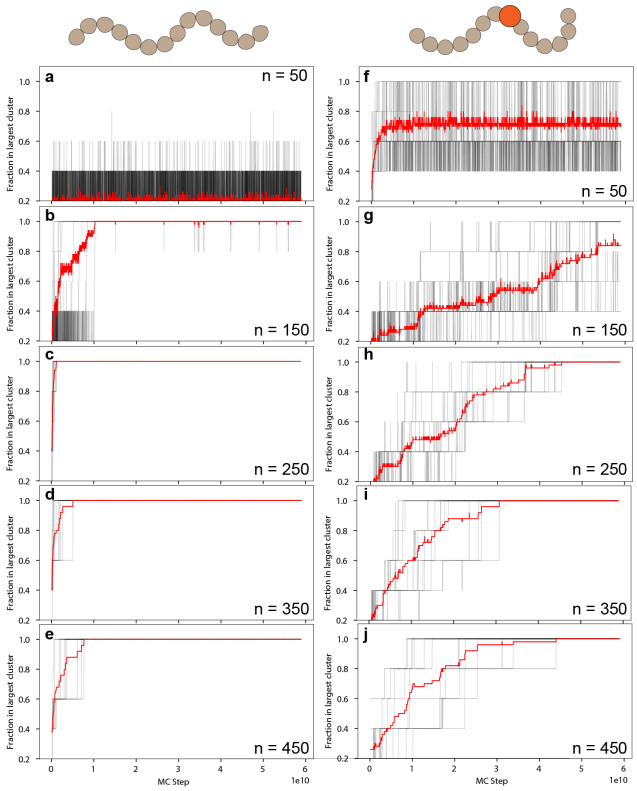


Fig. S18. Monte Carlo simulations reveal slow pseudo-kinetics of condensate fusion

Our simulations in **Fig. 6** reveal single-polymer condensates in the presence of a high-affinity binding site, whereas multi-chain droplets assemble in the absence of a high-affinity binding site. To further explore the origin of single-polymer condensates we ran extensive Monte Carlo simulations using an approximate kinetics scheme (that includes cluster translation moves) to examine the pseudo-kinetics of assembly. Black lines in each panel correspond to individual simulation trajectories, while red lines report on the average behavior over ten independent simulations. *n* reflects the number of binder chains in each simulation, and for each 5 separate polymers are present. To assess the apparent kinetics of assembly, we asked what fraction of the total number of polymers are found in the largest cluster. Under conditions in which a single droplet forms 100% of the polymer chains will be found in the largest cluster. Panels a,b,c,d,e report on behavior for

polymers without a high affinity binding site. In all cases within 10^9 Monte Carlo steps every independent simulation has converged on a single multichain droplet that represents the thermodynamic minimum expected for a two-phase equilibrium. Panels f,g,h,i,j report on identical simulations performed with a single high affinity binding site. While these simulations trend towards or reach a single multichain condensate, the presence of a high-affinity binding site substantially retards the assembly kinetics, revealing a large regime over which single-polymer condensates are metastable.

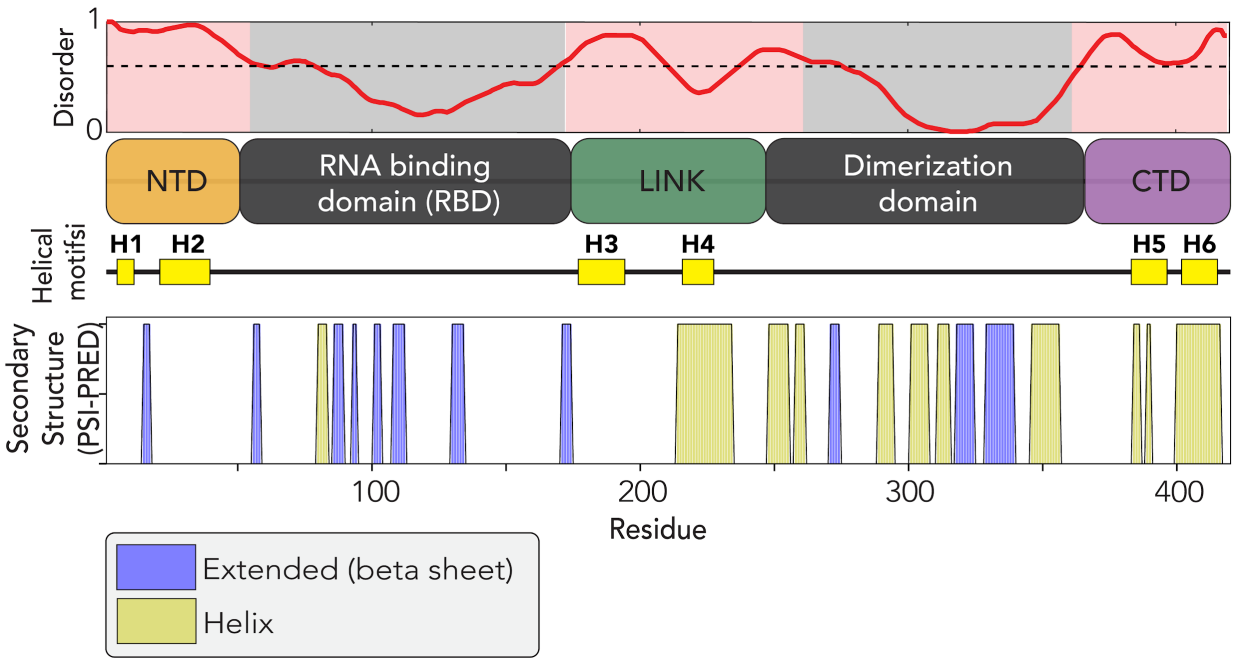


Fig. S19. Comparison of secondary structure in IDRs from bioinformatics predictions

We computed secondary structure propensities for the full-length protein using the PSI-PRED prediction server [741]. This analysis correctly identifies helices H4, H5 and H6, but fails to identify those H1, H2 and H3. Helix H3, H4 and H6 have been similarly identified by NMR and/or hydrogen-deuterium exchange mass spectroscopy [392,636,651]. These results demonstrate that our simulations are able to identify predicted helices but, furthermore, find helices that conventional structural bioinformatics software fails to correctly identify.

Chapter 6: The disordered N-terminal tail of SARS CoV-2 Nucleocapsid protein forms a dynamic complex with RNA

This chapter was published on Bioarxiv and Under Review at the journal Nucleic Acid Research as:

The disordered N-terminal tail of SARS CoV-2 Nucleocapsid protein forms a dynamic complex with RNA. Cubuk J, Alston JJ, Incicco JJ, Holehouse AS, Hall KB, Stuchell-Brereton MD, Soranno A. *bioRxiv* 2023.02.10.527914; doi: https://doi.org/10.1101/2023.02.10.527914

Contributions. J.C. expressed, purified, and labeled all protein constructs. J.C. performed all single-molecule experiments with single-stranded RNA and folding stability of RBD_L, including nanosecond FCS measurements. J.J.A. performed all single-molecule experiments with double-stranded RNA, folding stability measurements of RNA, and simulations. K.H. and J.J.A. *in-vitro* transcribed and purified RNA. J.C. and M.D.S.-B. designed the NTD-RBD_L and RBD_L nucleic acid binding assay. J.J.I. developed analytical tools for binding models. J.J.A. and A.S.H. developed computational tools for simulations. J.C., K.B.H, M.D.S.-B, J.J.I, J.J.A, A.S.H., and A.S. wrote the paper. M.D.S.-B. and A.S. supervised experiments and data analysis. J.J.I, J.C., M.D.S.-B., K.H., and A.S. conceived the experiments.

6.1 Abstract

The SARS-CoV-2 Nucleocapsid (N) protein is responsible for condensation of the viral genome. Characterizing the mechanisms controlling nucleic acid binding is a key step in understanding how condensation is realized. Here, we focus on the role of the RNA Binding Domain (RBD) and its flanking disordered N-Terminal Domain (NTD) tail, using single-molecule Förster Resonance Energy Transfer and coarse-grained simulations. We quantified contact site size and binding affinity for nucleic acids and concomitant conformational changes occurring in the disordered region. We found that the disordered NTD increases the affinity of the RBD for RNA by about 50-fold. Binding of both nonspecific and specific RNA results in a modulation of the tail configurations, which respond in an RNA length-dependent manner. Not only does the disordered NTD increase affinity for RNA, but mutations that occur in the Omicron variant modulate the interactions, indicating a functional role of the disordered tail. Finally, we found that the NTD-RBD preferentially interacts with single-stranded RNA and that the resulting protein:RNA complexes are flexible and dynamic. We speculate that this mechanism of interaction enables the Nucleocapsid protein to search the viral genome for and bind to high-affinity motifs.

6.2 Introduction

The SARS-CoV-2 virus is a positive-sense single-stranded RNA coronavirus with a genome of nearly 30000 nucleotides[742]. This large genome is packaged into small viral particles of ~100-□□ nm diameter[743]. Such a degree of packaging is mediated by the interaction of the viral genome with multiple copies of the Nucleocapsid (N) protein. The "beads on a string structures"[450,744] formed by the SARS-CoV-2 N protein inside the virion are at variance with previously proposed helical structures seen in other coronaviruses[745,746] and the mechanism of their formation is not well understood. From a biophysical standpoint, the compaction of a single viral genome and the phase separation of the protein with multiple nucleic acids potentially stem from the same set of interactions[441]. Independent experiments from many labs (including ours) have demonstrated that N protein can undergo phase separation with nucleic acid, both *in vitro* and in living cells[392,442,464,469,637,747–750]. Phase separation can be favored by specific RNA sequence motifs[442] and altered, in cells, by interactions with small molecules[751]. Quantifying the molecular interactions at play is therefore key to identifying the processes controlling condensation on the single- and multi- chain scale.

The SARS-CoV-2 N protein shares a similar domain architecture to analogous N proteins from other coronaviruses, including an RNA Binding Domain (RBD), a dimerization domain, and three intrinsically disordered regions (IDRs) that flank the folded domains. By combining single-molecule experiments and Monte Carlo simulations, we previously showed that N protein adopts a complex and dynamic conformational ensemble as a result of its disordered regions[441]. While many experiments have focused on the interaction of the two folded regions (RBD and dimerization domain) with RNA, little is known about the role played by the three disordered regions in aiding the capture and organization of the nucleic acid. The so-called fly-casting model[752] suggests that

IDRs have a larger capture radius compared to rigid proteins, resulting in an amplified recruitment of ligands. At the same time, recent experiments have pointed out the peculiarity of disordered regions in encoding for and modulating binding affinity, showing that complexes of oppositely charged biopolymers may achieve high affinity and retain fast dynamic ensembles[76].

Here, we focused our investigation on the RNA Binding Domain (RBD) of the SARS-CoV-2 N protein and studied its interaction with nucleic acids, in the presence and absence of the disordered N-Terminal Domain (NTD). We restricted our analysis to the RBD and the contiguous NTD (**Fig. 1** and **Supplementary Tables 1 and 2**) to identify the specific contributions of the IDR to the folded domain, which otherwise would be masked or altered by the effect of other domains. We hypothesized that the NTD plays an important role since it contributes to localization of the N protein into stress granules[392,469,753] in a RNA dose-dependent manner[469], suggesting that localization is also mediated by its interaction with nucleic acid.

Single-molecule Förster Resonance Energy Transfer (FRET)[754–756] provides an effective method to determine the affinity and stoichiometry of the binding of RNAs to both RBD and NTD-RBD, while monitoring conformational and dynamic changes occurring in the NTD within the same set of experiments. Single-molecule detection simplifies identification of the contact site size and affinity of the protein even for long nucleic acids since all protein:RNA complexes contain only one single protein (as monitored by Pulsed Interleaved Excitation[216]), whereas in typical ensemble experiments one has to account for the contribution of different protein:nucleic acid stoichiometries to the overall signal.

We examined RNA binding using both "nonspecific" and specific RNA molecules. *In cell* crosslinking experiments found that N protein is bound to mRNAs sites containing multiple rU's[757], while others found it dispersed over the viral genome, comprising both single-stranded and double-stranded regions[750,758]. Given the lack of consensus in the literature, we have opted for "nonspecific" poly(rU)$_n$ sequences that are well-behaved polyelectrolytes and, differently from poly(rA) and poly(rG), do not undergo stacking at high nucleic acid concentrations. As specific sequences, we have focused on a single-stranded RNA (ssRNA) element of 21 nucleotides that has been isolated from the 5' UTR of the viral genome (which we will refer to as V21) and on hairpins from the 5' UTR (SL5B) and a putative packaging signal NSP15[666] (**Supplementary Table 3**).

6.3 Material And Methods

Protein expression and purification

GST-His9-SARS-CoV2 NTD-RBDL and NTDL-RBD Nucleocapsid constructs were expressed recombinantly in Gold BL21(DE3) cells (Agilent). 4 L cultures were grown in LB medium with carbenicillin (100 ug/mL) to OD600 ~0.8 and induced with 0.25 mM IPTG for 3 hours at 37 ºC. Harvested cells were lysed with sonication at 4 ºC in lysis buffer (50 mM Tris pH 8, 300 mM NaCl, 10% glycerol, 10 mg/mL lysozyme, 5 mM βME, cOmplete™ EDTA-free Protease Inhibitor Cocktail (Roche), DNAse I (NEB), RNAse H (NEB)). The supernatant was cleared by centrifugation (140,000 x g for 1 hr) and bound to a HisTrap FF column (GE Healthcare) in buffer A (50 mM Tris pH 8, 300 mM NaCl, 10% glycerol, 20 mM imidazole, 5 mM βME). The column was then washed with High Salt Buffer (50 mM Tris pH 8, 2M NaCl, 10% glycerol, 5 mM βME) for ten column volumes followed by ten column volumes of Buffer A. GST-His9 -N protein fusion was eluted with buffer B (buffer A + 500 mM imidazole) and dialyzed into cleavage buffer (50 mM Tris

pH 8, 50 mM NaCl, 10% glycerol, 1 mM DTT) with HRV 3C protease, thus cleaving the GST-His9 -N fusion yielding N protein with two additional N-term residues (GlyPro). N protein was then bound to an SP sepharose FF column (GE Healthcare) and eluted using a gradient of 0-100% buffer B (buffer A: 50 mM Tris pH 8, 50 mM NaCl, 10% glycerol, 5 mM βME, buffer B: buffer A + 1 M NaCl) over 100 min. Purified NTD-RBDL and NTDL-RBD constructs were analyzed using SDS-PAGE and their concentrations were determined spectroscopically in 50 mM Tris (pH 8.0), 500 mM NaCl, 10% (v/v) glycerol using an extinction coefficient of 25200 M-1 cm-1 at 280 nm.

GST-His9-SARS-CoV2 RBD_L Nucleocapsid construct was expressed recombinantly in Gold BL21(DE3) cells (Agilent). 4 L cultures were grown in LB medium with carbenicillin (100 ug/mL) to OD600 ~ 0.6 and induced with 0.3 mM IPTG for 3 hours at 37 ºC. Harvested cells were lysed with sonication at 4 ºC in lysis buffer (50 mM Tris pH 7, 300 mM NaCl, 10% glycerol, 10 mg/mL lysozyme, 5 mM βME, cOmplete™ EDTA-free Protease Inhibitor Cocktail (Roche), DNAse I (NEB), RNAse H (NEB)). The supernatant was cleared by centrifugation (140,000 x g for 1 hr) and bound to a HisTrap FF column (GE Healthcare) in buffer A (50 mM Tris pH 7, 300 mM NaCl, 10% glycerol, 20 mM imidazole, 5 mM βME).The column was then washed with High Salt Buffer (50 mM Tris pH 7, 2M NaCl, 10% glycerol, 5 mM βME) for ten column volumes followed by ten column volumes of Buffer A. GST-His9 -N protein fusion was eluted with buffer B (buffer A + 500 mM imidazole) and dialyzed into cleavage buffer (20 mM Tris pH 7, 20 mM NaCl, 10% glycerol, 1 mM DTT) with HRV 3C protease, thus cleaving the GST-His9 -N fusion yielding N protein with two additional N-term residues (GlyPro). The N protein was then run over a HisTrap FF column (GE Healthcare) in Buffer A (20 mM Tris pH 7, 20 mM NaCl, 10% glycerol) and the flow through was collected. N protein was then bound to an SP sepharose FF column (GE Healthcare) and eluted

using a gradient of 0-100% buffer B (buffer A: 20 mM Tris pH 7, 50 mM NaCl, 10% glycerol, 5 mM βME, buffer B: buffer A + 1 M NaCl) over 100 min. Purified RBD_L construct was analyzed using SDS-PAGE and its concentration was determined spectroscopically in 50 mM Tris (pH 7.0), 300 mM NaCl, 10% (v/v) glycerol using an extinction coefficient of 25200 M^{-1} cm^{-1} at 280 nm. Plasmid DNA sequences for the constructs can be found in **Supplementary Information**.

Protein labeling

All Nucleocapsid variants were labeled with Alexa Fluor 488 maleimide (Molecular Probes, USA) under denaturing conditions in buffer A (10 mM Tris pH 7.3, 6 M Urea) at a dye/protein molar ratio of 0.7/1 for 2 hrs at room temperature. Single labeled protein was isolated via ion-exchange chromatography (Mono S 5/50 GL, GE Healthcare - protein bound in buffer A (+5 mM βME) and eluted with 0-40% buffer B (buffer A + 1 M NaCl) gradient over 70 min) and UV-Vis spectroscopic analysis to identify fractions with 1:1 dye:protein labeling. Single donor labeled N protein was then subsequently labeled with Alexa Fluor 594 maleimide at a dye/protein molar ratio of 1.3/1 for 2 hrs at room temperature. Double-labeled (488:594) protein was then further purified via ion-exchange chromatography (Mono S 5/50 GL, GE Healthcare).

RNA Preparation

Single-stranded RNAs were purchased from IDT (USA) and Horizon Discovery (USA). Hairpin RNAs were transcribed with T7 RNA polymerase from DNA oligonucleotides (IDT), using T7 RNA polymerase (NEB USA) in an optimized reaction mix. RNAs were purified by denaturing polyacrylamide gel electrophoresis (15% acrylamide, 19:1 bis, 8 M urea, Tris-Borate-EDTA), bands were visualized by UV shadowing and cut out. Gel slices were soaked in 0.3 M sodium acetate

overnight at 30 ºC in a rotating mixer, the solution was recovered and gel debris removed by centrifugation. RNA was precipitated overnight at -20 ºC in the presence of glycogen with 3X volume 100% ethanol, and the pellet resuspended in Milli-Q water (Millipore-Sigma, USA). Hairpins were annealed in 10 mM HEPES pH 6.5, 50 mM KCl buffer and their integrity and stability measured in UV melting experiments as a function of their concentration. RNA concentrations were determined spectrophotometrically employing their computed extinction coefficients at 260 nm.

Instrumentation

Single-molecule experiments were performed on a modified Picoquant MT200 instrument (Picoquant, Germany) using Pulsed Interleaved Excitation to enable identification of the donor- and acceptor-only as well as donor-acceptor populations. All data reported in this work are selected for the donor-acceptor population. Single-molecule measurements, unless otherwise stated, have been performed in 50 mM Tris, pH 7.4 at room temperature (23 ± 1 °C).

Analysis of binding experiments.

Binding of RNA ligands to labeled N protein constructs was monitored by following either the mean value of the transfer efficiency distribution or the fraction of bursts associated with the bound and unbound population (when they can be resolved).

In the first case, titration curves were analyzed according to:

$$\underline{E} - \underline{E_f} = (\underline{E_b} - \underline{E_f}) \frac{K_A[RNA]_{tot}}{1+K_A[RNA]_{tot}} \quad \text{(Eq. 1)}$$

where $\underline{E_f}$ and $\underline{E_b}$ are the mean transfer efficiencies for the free and bound protein, K_A is the association constant and [RNA] is the total concentration of RNA. Note that under all conditions the free RNA concentration is always much higher than the concentration of a bound complex because of the single-molecule concentrations used in the experiment.

In the second case, when the fraction of bound protein f_b is directly estimated, titration curves were analyzed according to:

$$f_b = \frac{K_A[RNA]_{tot}}{1+ K_A[RNA]_{tot}} \quad \text{(Eq. 2a)}$$

for the 1:1 binding cases, and:

$$f_{b1} = \frac{K_{A1}[RNA]_{tot}}{1+ K_{A1}[RNA]_{tot}+K_{A1}K_{A2}[RNA]^2_{tot}} \quad \text{(Eq. 2b)}$$

$$f_{b2} = \frac{K_{A1}K_{A2}[RNA]^2_{tot}}{1+ K_{A1}[RNA]_{tot}+K_{A1}K_{A2}[RNA]^2_{tot}} \quad \text{(Eq. 2c)}$$

for the 1:2 case treated in this work.

For the special case of the binding to the polynucleotide poly(rU), titration curves were obtained and analyzed as a function of the total concentration of nucleotide residues [poly(rU)], not RNA molecules. This is justifiable because under the experimental conditions employed, where the protein concentration is so much lower than RNA concentration, the McGhee-von Hippel formulation for the binding of large ligands to one dimensional lattices reduces to:

$$f_b = \frac{K_{int}[poly(rU)]}{1+K_{int}[poly(rU)]} \quad \text{(Eq. 3)}$$

where K_{int} is the intrinsic association constant.

Statistical Analysis

Values associated with multiple measurements are presented as mean and standard deviation of the measured points of at least two points. Results of model fit to the data are presented as best value and corresponding error of the fit as determined using non-linear regression algorithms in Mathematica (Wolfram Research Inc, USA).

Data Availability, Software, Algorithms

Data analysis of single-molecule data has been performed using the Fretica package for Mathematica (Wolfram Research Inc, USA) developed by the Schuler group (https://schuler.bioc.uzh.ch/wp-content/uploads/2022/07/Fretica20220630.zip). All single-molecule data reported in this work are deposited at https://github.com/holehouse-lab/supportingdata/tree/master/2023/cubuk_2023 Raw photon traces of single-molecule data will be made available upon request.

Simulations

Coarse-grained molecular dynamics (MD) simulations were performed in the NVT ensemble using the LAMMPS simulation engine with Mpipi model using the default parameters developed by Joseph et al.[759]. Mpipi is a one-bead-per residue coarse grained force field developed specifically for working with intrinsically disordered proteins. Non-bonded interactions are driven by a short-range potential and, where applicable, a long-range Coulombic potential. Bonded interactions are encoded via a simple harmonic potential. Simulations were performed with NTD-RBD, RBD, and NTD, with and without $(rU)_n$ of lengths n = 10, 12, 15, 17, 20, 25, 30, 35, 40 and 180 nucleotides. We also performed simulations of the Omicron variant of the NTD-RBD, with substitution of the proline

residue in position 13 with a leucine (P13L) and deletion of residues from 31 to 33 ($\Delta^{31\text{-}33}$). For assessing the role of each site, also we performed simulations of the P13L mutation alone versus the $\Delta^{31\text{-}33}$ alone. All simulations of the Omicron constructs were performed with and without$(rU)_{25}$.

All simulations were run with multiple independent repeats using a 30 nm^3 simulation box and periodic boundary conditions. As in previous work, folded domains were modeled as rigid bodies, whereas intrinsically disordered regions and ssRNA were described as flexible polymers[759,760]. For simulations where folded domains were present (i.e. those with the RBD), six distinct RBD conformations were taken from all-atom simulations of the RBD performed using the Folding@Home distributed computing platform[441,525,761]. This enables us to ensure conclusions obtained are not dependent on a specific RBD conformation. For the six independent starting configurations, five repeats were performed, with 300 million steps per repeat, such that 30 independent simulations were run for each unique protein/RNA combination. Simulation configuration data was recorded every 100,000 steps, and the first 600,000 steps (0.2% of the simulation) discarded as equilibration. Across the 30 independent simulations for each protein/RNA combination we generated approximately 270,000 frames. A summary of the simulations performed is provided in Supplementary Table 4.

Simulations were analyzed using SOURSOP (https://soursop.readthedocs.io/) and MDTraj [715]. All analysis code for simulations is provided at https://github.com/holehouse-lab/supportingdata/tree/master/2023/cubuk_2023. Simulation trajectory data is available at DOI:10.5281/zenodo.7631327. For more details on the simulations see extended materials and methods in the Supplementary Information.

Database Referencing

Sequence data for the Nucleocapsid variants, including the Omicron variant, were obtained from the GISAID lineage-comparison database: https://gisaid.org/lineage-comparison/

Extended description of experimental procedures, material and methods, and data analysis are presented in Supplementary Information.

6.4 Results

In order to investigate the binding and conformational changes of the N-terminal disordered tail and RNA-binding domain of SARS-CoV-2 Nucleocapsid protein *via* single-molecule FRET, we created two truncated constructs, one spanning the full N-terminal segment of the protein comprising both the NTD and RBD and another comprising the RBD alone (**Fig. 1**). Cysteine mutations were introduced in the wild-type sequence to enable fluorophore addition to the constructs *via* maleimide-thiol chemistry. Specifically, we introduced cysteine mutations in the RBD sequence in positions 68 and 172 of the NTD-RBD and RBD constructs to monitor conformations of the RBD. In contrast, we introduced cysteine residues in positions 1 and 68 of the NTD-RBD construct to monitor conformations of the NTD (**Fig. 1**). We will refer to these constructs as RBD_L, NTD-RBD_L, and NTD_L-RBD respectively, where the L subscript identifies the region probed by the labels. All constructs have been expressed in *E.coli*, purified, and labeled with Alexa Fluor 488 and Alexa Fluor 594.

6.5 Folding stability of RBD.

As a preliminary step, we tested whether truncation of the NTD impacts the conformations adopted by the RBD and its folding stability, since this would alter the ability of the domain to interact with nucleic acids. Our previous single-molecule experiments[441] showed that the RBD is equally stable when it is part of the full-length protein or of the isolated NTD-RBD construct, suggesting that the linker region does not impact its folding stability. Following this earlier work, we next directly measure the stability of the RBD in the absence of the NTD.

Single-molecule FRET measurements of the RBD construct show a single peak with high transfer efficiency (**Fig. 2**) that is compatible with previous observations of the completely folded RBD in the context of the NTD-RBD and full-length protein[441]. To confirm the observation, we further quantified the folding stability of the RBD in the absence of the NTD by titrating Guanidinium Chloride (GdmCl) into the RBD_L construct. Increasing the concentration of denaturant revealed the appearance of up to two species, which mirrors previous observations of an intermediate and unfolded state identified for the same domain[441]. An estimate of the relative abundance of each species can be computed by comparing the relative areas of the distinct populations. The data can be well described assuming a thermodynamic equilibrium between three states with ΔG_{UI} = 2.8 ± 0.1 kcal mol^{-1} and $c_{UI,1/2}$ = 1.26 ± 0.03 M and ΔG_{IF} = 7.6 ± 0.4 kcal mol^{-1} and $c_{IF,1/2}$ = 1.21 ± 0.01 M (**Fig. 2 and Supplementary Information**). Overall, our observations confirm that RBD is completely folded under aqueous buffer conditions. Compared to the full-length protein, truncation of the tail slightly shifts the unfolding transition towards lower GdmCl concentrations, but does not significantly affect the fraction folded in the absence of denaturant (**Supplementary Table 5**).

6.6 Binding of nonspecific RNA to RBD.

Given our goal is to quantify and compare the binding affinity of the RBD for RNA, we sought to develop a single-molecule assay that would let us quantify the fraction of bound protein as a function of RNA concentration. We first tested whether binding of RNA to RBD can be visualized via changes in transfer efficiency. With increasing concentration of a ~200 nucleotide long poly(rU), we noticed a small but measurable shift toward higher values of transfer efficiencies, from a mean transfer efficiency of ~ 0.87 to ~ 0.90. (**Fig. 3**)

A plot of the deviation in mean transfer efficiency as a function of nucleic acid concentration reveals a sigmoidal trend that saturates at high concentration, as expected for a binding isotherm of the RNA to RBD on a logarithmic scale. We note that in typical ensemble experiments, a 1:1 protein:nucleic acid binding stoichiometry cannot be automatically assumed when titrating a long nucleic acid with multiple binding sites against protein. However, here the 1:1 binding stoichiometry can be invoked because of the single-molecule nature of the experiments, where only labeled proteins are present in the solution and only one labeled protein per time is observed in the confocal volume. This is confirmed by Pulsed Interleaved Excitation, which provides a quantification of the labeling stoichiometry of the measured molecules and supports that the protein remains "monomeric" across the whole titration. This does not exclude the possibility of two unlabeled nucleic acids binding to the protein, though we would expect a change in the concentration-response (see for comparison binding of NTD-RBD_L to specific single-stranded RNA). A fit of the mean transfer efficiencies across the titration to the 1:1 binding model reveals an intrinsic association constant K_{int} of (6 ± 2) x 10^{-2} μM^{-1} (**Fig. 3, Supplementary Table 6**) at the standard buffer conditions of 50 mM Tris, pH 7.4.

To further test whether the signal does indeed report on binding, we investigated the effect of nucleic acid length on the detected binding affinity. A decrease in the length of the nucleic acid is expected to result in weaker binding affinities because of the reduction in productive binding configurations for short oligonucleotides. When repeating the same titration, for $(rU)_n$ oligonucleotides with length n = 10, 12, 15, 17, 20, 25, 30, and 40 nucleotides, we observe an analogous response of the transfer efficiency distribution, with the mean transfer efficiency increasing with increasing RNA concentration (**Fig. 4** and **Supplementary Fig. 1**). As for poly(rU), each titration curve can be well described by a 1:1 binding model and the corresponding equilibrium binding constants can be estimated. When plotted against the length of the oligonucleotide, a clear increase in the association constant K_A (per molecule) is observed with increasing length of the RNA, ranging from $(4 \pm 3) \times 10^{-2}$ μM^{-1} to (1.2 ± 0.3) μM^{-1} (**Supplementary Table 7**).

Assuming a simple unidimensional lattice model with an intrinsic association constant K_{int}, a given length of the nucleic acid n, and a contact site size of M nucleotides (the number of contiguous nucleotides involved in the interaction when a "complete" contact is realized with protein), we expect a linear trend as a function of n extrapolating through the x-axis (the length of the nucleic acid) at $(M-1)$, i.e.

$$K_a = K_{int}(n - M + 1) \qquad \textbf{(Eq. 4)}$$

Indeed, measured association constants follow a linear trend and fit to **Eq. 4** results in an intrinsic association constant $K_{int} = (4.5 \pm 0.5) \times 10^{-2}$ μM^{-1} and a contact site size $M = 12 \pm 2$. The model can be further developed to incorporate the contribution of partial interactions of the protein with the nucleic acid and include overhang effects, which in a first approximation can be described by:

$$K_A = K_{int,M}(n - M + 1) + 2\sum_{j=1}^{M-1} \quad K_{int,j} \quad for\ M < n \qquad \textbf{(Eq. 5a)}$$

$$K_A = K_{int}(M - n + 1) + 2\sum_{j=1}^{M-1} \quad K_{int,j} \quad for\ M \geq n \qquad \textbf{(Eq. 5b)}$$

where $K_{int,j}$ represents a modified K_{int} to account for the overhang effects (**Supplementary Information**).

The equation provides a quantitative representation of the complete dataset and identifies a K_{int} = (5.2 ± 0.4) x 10^{-2} μM^{-1} and a contact site size M = 23 ± 2. Note that K_{int} is within error of the value determined with **Eq. 4** and is consistent with the corresponding intrinsic association constant measured with the ~200 nucleotide-long poly(rU). However, introducing partial binding at the ends of the chain leads to an increase in the estimate of the site size. This is a reflection of a strong assumption in the model, i.e. that the same average interaction is realized through all amino acids and nucleotides across the contact site (**Supplementary Information**). This obviously is an oversimplification that does not account for the contribution of ion release to the association constant as well as sequence-specific effects of the contact site. Therefore, the absolute value of the contact site size is likely to be overestimated by the fit to **Eq. 5.** The value falls between the estimates obtained with **Eq. 4** and **Eq. 5**. Having estimated the association constant and contact site size for the RBD, we then proceeded to investigate how the addition of the NTD alters these interaction parameters.

6.7 Binding of nonspecific RNA to NTD-RBD.

To test whether the addition of the disordered tail leads to a change in the binding affinity, we measured the association of the same poly(rU) using the construct NTD-RBD_L. Titration of the RNA reveals a shift in the mean transfer efficiency that is analogous to the one observed for the

RBD_L, but the transition associated with binding is now shifted to low nanomolar concentrations. Fit of the mean transfer efficiency with a 1:1 binding model reveals a $K_{int} = (2.0 \pm 0.4)\ \mu M^{-1}$.

To confirm that this effect is due to the disordered tail, we turn to a second construct, the NTD_L-RBD with labels in positions 1 and 68, which has been shown previously to report on the configurations of the disordered N-terminal tail and is in good agreement with the results from atomistic Monte Carlo simulations[441]. In the absence of RNA, this NTD_L-RBD construct in aqueous buffer conditions reports on one narrow distribution that reflects the fast averaging over the conformational ensemble of the disordered tail. We proceed by testing if the same construct can report on RNA binding. With increasing concentration of poly(rU), we observe a modulation of the transfer efficiency distribution with a shift toward lower transfer efficiencies, from a mean transfer efficiency $E = 0.709 \pm 0.009$ in absence of RNA to $E = 0.542 \pm 0.003$ in presence of 10 μM of poly(rU) (**Fig. 3**). This observation clearly supports that the disordered tail is directly affected by the binding of RNA.

Analogous to the case of NTD-RBD_L and RBD_L, an estimate of the binding affinity can be obtained by plotting the mean transfer efficiency (as fitted by a Gaussian distribution) as a function of the RNA concentration. Such analysis can be interpreted in terms of a simple 1:1 binding model, resulting in a $K_{int} = (3.7 \pm 0.4)\ \mu M^{-1}$. By a careful inspection of the width of the distribution, a broadening is observed for intermediate concentrations of RNA, suggesting that the measured distribution is indeed the resulting average of an unbound and bound population. Under this assumption, data can be refitted using two Gaussian distributions and the corresponding areas can be used to infer the fraction bound and unbound (**Fig. 3**). These quantities can be further analyzed

to extract binding affinity for the nucleic acid, K_{int} = (4.0 ± 0.3) μM^{-1} , which is in very good agreement with the one obtained from the mean value of the distribution. Both estimates of intrinsic association constants for the NTD_L-RBD constructs are in close agreement with the one obtained for NTD-RBD_L, confirming both constructs report on the same RNA binding independent of the labeling position. Based on these observations, the affinity of the NTD-RBD constructs appears to be ~40-80 times tighter than that of the RBD alone, pointing to a direct contribution of the disordered region in favoring RNA binding.

Since the tail unequivocally favors binding, the conformations of NTD_L-RBD upon RNA binding represent direct interactions of the tail with RNA. This poses a further question of whether the conformational change of the NTD represents a specific structural rearrangement due to an intrinsic encoded bound conformation or whether the conformational change reflects a dynamic conformational ensemble for the NTD-RBD/RNA complex. In the first case scenario, we expect that altering the length of the homo-polynucleotide sequence would possibly result in a change of affinity, but would not alter the mean transfer efficiency. In the second case scenario, instead, we expect to observe a change in both affinity and mean transfer efficiency.

To test this hypothesis, we investigated the binding of $(rU)_n$ oligonucleotides with *n* ranging from 10 to 40 nucleotides (**Fig. 4**). For all of the sequences we observe a continuous shift in the mean transfer efficiency, reflecting binding of the RNA. Significantly, the mean transfer efficiency corresponding to the bound state depends on the length of the nucleic acid. The dependence of the mean transfer efficiency with the length of nucleic acid suggests a saturation effect that is reached for sufficiently long RNA. Inspection of the binding equilibrium constant as a function of length reveals two distinct regimes, which - as a first approximation - can be described by using **Eq. 4** and

2. A linear fit using **Eq. 4** for RNAs with length between 20 and 40 nucleotides results in a K_{int} = (4.2 ± 0.4) μM^{-1} and M = 21 ± 1 nucleotides. A complete fit of the dataset using **Eq. 5** results in an intrinsic association constant K_{int} = (4.3 ± 0.2) μM^{-1} and M = 25 ± 2 nucleotides. The change in slope at approximately 20 nucleotides indicates that this length of nucleic acid is required to satisfy all the contacts between the nucleic acid and the NTD_L-RBD construct, which results in a larger contact site size. In addition to a larger contact size, the interaction per nucleotide is tighter than the one determined for the RBD alone, as indicated by the NTD-RBD K_{int}. Interestingly, a shift in transfer efficiency is observed for lengths shorter than the contact site size of RBD, implying that even for short oligos not all the contacts occur within the folded domain, and interactions with the tail need to be formed.

Taken together with the tighter K_{int} observed for NTD-RBD, these observations indicate that the complex between RNA and NTD-RBD is not solely initiated by contacts with the RBD domain but instead relies on dynamic interactions between the RNA and both RBD and NTD. Furthermore, the transfer efficiency shift does not saturate at the contact site size of the NTD-RBD construct (20 nucleotides); instead, a continuous change is observed for longer lengths, approaching saturation at approximately 40 nucleotides. These observations further suggest a dynamic complex between the protein and RNA, where the position of the contacts formed depends on the number of available nucleotides and the contact site size represents a mean number of minimum contacts that are formed above a given length of the oligo.

To test this hypothesis, we performed ns-FCS measurements of the NTD_L-RBD in the presence of RNA. We previously showed that the NTD region in absence of RNA is flexible and dynamic[441]. ns-

FCS measurements of the NTD_L-RBD in the absence of RNA reveals a reconfiguration time of approximately 110 ± 20 ns, which is marginally affected upon binding RNA, with a reconfiguration time of the NTD spanning a range between 94 and 108 ns across the different lengths tested from $(rU)_{10}$ to $(rU)_{40}$ (**Supplementary Fig. 2**). This indicates that the NTD remains largely dynamic and contacts must occur only across a small set of nucleotides.

6.8 Simulations of RNA binding to NTD-RBD.

To gain a molecular understanding of the interaction between RNA and the NTD-RBD, we turned to coarse-grained molecular dynamics simulations. We utilized the Mpipi force field, a recently-developed model that combines short-range interactions and long-range electrostatics and encodes each amino acid or nucleotide as a chemically-distinct entity (**Fig. 5A**) [759]. Mpipi was specifically developed with intrinsically disordered regions in mind[759]. Previous work has shown good agreement between simulations and experiments when this model has been used to assess non-specific protein-protein and protein-RNA interactions leading to phase separation [759,762,763].

We first simulated RBD with $(rU)_{10}$ to identify residues on the folded domain that contribute to ssRNA binding (**Fig. 5B**). We calculated protein:RNA contacts from these simulations and observed reasonable agreement with previously-reported NMR chemical shift perturbation experiments of the RBD with ssRNA, performed with a 10-mer RNA of 5'-UCUCUAAACG-3'[525]. This result suggests that our simulations, at least qualitatively, are able to recapitulate experimentally measured protein:RNA interactions (**Fig. 5B**).

Having first performed simulations of $(rU)_{10}$ with the RBD, we next performed simulations of NTD-RBD and $(rU)_{10}$. In addition to the previously observed RBD interactions with $(rU)_{10}$, we now observed additional interactions between the disordered NTD and $(rU)_{10}$ (**Fig. 5B, Supplementary Fig. 3**). The NTD remains fully disordered in the bound state of NTD-RBD:$(rU)_{10}$ (**Supplementary Fig. 4**) and the pattern of RBD – $(rU)_{10}$ interactions is comparable in both the presence and absence of the disordered NTD. While the same RBD residues engage with RNA in the presence vs. absence of the NTD, the frequency is altered. Specifically, the NTD enhances interactions between residues 89 – 107 of the RBD with RNA (**Supplementary Fig.3**). This region maps to the β-extension previously identified as engaging in RNA interactions [525]. Within the NTD, residues 30-50 contain five positively charged amino acids (four arginines and one lysine) and interact directly with $(rU)_{10}$, in good agreement with recently published NMR experiments[527] (**Fig. 5B**). Taken together, these results suggest that the presence of the NTD potentiates RBD:RNA interactions as well as engaging in a new set of interactions with RNA.

We then tested whether our simulations capture the enhanced affinity of NTD-RBD with RBD and the length dependence of the binding model. By defining the fraction of the simulation in which the protein and RNA are bound to one another, we can calculate an apparent binding association constant (K_A) for simulations with either RBD or NTD-RBD and compare the relative values (see **Supplementary Tables 8-9** and **Supplementary Fig. 5**). Comparing the binding of these two constructs to $(rU)_{25}$ (which is larger than the measured contact site size of RBD and equivalent to the upper limit of the one of NTD-RBD), the presence of the NTD increases the K_A by a factor of 4.7 ± 0.4 %, in good agreement with our experimentally measured ratio of association constants of 3 ± 1 % for $K_{A,RBD}/K_{A,NTD\text{-}RBD}$ (**Fig. 5D, Supplementary Table 10**). Intriguingly, simulations of NTD

alone with $(rU)_{25}$ revealed substantially weaker binding compared to either the RBD or NTD-RBD (**Fig. 5D-G**). This suggests that the NTD's ability to enhance RNA binding – at least in the context of poly(rU) – is an emergent property of the NTDs location relative to the RBD, as opposed to solely an intrinsic ability to bind RNA tightly.

Our single-molecule FRET experiments revealed an expansion of the NTD upon binding to RNA, where longer single-stranded RNAs lead to a higher degree of NTD expansion (**Fig. 4**). This is in contrast to simple expectations for polyelectrolyte condensation, where oppositely charged polymers are expected to compact upon interaction with one another[324,764]. This RNA-dependent expansion of the NTD is reproduced in our simulations, where we observed an increase in the root mean square distance (RMSD) between residues 1 and 68 of the NTD upon RNA binding, followed by a modest increase in RMSD as the RNA length increases up to $(rU)_{20}$ (**Fig. 5C**). These trends are in qualitative agreement with the single-molecule FRET measurements (**Fig. 4F**). These results confirm our ability to capture the conformational behavior of the NTD upon RNA binding, while adding further evidence of RNA length dependent expansion of the NTD.

Importantly, in all the simulations the bound state is a dynamic complex that is compatible with the dynamics observed in nsFCS experiments (**Fig. 5A, Supplementary Movie 1**). Taken together, our results suggest that NTD-RBD interacts with RNA forming a disordered "fuzzy" complex largely driven by the interaction with positively charged groups in the NTD and RBD.

6.9 Effect of salt.

Protein-RNA interactions are known to be sensitive to salt concentrations due to the large contribution of electrostatics. A significant contribution to binding can arise from condensed ions on protein and RNA, which can be released upon binding. To estimate the extent of ion release, we measured the association constant as a function of the salt concentration. We restrict our investigation to $(rU)_{20}$ and $(rU)_{40}$, where we can quantify affinities up to 200 mM KCl in the range of available concentrations of the ligand. As shown in **Supplementary Fig. 6** and **7**, the mean transfer efficiency of the NTD_L-RBD is marginally altered by salt screening in absence of the ligand, which is consistent with previous observations[441].

NTD_L-RBD was titrated with $(rU)_n$ at different KCl concentrations. Representative histograms and the observed dependence of K_A on salt concentration are shown in **Fig. 6** and **Supplementary Fig. 6-7**. Both $(rU)_{20}$ and $(rU)_{40}$ datasets reveal a linear trend on the log-log plot of K_A and K^+ concentration. Analogous results are obtained when considering the total concentration of cations K^+ and $Tris^+$ (**Supplementary Fig. 8**). The lack of curvature in K^+ titration suggests that interactions with $Tris^+$ ions do not contribute substantially to ion release. The slope of the linear trend is equal to -5.1 ± 0.4 and -5.0 ± 0.5 for $(rU)_{20}$ and $(rU)_{40}$, respectively, indicating a net release of ~ 5 ions upon interaction[765] (see **Supplementary Table 11**). Finally, our measurements also provide a quantification of the RNA binding association constants at the physiological concentrations found in cells (~150 mM K^+). When compared to corresponding values observed in the reference buffer condition, we observe an decrease of the association constant K_A to (0.17 ± 0.02) μM^{-1} for $(rU)_{20}$ and

(0.38 ± 0.04) μM^{-1} for $(rU)_{40}$, corresponding to a weaker affinity in higher salt concentration (see **Supplementary Table 12**).

6.10 Interaction with specific single-stranded RNA.

To test whether sequence specificity can affect affinity and mode of binding of the specific RNA with the disordered region, we studied the interactions with a 21 nucleotide sequence (V21) from the 5' UTR of the viral genome. This region of the genome was previously found interacting with the N protein in *in cell* crosslinking studies [442] and has been confirmed to adopt no secondary structure at room temperature[766].

We quantified binding of V21 using the NTD_L-RBD construct. As for the case of nonspecific single-stranded RNA, at increasing concentration of V21, we notice a shift of the mean transfer efficiency that reaches a saturating value at ~ 1 μM RNA concentration, which we interpret as representing the binding between one protein and one RNA strand. However, at concentrations of V21 higher than 1 μM, we observe the appearance of a second population at lower transfer efficiency, which is consistent with a second binding event of the nucleic acid to the protein, i.e. a 2:1 RNA:protein stoichiometry. This conformational change is associated with a mean transfer efficiency that is significantly lower than any of the mean transfer efficiencies that has been observed for poly(rU) (E ~ 0.37), indicating a distinct mode of binding and structural organization of the NTD. We interpret such an extended configuration as an expansion of the tail to accommodate two nucleic acid molecules. Since we observe this second mode of binding only for V21 but for none of the poly(rU) sequences, we propose that this second bound state is the result of a partial hybridization of the V21 sequence.

To quantify the association constants corresponding to the different binding events, we globally fit the change in the mean transfer efficiency associated with the first binding event and the change in relative area of the second population associated with the second binding event (**Fig. 7, Supplementary Table 13**). Data are globally fit to a model that accounts for two distinct bound states with corresponding association constants K_{A1}^{V21} of (6.2 ± 0.3) μM^{-1} and K_{A2}^{V21} of (0.15 ± 0.10) μM^{-1}. K_{A1}^{V21} is ~ 50% larger than the corresponding association constant for r(U)$_{20}$, K_{A1}^{rU20} = (4.3 ± 0.3) μM^{-1}, whereas the mean transfer efficiency of the bound state appears only slightly smaller than that for r(U)$_{20}$. To better understand if the second mode of binding is compatible with double-stranded sequences, we turned to the investigation of specific double-stranded RNA sequences.

6.11 Interaction with specific RNA hairpins.

The 5' UTR of the SARS-CoV2 genome contains short single-stranded regions and various conserved hairpins, which can offer additional binding sites to the NTD-RBD. In addition, double-stranded regions of the genomic RNA have been proposed as putative packaging signals[666], including the SL5B hairpin in the 5' UTR and the NSP15 hairpin from the mRNA of the Nonstructural Protein 15[632,666,767] (see **Fig. 8, Supplementary Fig. 9, Supplementary Table 14**). Given the potential role of these regions in driving condensation of the nucleic acid, we focused on these two archetypal sequences. NSP15 and SL5B were transcribed *in vitro*, and their hairpin structure at room temperature was confirmed by thermal melting experiments (**Supplementary Fig. 10**).

Single-molecule FRET measurements of the NTD$_L$-RBD construct bound to either SL5B or NSP15 reveal a clear shift of the transfer efficiency distribution toward lower values, i.e. more extended

configurations. Deviation of mean transfer efficiency can be fit as in the case of single-stranded RNA to determine the association constants: K_A^{NSP15} = (7.8 ± 0.7) x 10^{-1} μM^{-1} and K_A^{SL5B} = (5.3 ± 0.4) x 10^{-1} μM^{-1}. These values are compatible with the one associated with the second binding mode of V21, K_{A2}^{V21}, supporting the hypothesis that this binding mode is due to hybridization of a double-stranded RNA. Interestingly, the conformational changes of NTD_L-RBD bound to the hairpins appear to be larger than what is observed for the majority of single-stranded RNA, even if the binding affinity is weaker. We attribute the increased expansions of the disordered tail to the larger excluded volume of the double-stranded hairpin.

Finally, we turned to investigate which regions of the hairpins may contribute to the binding. Due to the similar affinity of these sequences to that of $(rU)_{10}$, we hypothesized that NTD-RBD may preferentially bind to the RNA hairpin through its loop region. We chose the NSP15 sequence as a reference and designed RNA hairpins (hpRNA) with perfect duplex stems and loops of either 4 or 10 nucleotides (**Fig. 7**). We refer to these constructs as TetraLoop and DecaLoop. The four nucleotide loop in the TetraLoop is cUUCGg, and is expected to result in a unique and stable structure, while the ten nucleotide loop contains seven U's and is unlikely to form internal structure. We found that the binding affinity of these two hpRNAs does seem to depend on the length of the loop, with a $K_A^{Tetraloop}$ = (6.7 ± 0.8) x 10^{-1} μM^{-1} and a $K_A^{Decaloop}$ = (3.4 ± 0.5) μM^{-1}, suggesting that the single-stranded loop does influence the affinity and, therefore, could be the main site of interaction. However, affinity is stronger than that of $(rU)_{10}$, indicating that binding involves both single- and double-stranded regions of the nucleic acid.

To probe the possible roles of defects in double-stranded regions, we tested whether introducing an unpaired A in the tetraloop hairpin stem would affect binding. We do not find significant differences from the perfect stem ($K_A^{Tetrabulge}$ = (3.4 ± 0.7) x 10^{-1} μM^{-1}), suggesting that small defects in the duplex do not influence the NTD-RBD region. Larger internal loops could act as binding sites, but these would depend on sequence and context.

6.12 Omicron variant.

Many mutations in the N protein occur within the disordered regions[768]. The Omicron variant offers a convenient point of comparison, with three key mutations found in the NTD. More than 90% of sequences on the GISAID database (accessed on February 8 2023) report a proline to leucine substitution in position 13 and deletion of three residues between positions 31 and 33[769] (**Supplementary Table 2**). Residue 13 is part of a predicted short helix motif [441] that may offer an interaction site for RNA binding, whereas residues 31 and 32 contain two oppositely charged residues. To test the impact of these mutations, we expressed, purified, and labeled the Omicron NTD_L-RBD ($^{Om}NTD_L$-RBD).

We first characterized the conformations of the tail in absence of RNA. Given the small variations in the sequence, both in terms of hydrophobicity and net charge, we expect negligible variations. Indeed, we observed no significant shift in transfer efficiency (**Fig. 9**). We then performed binding experiments at increasing concentrations of poly(rU). We observed an identical mean transfer efficiency at saturation concentrations of poly(rU) and K_A = (9 ± 1) 10^{-1} μM^{-1}, approximately 4 times weaker binding affinity than for the wild-type sequence. These observations overall support that the mode of binding of RNA is similar between NTD_L-RBD (Wuhan-Hu-1) and $^{Om}NTD_L$-RBD (as

supported by the same transfer efficiency in the bound state), but with different affinities (as indicated by the concentration dependence).

We further investigate molecular insights by performing corresponding coarse-grained simulations. Here, we observed a decrease in binding affinity between Wuhan-Hu-1 and the Omicron variants. We then tested whether this difference is driven by the lack of the proline substitution or by the charge suppression (**Fig. 9**). Mutating only the proline to leucine in our simulations resulted in no detectable change in the binding affinity. In contrast, maintaining the proline and deleting residues 31 to 33 results in a suppression of binding affinity, suggesting that the change in RNA binding affinity observed for Omicron NTD-RBD is dominated by charge effects **(Supplementary Table 15)**. Overall, our observations indicate that small changes in the sequence composition of NTD may not alter the overall conformational behavior of the chain, but can significantly impact the binding affinity.

6.13 Discussion

The NTD is essential for RBD function.

The N protein is responsible for packaging the SARS-Cov-2 genome, but the molecular mechanism of this process remains underdetermined. While previous work has focused on folded domains of the protein as possible centers for interactions, here we have been exploring the role played by one of the disordered regions to determine if the disordered region is a disposable appendage to the folded domain or plays a role in determining protein function. In particular, we investigated the NTD-RBD region and quantified how the disordered NTD contributes to the mode of binding and affinities for RNA. Through our experiments, we have discovered that the RBD alone binds very

weakly to single-stranded RNAs, while the NTD significantly increases RNA binding affinity. Altogether, our data suggest that the RBD alone cannot be considered a primary determinant of RNA binding, and association is most likely the result of the concerted interaction of the RBD and surrounding disordered regions with RNA.

The NTD-RBD forms a dynamic complex with RNA

Our data confirm the previous observations that the NTD is a flexible and dynamic region[441], whose large degree of conformational heterogeneity is retained when the protein is bound to RNA. Thus in defining the interactions between the NTD and RNA, we cannot model the complex as a rigid body with fixed interactions; rather, we have to consider the points of interaction that can be sampled by the disordered protein and nucleic acid. Inspection of the sequence composition (**Supplementary Table 1**) reveals 7 positive charged residues (6 Arg and 1 Lys) and 2 hydrophobic residues (1 Phe and 1 Trp), which offer possible sites of interaction with the nucleic acid. Indeed, arginines can neutralize phosphate groups on the RNA and aromatic groups of Phe and Trp can stack with RNA bases. From a point of view of the sequence pattern, two Arg and one Phe residues occur in a putative helix (identified in our previous simulations[441]) that span from residue 10 to 16, one Trp and Phe are positioned at the junction between the NTD and RBD, and the remaining Arg and Lys residues are clustered between position 30 and 50.

Our coarse-grained simulations point to a key role of electrostatic interactions in regulating the binding of the nucleic acid to the NTD-RBD region, in particular, the stretch between residues 30 and 50 in the NTD and between residues 85 and 110 in the RBD (**Supplementary Fig. 11 and 12**). These RBD residues comprise the positively charged β-extension, a flexible pair of beta strands that

prior work has identified as wrapping around single-stranded RNA during binding [525]. Previous computational work proposed that the interplay between charged residues on the RBD surface and in the NTD can tune NTD conformational behavior[770]. An additional explanation for these previous observations could be one in which N protein has evolved across coronaviridae to ensure high-affinity RNA binding, with compensatory/co-evolutionary changes in the NTD and RBD ensuring that non-specific electrostatically-driven interactions are conserved in spite of sequence variation in both the NTD and RBD.

Our simulations also allow us to deconvolve the relative contributions of the NTD and RBD to RNA binding, illustrating the benefit of a combined, multi-pronged approach in molecular dissection [544]. Although the addition of the NTD to the RBD leads to a substantial increase in binding affinity, our simulations predict that, in isolation, the NTD binds RNA more weakly than either the RBD or the NTD-RBD. With this in mind, the impact of the NTD appears to be mediated by its position relative to the positively-charged β-extension on the RBD. The resulting orientation offers a dynamic, positively charged binding surface, such that the emergent binding affinity is substantially higher than would be naively expected, likely through both an avidity effect and by prepaying the entropic cost of bringing two positively charged protein regions into relatively close contact with one another.

In addition, the simulations corroborate the experimental intuition of a dynamic complex where not only the protein but also the nucleic acid is exploring heterogeneous conformations in the bound state. Overall, these observations ascribe the NTD-RBD:RNA complex to the category of so-called "fuzzy" complexes. The strong electrostatic nature of the interactions is consistent with the recent

observation of highly dynamic complexes formed by oppositely charged biopolymers[771], as for the case of prothymosin alpha and histone H1[56,772].

The NTD-RBD region prefers single-stranded RNA

Our data clearly support the conclusion that the NTD-RBD exhibits some discrimination among RNA targets. We find a generally higher affinity for both specific and non-specific sequences of single-stranded RNA. This is consistent with previous studies of N protein[442,773], including *in cell* crosslinked studies of the protein to the 5' UTR[442], where single-stranded regions, several large loops and junctions predominated the interactions. Additional studies also identified short U-tracts as possible targets of the interaction. Compared to single-stranded RNA, our work finds lower affinities for double-stranded RNA sequences. In particular, our investigation of model hairpins based on the NSP15 genome region tested the role of RNA duplexes, hairpin loops, and duplex deformations in NTD-RBD association. We found that small deformations in the duplex do not significantly alter the interaction with the protein, whereas an increase in the size of the loop region results in an increase of the binding affinity, confirming a preferential interaction of this protein region with single-stranded RNA.

NTD mutations alter RNA binding.

A high number of mutations occur in disordered regions of the Nucleocapsid protein[768]. Our results on the impact of the Omicron NTD mutations clearly show that alterations of three amino acids in this IDR are sufficient to decrease the interaction affinity between the construct and the nucleic acid. This implies not only that the N protein IDRs play a role in the interaction of the protein with nucleic acids, but that mutations in the same regions can effectively alter the function of the protein.

Moreover, while it is often assumed that small changes in IDRs may not substantially influence molecular function, our results here provide a clear counter-example, whereby a 4-times change in binding affinity is driven by just a few mutations. The sensitivity of RNA binding to small sequence changes that alter the charge of the protein also raises the possibility that phosphorylation may play a role in tuning RNA binding affinity, as has been proposed previously [392,443].

The fact that mutations minimally alter the conformational ensemble, but do alter interaction with the nucleic acid suggests an additional layer of complexity encoded in disordered proteins: on one side, the overall conformations of the protein may impact the capturing radius of the protein, whereas the specificity of residues in the sequence may modulate the binding affinity. This is particularly interesting since the properties of disordered regions can be robust to sequence mutations, as different residues can encode for similar properties of protein conformations, dynamics, and interactions. Indeed, available sequences of the SARS-CoV-2 genome are derived from patients and, therefore, are intrinsically biased to be functionally active (genome must be packaged and virus must be infective). Future studies will be required to understand what type of sequence mutations in IDRs can be tolerated by the virus to maintain the ability of condensing the nucleic acid.

6.14 Conclusions

Overall, our measurements support a model in which the disordered NTD favors binding of the RNA to the RBD by directly participating in the interaction with the ligand and conformations are adapted based on the length of the nucleic acid. The dynamic nature of the complex combined with the preference of single-stranded RNAs may serve as a searching mechanism along the viral genome

for identifying high affinity regions. The ability of the NTD domain to accommodate more than one RNA, possibly harnessing the hybridization of the sequence, may contribute to the packaging of the viral genome.

6.15 Acknowledgements.

We thank Tim Lohman and Roberto Galletto for useful insights and discussions on nucleic acid binding and SARS-CoV-2 Nucleocapsid, Ben Schuler and Daniel Nettels for developing and maintaining the Fretica package used for the analysis of single-molecule data, Silvia Jansen for sharing reagents, Vaclav Veverka and Evzen Boura for sharing their chemical shift perturbation data, and Giulio Tesei for useful insights regarding the simulation analysis. This research was supported by the NIH National Institute on Allergic and Infectious Diseases with R01AI163142 (to A.S., K.B.H., and A.S.H.) and by the NIH National Cancer Institute F99CA264413 (to J.J.A.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

6.16 Figures

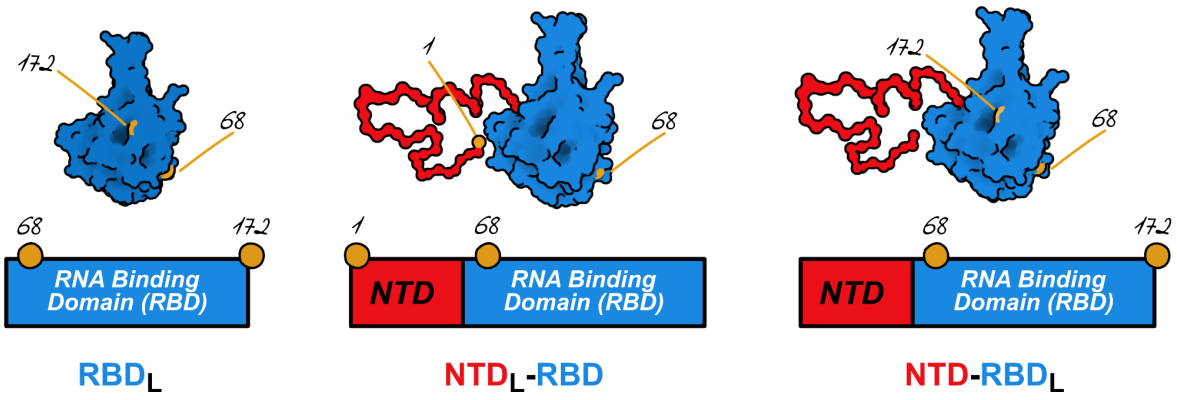


Figure 1. Nucleocapsid protein constructs in this study

(left) RNA Binding Domain (RBD) with dyes in position 68 and 172. (center) NTD-RBD construct with dyes in position 1 and 68, sampling the disordered region. (right) NTD-RBD construct with dyes in position 68 and 172 to sample conformational changes and interactions in the RBD domain.

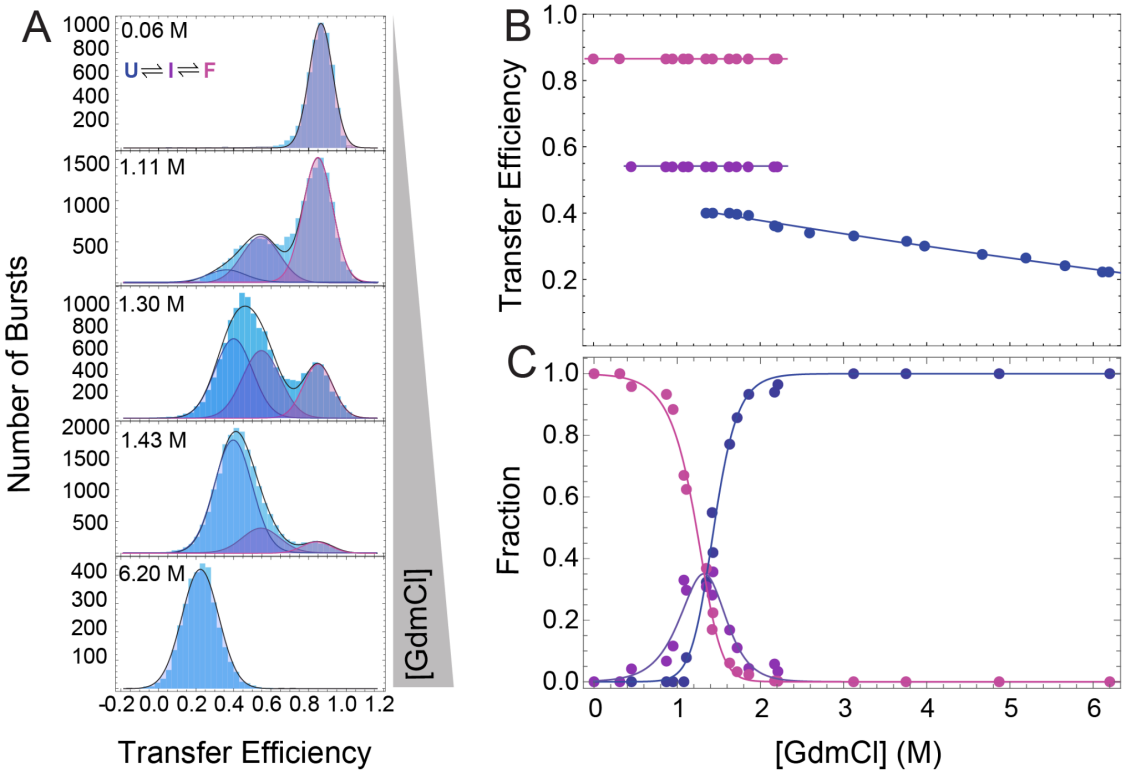


Figure 2. RNA Binding Domain (RBD) folding

A. Representative distributions of transfer efficiencies at different GdmCl concentrations. The transfer efficiency distributions are fitted with up to three Gaussian distributions. The folded configuration with high mean transfer efficiency is converted into an intermediate and unfolded state with lower mean transfer efficiencies with increasing GdmCl concentration. **B.** Mean transfer efficiencies obtained from a global fit of the histograms (see **Supplementary Information**) for the folded (magenta), intermediate (purple), and unfolded (blue) populations. Lines are guides for the eyes. **C.** Corresponding fractions of the folded (magenta), intermediate (purple), and unfolded (blue) populations. Lines represent a fit to the corresponding thermodynamic equilibrium according to **Eq. S6** and **S7.**

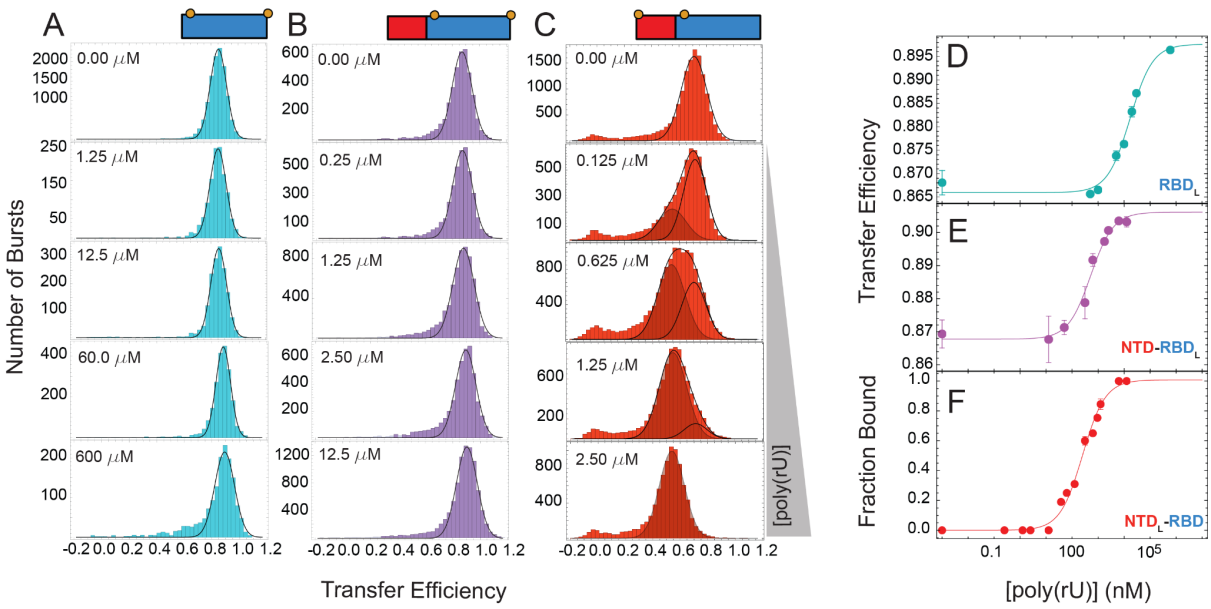


Figure 3. poly(rU) binding to RBD and NTD-RBD

A. Representative distributions of transfer efficiencies at different concentrations of poly(rU) for RBD_L. Distributions are fitted to a single Gaussian distribution. **B.** Representative distributions of transfer efficiencies at different concentrations of poly(rU) for NTD-RBD_L. Distributions are fitted to a single Gaussian distribution. **C.** Representative distributions of transfer efficiencies at different concentrations of poly(rU) for NTD_L-RBD. Distributions are fitted to two Gaussian distributions. **D.**Variations in the mean transfer efficiency of RBD_L upon binding poly(rU). **E.** Variations in the mean transfer efficiency of NTD-RBD_L upon binding poly(rU). **F.** Fraction bound of NTD_L -RBD as a function of poly(rU) concentration. Solid lines represent the fit to the binding equations **Eq. 3**. Best fit values of K_{int} are shown in Supplementary Table 6.

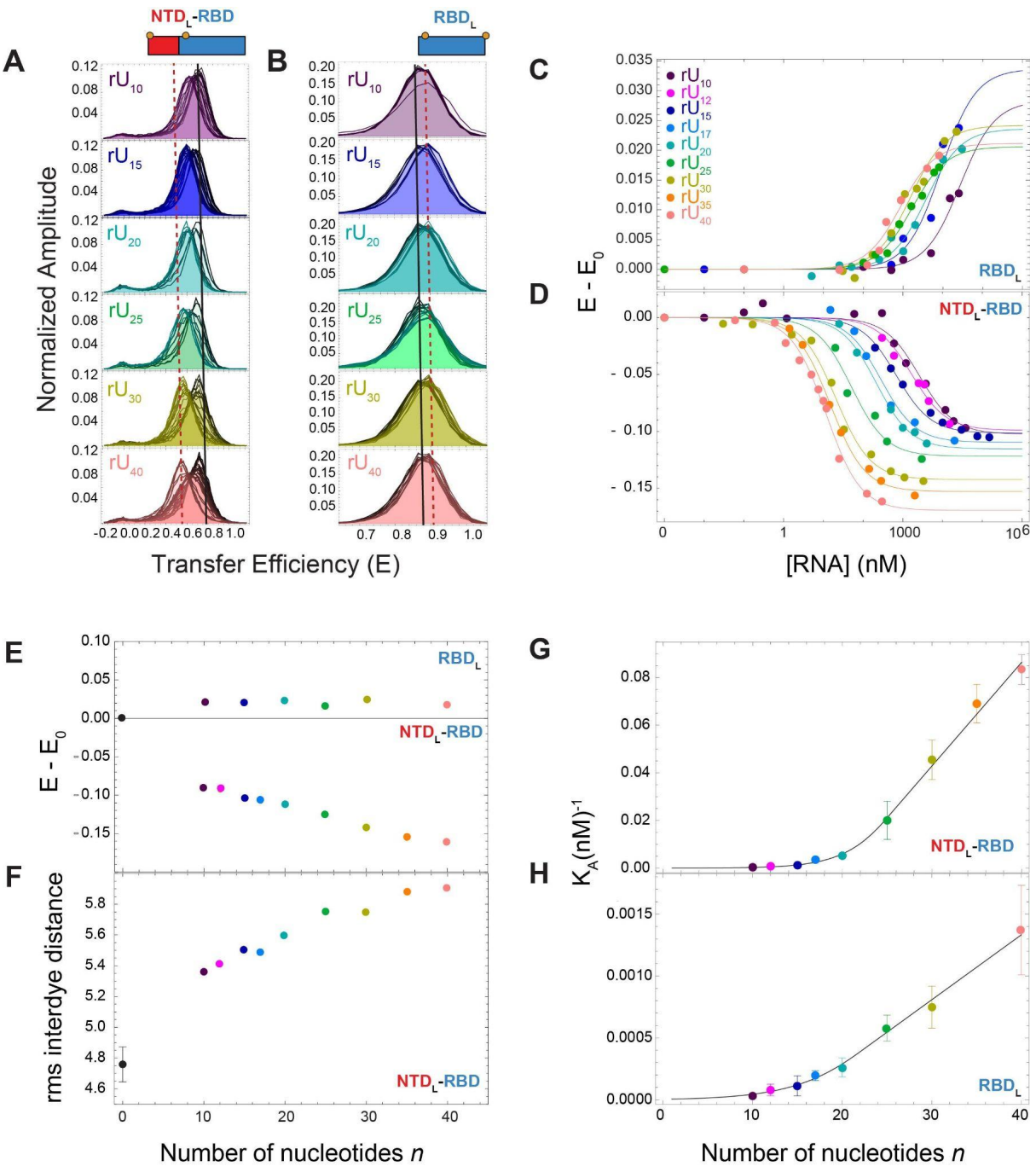


Figure 4. Length dependence of poly(rU) binding to NTD-RBD and RBD

A-B. Representative histograms of NTD$_L$-RBD (**A**) and RBD$_L$ (**B**) for rU$_n$ with nucleotide length n equal to 10, 15, 20, 25, 30, 40. The line of the transfer efficiency distribution varies from black (no RNA, starting condition) to the representative color of the specific length with increasing

concentration of RNA. Black solid vertical line identifies the mean transfer efficiency at the starting condition (E_0), red vertical dashed line identifies the mean transfer efficiency at "saturation". **C-D.** Transfer efficiency changes upon $(rU)_n$ binding, $E\text{-}E_0$, for RBD_L (**C**) and NTD_L-RBD (**D**) for all nucleotide lengths. Compare with single titrations in **Supplementary Fig. 1** for replicates and errors associated with each point. Solid lines are fit to **Eq. 1**. **E.** Variation range of transfer efficiency E with respect to the transfer efficiency E_0 measured in absence of ligands for both NTD_L-RBD and RBD_L constructs. **F.** Root-mean-square (rms) interdye distance of the disordered tail as measured by the labeling positions in NTD_L-RBD and as a function of nucleic acid length. **G-H**. Association constants as a function of the number of nucleotide bases in $(rU)_n$.

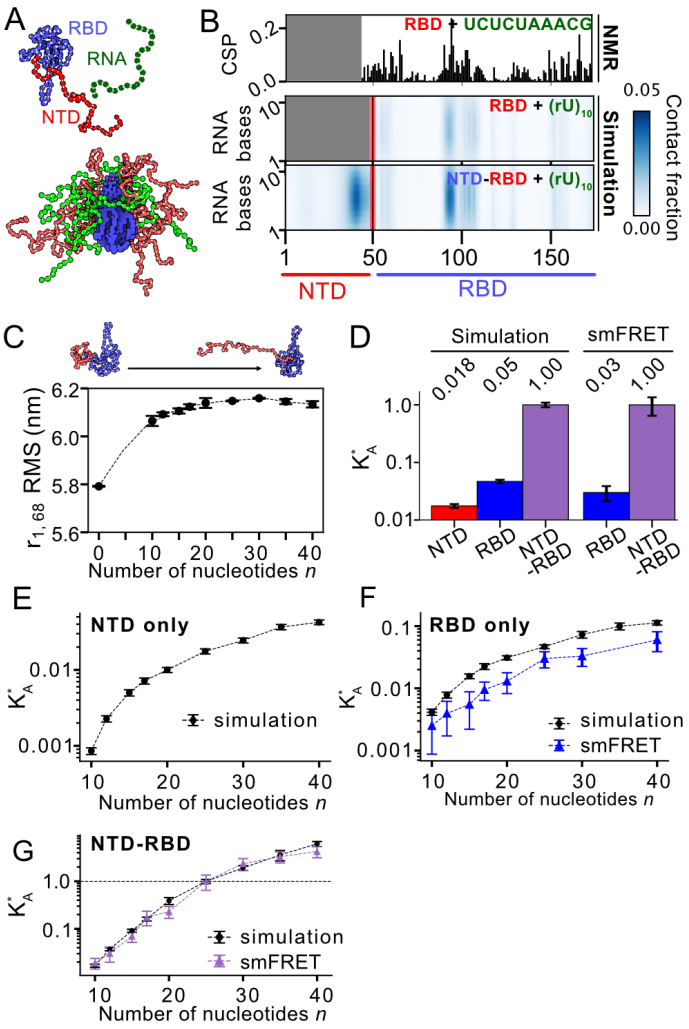


Figure 5. Coarse-grained simulations of the Nucleocapsid protein with ssRNA

A. The Mpipi forcefield is used to model SARS-CoV 2 N-protein interactions with ssRNA $(rU)_n$[759]. Each amino acid and nucleotide is represented as a single bead (see Methods). The Nucleocapsid-RNA bound state is highly dynamic (bottom). **B.** Simulations of RBD + $(rU)_{10}$ (middle) or NTD-RBD + $(rU)_{10}$ (bottom) enable the assessment of which residues engage in direct RNA interactions. Protein:RNA contacts are quantified by calculating the contact fraction, defined as the fraction of the simulation in which each amino acid-nucleotide pair is under a threshold distance of 14 Å. The specific threshold chosen does not alter which residues are identified as RNA-interacting

(**Supplementary Fig. 3**). The pattern of residues identified from simulations shows qualitative agreement with chemical shift perturbation data of the RBD (amino acids 44-173) observed upon binding to a 10-mer ssRNA (5'- UCUCUAAACG-3')[525]. **C.** Root-mean-square distance (RMSD) between residues 1 and 68 increases upon ssRNA binding, with a modest increase observed in the RNA-bound state as a function of RNA length up to $(rU)_{20}$. **D.** The normalized binding affinity (K_A^*) of the NTD, RBD, or NTD-RBD binding to $(rU)_n$ is calculated as the apparent binding affinity divided by the apparent binding affinity for NTD-RBD binding $(rU)_{25}$. K_A^* can be calculated in a self-consistent manner for simulations (left) and experiment (right). **E.** Length dependent K_A^* of the NTD + $(rU)_n$. **F.** Length dependent K_A^* of the RBD + $(rU)_n$. **G.** Length-dependent K_A^* of the NTD-RBD + $(rU)_n$. For **E**,**F** and **G,** K_A^* is calculated by dividing the apparent K_A from the specific $(rU)_n$ length by the apparent K_A from the NTD-RBD + $(rU)_{25}$ simulation.

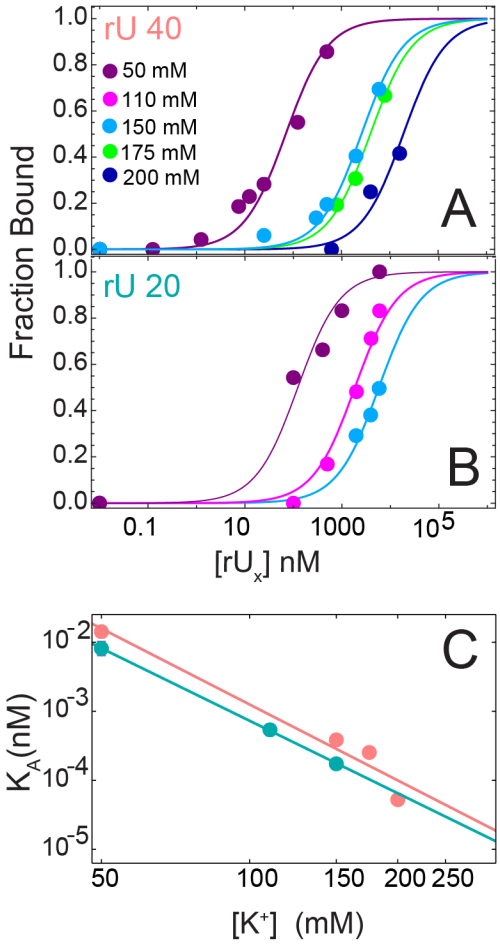


Figure 6. Salt dependence of binding association constant

Fraction bound is determined from single-molecule FRET experiments of the NTD_L-RBD as a function of $(rU)_{40}$ (**A**) and $(rU)_{20}$ (**B**) concentration. Each curve is measured in 50 mM Tris buffer and increasing KCl concentration: 50 mM (purple), 110 mM (magenta), 150 mM (cyan), 175 mM (green), 200 mM (blue) KCl. See corresponding histograms in **Supplementary Fig. 6-7** and **3**. Solid lines are fit to **Eq. 2a**. **C.** Association constants determined from the measurements in panel A ($(rU)_{40}$, pink) and panel B ($(rU)_{20}$, cyan) are plotted against the concentration of K^+ ions on a log-log plot. Solid lines represent the linear fit of Log(K_A) as a function of Log([K^+]). Results for total ion concentration are reported in **Supplementary Fig. 8**. The similar slope of $(rU)_{40}$ and $(rU)_{20}$ data

suggests that the same net ion release occurs upon binding of the two different lengths of nucleic acids (see **Supplementary Table 11**).

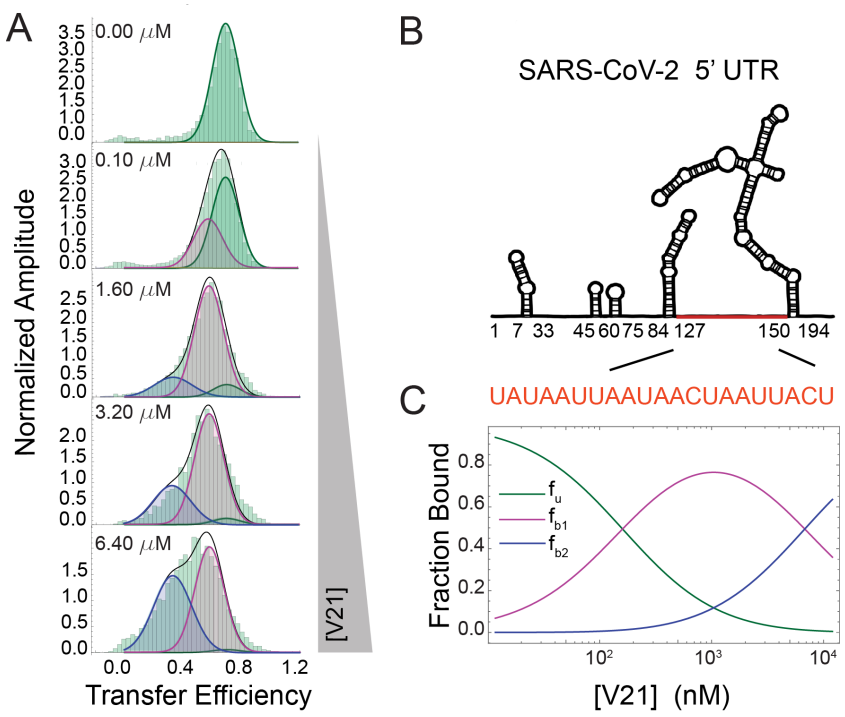


Figure 7. Specific ssRNA binding to NTD_L-RBD

A. Representative distributions of transfer efficiencies upon binding of V21. Increasing concentration of RNA leads to a first conformational change of the tail that appears to be largely completed at ~3 μM. Further increasing the concentration of V21 leads to a second conformational change of the disordered region, indicating that the protein is binding two copies of the nucleic acids. Areas are fitted according to **Eq. 2b** and **2c**. **B.** Graphical representation of the SARS-CoV-2 5' UTR based on Iserman et al.[442], highlighting the region corresponding to V21. **C.** Fraction of each state: unbound (f_u), bound to one V21 molecule (f_{b1}), and bound to two V21 molecules (f_{b2}). Corresponding values of the fit are reported in **Supplementary Table 13**.

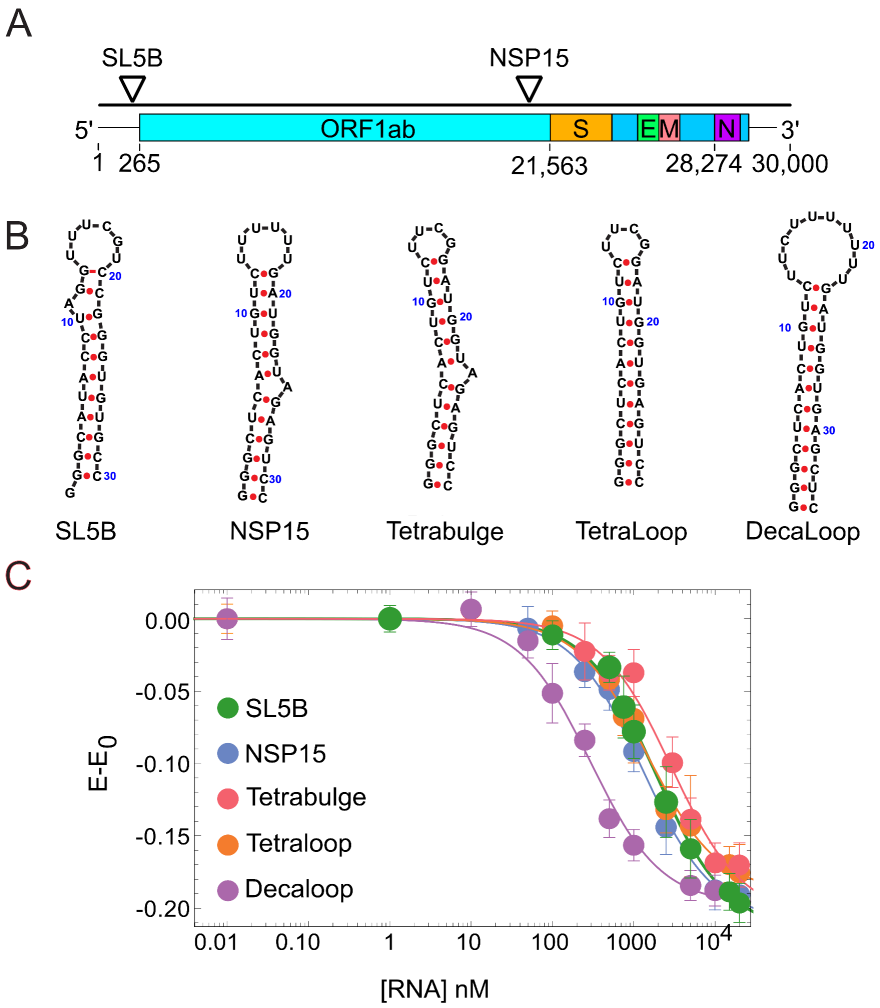


Figure 8. Specific hairpin RNA (hpRNA) binding to NTD-RBD

A. Position of studied hpRNA sequences in the viral genome. **B.** Hairpin structure and sequence. **C.** Variation in the mean transfer efficiencies of the NTD_L-RBD as a function of hpRNA concentration. When no hpRNA is present, transfer efficiency is ~0.68 (compare with **Supplementary Figure 8**). Solid lines are fit to **Eq. 1**.

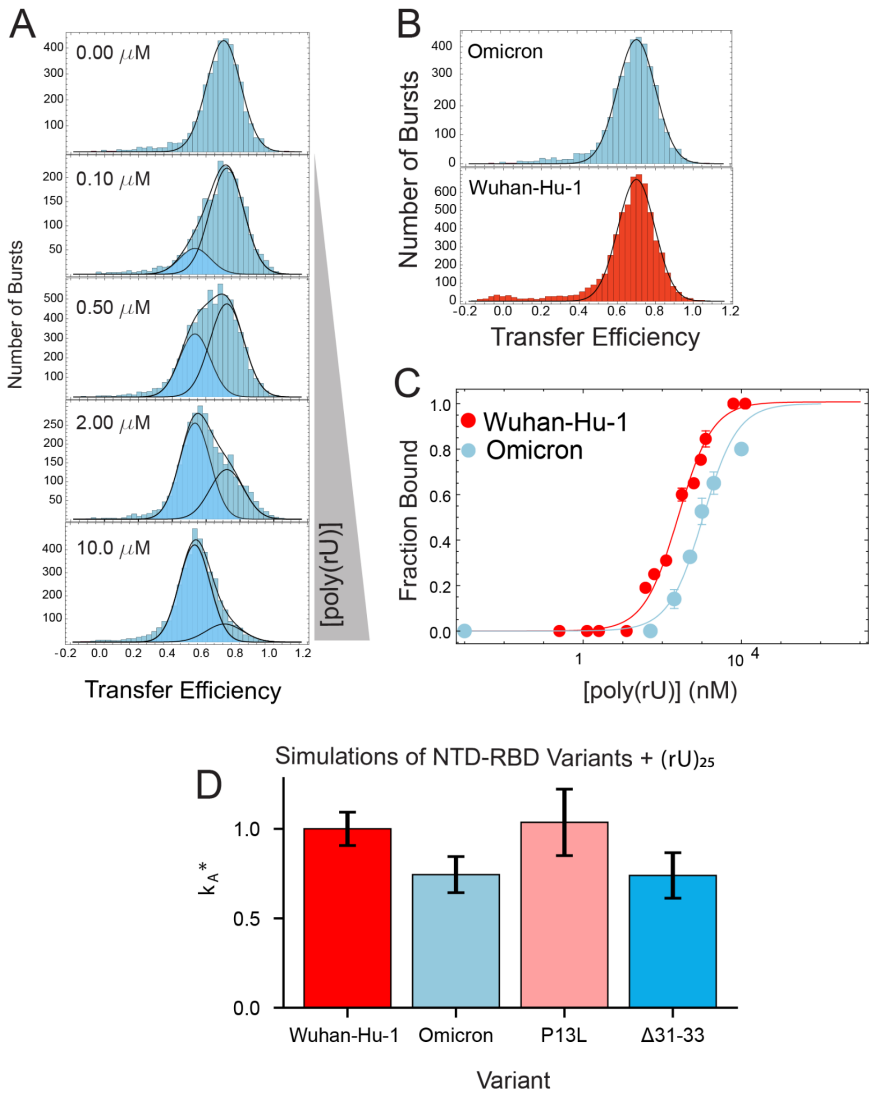


Figure 9. Omicron variant

A. Transfer efficiency distributions for the Omicron variant as function of poly(rU) concentration. Distributions are fitted with up to two Gaussian distributions to quantify the mean transfer efficiency and relative fraction of bound and unbound fractions. **B.** Comparison of unbound configuration of disordered tail for Wuhan-Hu-1 (red) and Omicron variant (cyan) reveals no significant variations in overall conformations. **C.** Comparison of binding affinity for Wuhan-Hu-1 (red) and Omicron variant (cyan) reveals different affinities for poly(rU). Solid lines are fit to **Eq.**

2a. **D.** Trend of the normalized binding affinity (K_A^*) predicted by simulations with Mpipi model for the Omicron mutant and additional variants.

6.17 Supplementary Information

Experimental setup and procedure for single-molecule fluorescence experiments. Single-molecule confocal fluorescence measurements are performed on a Picoquant MT200 instrument (Picoquant, Germany). To enable Pulsed Interleaved Excitation (PIE), we synchronize a diode laser (LDH-D-C-485, PicoQuant, Germany) and a supercontinuum laser (SuperK Extreme, NKT Photonics, Denmark), filtered by a z582/15 band pass filter (Chroma) pulsed at 20 MHz such that a delay of approximately 25 ns occurs between each laser pulse. Lasers are focused in the sample through a 60x1.2 UPlanSApo Superapochromat water immersion objective (Olympus, Japan). Emitted photons are collected through the same objective, passed through a dichroic mirror (ZT568rpc, Chroma, USA), and further filtered by a long pass filter (HQ500LP, Chroma Technology) to suppress scattering light. After passing through the confocal pinhole (100 μm diameter), the emitted photons are separated into four channels by a polarizing beam splitter (which differentiates between perpendicular and parallel polarization), followed by a dichroic mirror (585DCXR, Chroma) that further discriminates between donor and acceptor photons. Donor and acceptor emission is then filtered using band pass filters, ET525/50m or HQ642/80m (Chroma Technology), respectively, and finally focused on SPAD detectors (Excelitas, USA). The arrival time of every photon is recorded with a HydraHarp 400 TCSPC module (PicoQuant, Germany). FRET experiments are performed by exciting the donor dye with a laser power of 100 μW (measured at the back aperture of the objective), whereas acceptor direct excitation is adjusted to match a total emission intensity after acceptor excitation to the one observed upon donor excitation (between 50 and 70 μW). Single-molecule FRET efficiency histograms are acquired at labeled protein concentrations between 50 pM and 100 pM, estimated from dilutions of samples with known concentration, as previously determined *via* absorbance measurements.

All measurements, unless differently specified, were performed in 50 mM Tris pH 7.4, 200 mM β-mercaptoethanol (for photoprotection), 0.001% Tween20 (for surface passivation) and GdmCl at the reported concentrations. All measurements were performed in uncoated polymer coverslip cuvettes (Ibidi, Wisconsin, USA). When using denaturant or salt, the exact concentration is determined from measurement of the solution refractive index with an Abbe refractometer (Bausch & Lomb, USA).

Each sample was measured for at least 10 min at room temperature (295 ± 0.5 K) and all measurements were performed at least in duplicate (independent replicates from a new sample preparation) to confirm reproducibility of the results.

Construction of transfer efficiency histograms.

Fluorescence bursts were identified by time-binning photons in bins of 1 ms and accepting bursts whose total number of photons after donor excitation was larger than at least 10 photons in each bin. Contiguous bins were merged if the total number of photons was larger than at least 20 photons. The exact threshold was selected based on the background contribution identified in the photon counting histograms with 1 ms binning. A minimum common threshold across constructs has been used to minimize variations in the width of the transfer efficiency distributions due to the difference in the acceptance thresholds, as expected for a shot-noise-limited system. Transfer efficiencies for each burst were calculated according to

$$E = n_A/(n_A + n_D) \qquad \text{(Eq. S1)}$$

where n_A and n_D are the numbers of donor and acceptor photons, respectively.

Reported transfer efficiencies are corrected for background, acceptor direct excitation, channel crosstalk, differences in detector efficiencies, and quantum yields of the dyes.

Similarly to transfer efficiency, the labeling stoichiometry ratio S is computed accordingly to:

$$S = I_D/(I_D + \gamma_{PIE}\, I_A) \qquad \text{(Eq. S2)}$$

where I_D and I_A represent the total intensities observed after donor and acceptor excitation and γ_{PIE} provides a correction factor to account for the differences between donor and acceptor in detection efficiency and laser intensities. In the histograms, we present the bursts with stoichiometry corresponding to 1:1 donor:acceptor labeling (in contrast to donor and acceptor only populations), which are selected according to the criterion $0.3 < S < 0.7$. Variations in the selection criteria for the stoichiometry ratio do not impact significantly the observed mean transfer efficiency (within experimental errors).

Fit of transfer efficiency distributions

To estimate the mean transfer efficiency and extract multiple populations from the transfer efficiency histograms, each population was approximated with either a Gaussian or a LogNormal distribution function. When fitting more than one peak, the histogram is analyzed with a sum of Gaussian and/or LogNormal functions. When analyzing multiple overlapping populations, in order to limit the model parameters and potential overfitting, we favored the use of global fit analysis, where some parameters are shared across multiple or all concentrations.

Determination of root mean square interdye distances from mean FRET transfer efficiencies.

Conversion of mean transfer efficiencies to an interdye distance for fast rearranging ensembles requires the assumption of a distribution of distances. Here we employed the Gaussian model (see Cubuk *et al.* 2021 [441] where we compared this model to the self avoiding random walk model). In the Gaussian model, the conversion rely on one single fitting parameter, the root mean square interdye distance $r = < R^2 >^{1/2}$.

Estimates of this parameter is obtained by numerically solving:

$$< E > = \int_0^\infty \quad E(R)\, P(R)\, dR \qquad \text{(Eq. S3)}$$

where *R* is the interdye distance, *P(R)* represents the chosen distribution, and *E(R)* is the Förster equation for the dependence of transfer efficiency on distance *R* and Förster radius R_0:

$$E(R) = \frac{R_0^6}{R_0^6 + R^6} \qquad \text{(Eq. S4)}$$

The Gaussian chain distribution is given by:

$$P(R) = 4\pi R^2 \left(\frac{3}{2\pi r^2}\right)^{3/2} exp\left(\frac{-3R^2}{2r^2}\right) \qquad \text{(Eq. S5)}$$

Eqs. S4 and **S5** are substituted into **Eq. S3** and *r* is numerically optimized such that in integral equals the experimentally determined value for mean transfer efficiency.

Folding equilibrium of the RBD

The folding equilibrium of the RBD revealed the occurrence of three distinct states: native (N), intermediate (I), and unfolded (U). To quantify the thermodynamic properties of the three-state

equilibrium N ⇌ I ⇌ U, the corresponding fraction folded, intermediate, and unfolded can be written in terms of the equilibrium constant K_{UN} and K_{NI} as:

$$f_U = 1/(1 + K_{UI} + K_{UI}K_{IN}) \quad \text{(Eq. S6a)}$$

$$f_I = K_{UI}/(1 + K_{UI} + K_{UI}K_{IN}) \quad \text{(Eq. S6b)}$$

$$f_N = K_{UI}K_{IN}/(1 + K_{UI} + K_{UI}K_{IN}) \quad \text{(Eq. S6c)}$$

T

he equilibrium constant K_{UN} and K_{NI} can be expressed as:

$$K_{UI} = exp\left[\Delta G_0^{UI}/RT\ \left(c - c^{UI}\right)/c^{UI}\right] \quad \text{(Eq. S7a)}$$

$$K_{IN} = exp\left[\Delta G_0^{IN}/RT\ \left(c - c^{IN}\right)/c^{IN}\right] \quad \text{(Eq. S7b)}$$

where ΔG_0^{UI} and ΔG_0^{IN} are the free energy differences in aqueous buffer conditions between the U and I and I and N states, respectively, and c^{UI} and c^{IN} are the concentrations where the corresponding fraction curves cross each other. It is important to note that whereas in the case of a simple two-state system N ⇌ U, the corresponding c^{UN} represents the midpoint of the folding transition, in the general case with more than two states, the crossing points do not necessarily occur at the midpoint (50%) of the transition.

Equilibrium binding models

Here, we describe the assumptions behind the models for nonspecific interaction of monomeric N-protein with ssRNA. The models are derived for the specific case of the performed single-molecule experiments, where binding experiments were conducted at concentrations of protein much lower than the concentration of nucleic acid.

In all cases, we assume that binding takes place in a single orientation of the protein relative to the

nucleic acid 3´-5´ polarity (though this can be easily extended to the more general case and does not significantly affect the interpretation of our results).

The observed association constants are expressed as the sum of intrinsic association constants for binding of the protein to each available position along the nucleic acid strand. We define a "position" as contiguous stretch of nucleotides that represent the protein's binding footprint, i.e.:

$$K_A = \frac{\sum_i \quad [(PR_M)_{position\ i}]}{[P][R_M]} = \sum_i \qquad K_{position\ i} \qquad \text{(Eq. S8)}$$

In these models, we assume that the oligonucleotides are homogeneous, made of repetitive superimposed segments presenting the same affinity for the protein, with periodicity length equal to 1 nucleotide; Coulombic end effects on counterion condensation and protein binding on the nucleic acid are not considered (see, for example, the work by Shkel, Ballin and Record [774]). The value of the intrinsic association constant $K_{int,m}$ for each available position is only dependent on the site size m (i.e., the number of contiguous nucleotides involved in the interaction) but not on its position along the nucleic acid (we neglect end effects and position specificity). Under these assumptions the association constant can be written as

$$K_A = \frac{\sum_i \quad [(PR_M)_{position\ i}]}{[P][R_M]} = \sum_i \qquad K_{position\ i} =$$

$$\sum_{m=i}^{M} \quad (\#\ positions\ with\ m\ cont.nt) K_{int,m} \qquad \text{(Eq. S9)}$$

Single binding mode, no overhangs

We first consider the case of a single binding mode with no overhangs. In this scenario:

- the protein only binds if the oligonucleotide length M is equal or longer than its

contact site size, *n*; if $M < n$, it does not bind, i.e. $K_A = 0$;

- it binds with equal affinity, K_{int}, to all possible contiguous stretches of *n* nucleotides on the oligonucleotide, which can be counted to be $M - n + 1$;
- it does not bind through stretches of contiguous nucleotides shorter than the contact site size *n*.

Under these assumptions, the association constant can be written as:

$$K_A = \frac{\sum_i \ [(PR_M)_{position\ i}]}{[P][R_M]} = \sum_i \quad K_{position\ i} = K_{int}0 \quad for\ M < n \qquad \text{(Eq. S10a)}$$

$$K_A = \frac{\sum_i \ [(PR_M)_{position\ i}]}{[P][R_M]} = \sum_i \quad K_{position\ i} = K_{int}(M - n + 1) \quad for\ M \geq n \qquad \text{(Eq. S10b)}$$

Single binding mode, with 'overhangs'

In this version of the model, the protein can bind to oligonucleotides of any length:

- if $M \geq n$, the protein can either bind in full length sites, spanning *n* nucleotides, or to ends of the oligonucleotide, making contacts with a number of nucleotides smaller than *n*, leaving a protein 'overhang' that does not make contact with the nucleic acid;
- if $M < n$, the oligonucleotide can bind in different positions on the protein, spanning different portions of its nucleic acid binding site; these short oligos can bind within the binding site on the protein, or on the edges of the binding site leaving unbound nucleotide overhangs;
- in all cases, the protein interacts with a stretch of contiguous nucleotides, and the association constant for binding with a given number *m* of contiguous nucleotides is equal to the product of an intrinsic association constant $K_{int,\ m}$, times the number of possible configurations for the given values of *M* and *n;*

- the protein interacts with stretches of length $m < n$ only if there is no available nucleotides in one of the sides of the stretch; *i.e.*, only if binding to an end of an oligo or to an oligo with total length $M < n$;
- in addition to the fixed binding polarity, it is assumed that the nucleic acid binding site in the protein, able to interact with a contiguous stretch of nucleotides of length n, interacts with a short stretch of contiguous nucleotides, $m < n$, independently on where the stretch is located along the nucleic acid binding site; therefore,

$$K_A = K_{int,M}(n - M + 1) + 2\sum_{j=1}^{M-1} K_{int,j} \quad for\ M < n \qquad \text{(Eq. S11a)}$$

$$K_A = K_{int}(M - n + 1) + 2\sum_{j=1}^{n-1} K_{int,j} \quad for\ M \geq n \qquad \text{(Eq. S11b)}$$

- the values of intrinsic association constants with stretches of nucleotides shorter than n, $K_{int,\,m}$, are given by

$$K_{int,m} = K_{in,m}K_{tg} = (K_{in,n})^{\frac{m}{n}}K_{tg} = exp[\tfrac{\Delta G_{in,m}+\Delta G_{tg}}{RT}] = exp[\tfrac{(\Delta G_{in}/n)\,m\,+\Delta G_{tg}}{RT}] \qquad \text{(Eq. S12)}$$

The terms in the equation can be conceptualized with the following reaction scheme:

$$\mathrm{P} + \mathrm{R}_m \overset{K_{tg}}{\rightleftharpoons} \mathrm{P{\cdot}R}_m \overset{K_{in,m}}{\rightleftharpoons} \mathrm{PR}_m \qquad \text{(Eq. S13)}$$

where the first step is the bimolecular encounter of the protein P and the stretch of m nucleotides R_m, in the proper orientation for binding, and the second step is the actual establishment of the interactions between the protein and the nucleic acid site (compare with concepts derived by Lou and Sharp [775], equations 1-9)

This equation for $K_{int,m}$ involves the assumption that the translational-rotational entropic cost of the bimolecular encounter in the proper orientation (ΔG_{tg} = -RT LogK_{tg}) is independent of m. Also it involves the assumption that the contribution of the actual interactions established upon

binding, -enthalpic and entropic components, such as counterion release- to the binding free energy is equally subdivided per nucleotide constituting the protein-nucleic acid binding interface ($\Delta G_{in,m} = m/n\, \Delta G_{in}$, or in terms of equilibrium constants, $K_{in,m} = (K_{in})^{m/n}$). Therefore:

$$K_A = K_{tg}\{(K_{in})^{M/n}(n - M + 1) + 2\sum_{j=1}^{M-1} (K_{in})^{j/n}\} \quad for\ M < n$$

(Eq. S14a)

$$K_A = K_{tg}\{K_{in}(M - n + 1) + 2\sum_{j=1}^{n-1} (K_{in})^{j/n}\} \quad for\ M \geq n$$

(Eq. S14b)

$$K_A = \frac{K_{int}}{K_{in}}\{(K_{in})^{M/n}(n - M + 1) + 2\sum_{j=1}^{M-1} (K_{in})^{j/n}\} \quad for\ M < n$$ (Eq. S14c)

$$K_A = \frac{K_{int}}{K_{in}}\{K_{in}(M - n + 1) + 2\sum_{j=1}^{n-1} (K_{in})^{j/n}\} \quad for\ M \geq n$$

(Eq. S14d)

where the first term represents the binding to the longest available stretch of nucleotides (*M* if $M<n$, or *n* if $M>n$) and the summation on the second term represents the binding to ends of the oligo with ends of the nucleic acid binding site on the protein involving shorter stretches of nucleotides.

The summation terms can be conveniently replaced by

$$\sum_{j=1}^{M-1} (K_{in})^{j/n} = (K_{in})^{M/n}\frac{1-(K_{in})^{(1-M)/n}}{(K_{in})^{1/n}-1}$$ (Eq. S15a)

$$\sum_{j=1}^{n-1} (K_{in})^{j/n} = K_{in}\frac{1-(K_{in})^{(1-n)/n}}{(K_{in})^{1/n}-1}$$ (Eq. S15b)

Then we have:

$$K_A = K_{int}\{(K_{in})^{M/n}[(n - M + 1) + 2\frac{1-(K_{in})^{(1-M)/n}}{(K_{in})^{1/n}-1}]\}\ for\ M < n$$ (Eq. S16a)

$$K_A = K_{int}(M - n + 1) + 2\frac{1-(K_{in})^{(1-n)/n}}{(K_{in})^{1/n}-1} \; for\; M \geq n \qquad \text{(Eq. S16b)}$$

or, in terms of free energy,

$$K_A = K_{int}e^{\frac{\Delta G_{in}}{RT}(\frac{n-M}{n})}[(n - M + 1) + 2\frac{e^{\frac{\Delta G_{in}}{RT}(\frac{M}{n})}(1-e^{\frac{\Delta G_{in}}{RT}(\frac{1-M}{n})})}{e^{\frac{\Delta G_{in}}{RT}(\frac{1}{n})}-1}]\; for\; M < n \qquad \text{(Eq. S17a)}$$

$$K_A = K_{int}[(M - n + 1) + 2\frac{e^{\frac{\Delta G_{in}}{RT}}(1-e^{\frac{\Delta G_{in}}{RT}(\frac{1-n}{n})})}{e^{\frac{\Delta G_{in}}{RT}(\frac{1}{n})}-1}]\; for\; M \geq n \qquad \text{(Eq. S17b)}$$

Nanosecond FCS analysis

Autocorrelation curves of acceptor and donor channels and cross-correlation curves between acceptor and donor channels were calculated with the methods described previously[170,736]. All samples were measured at single-molecule concentrations (~100 pM), and bursts corresponding to the donor-acceptor population transfer efficiency were selected to eliminate the contribution of donor-only to the correlation amplitude. Finally, the correlation was computed over a time window of 5 μs, and characteristics timescales were extracted according to:

$$g_{ij}(\tau\ \ ) = 1 + \frac{1}{N}(1 - c_{AB}Exp[-(\tau - \tau_0)/\tau_{AB}]) \times$$

$$\times (1 + c_{CD}Exp[-(\tau - \tau_0)/\tau_{CD}])(1 + c_TExp[-(\tau - \tau_0)/\tau_T]) \qquad \text{(Eq S18)}$$

where N is the mean number of molecules in the confocal volume and i and j indicate the type of signal (either from the Acceptor or Donor channels). The three multiplicative terms describe the contribution to amplitude and timescale of photon antibunching (AB), chain dynamics (CD), and triplet blinking of the dyes (T). τ_{CD} is then converted in the reconfiguration time of the interdye distance $\tau_r$correcting for the filtering effect of FRET as described previously [197].

Coarse-grained simulations

Coarse-grained simulations were performed using the Mpipi model[759]. In Mpipi, each bead (amino acid or nucleotide) is chemically unique, and inter-bead interactions contain contributions from a short-range Wang-Frenkel potential and, where applicable, a long-range Coulombic potential for beads with a net charge[776]. The Coulombic potential takes the ionic strength into account, and simulations were performed at an equivalent of 50 mM NaCl. The parameters associated with the inter-bead Wang-Frenkel potential were determined through a combination of all-atom and quantum mechanical simulations and capture a mixture of Van der Waal interactions, cation-pi and pi-pi interactions.

As in previous work, folded domains were modeled as rigid bodies, whereas intrinsically disordered regions and ssRNA were described as flexible polymers[759,760]. Beads found within the core of globular domains ("buried" residues) have their interaction strength scaled down, as in the original Mpipi implementation.

Calculating apparent association constants from simulations

To determine the apparent association constants (K_A) for simulations, we calculated the fraction of frames in which protein and RNA were bound. To determine the bound fraction requires a definition for protein:RNA binding. We applied a measure whereby binding is determined based on consecutive simulation frames in which the protein and RNA centers-of-mass (COM) are under an RNA-length dependent threshold. This approach is motivated by the fact that histograms of the protein:RNA COM clearly show two distributions; a bound COM distance distribution and an unbound COM distance distribution (**Supplementary Fig. 5B**). As the RNA becomes longer, the separation between these two peaks changes (as the peak of the bound

distribution shifts to larger values due to the larger RNA molecule). These histograms enable us to define an RNA-length-specific distance threshold for each simulation. With this naive cutoff defined, we define binding as five or more consecutive frames where the protein and RNA COM are under the predefined threshold distance. The use of a minimum number of consecutive frames enables us to distinguish transient random encounters between the protein and RNA from *bona fide* binding events, where protein and RNA are directly engaging (**Supplementary Fig. 5C,D)**.

Having determined the fraction bound, we then calculated an apparent K_D with the expression:

$$K_D = \frac{(1-f_{bound})^2}{N_A V f_{bound}} \quad \text{(Eq. S19)}$$

where f_{bound} is the fraction of the simulation in which the two species are bound, N_A is Avogadro's constant, and V is the simulation box volume in liters, returning a K_D in mol/L. The K_A is then calculated as 1 / K_D. This approach is analogous to that of Tesei *et al.*, albeit using a different strategy to define if two molecules are bound vs. unbound[777]. Finally, having calculated the apparent association constants, we can ask how protein:RNA affinity varies across simulations of the NTD alone, RBD alone, and NTD-RBD with different lengths of $(rU)_n$.

When comparing the K_A values from simulations with experiment, we found poor agreement between the absolute values of the association constants, a feature that is commonly seen for coarse-grained models[778]. To enable a direct comparison between experiments and simulations, we calculate a normalized binding affinity (K_A^*), which we define as the ratio between the simulation (or experimental) apparent K_A for a given protein:RNA combination divided by the corresponding simulation (or experimental) apparent K_A for NTD-RBD binding to $(rU)_{25}$. This,

in effect, sets the NTD-RBD + $(rU)_{25}$ binding affinity as a reference point, and all other K_A^* values are defined as either greater than 1 (stronger binding than NTD-RBD + $(rU)_{25}$) or less than 1 (weaker binding than NTD-RBD + $(rU)_{25}$). By using this ratio, we can plot data from simulations and experiments on the same axes and compare the relative binding affinities (as a function of RNA length, protein construct, or protein sequence). This analysis reveals relatively good agreement between simulations and experiments (**Fig. 5D, F, G**), despite the many assumptions made in the coarse-grained force field.

Measuring the stability of double-stranded RNA

Absorbance is measured using a UV-Vis spectrophotometer. The sharp increase in absorbance reports on the hyperchromicity of the hairpin RNA as it converts from double-stranded (ds) to single-stranded (ss) RNA. Melting temperatures are determined by fitting absorbance values as a function of temperature to:

$$Abs = \frac{(\alpha_{ds} + \beta_{ds}T) + (\alpha_{ss} + \beta_{ss}T)e^{-m(T-T_m)}}{1 + e^{-m(T-T_m)}} \quad \text{(Eq. S20)}$$

where α_{ds} and α_{ss} refer to the absorbance of the RNA in the ds and ss state at initial temperature. β_{ds} and β_{ss} are the rate of change of the absorbance in each state as a function of temperature (T) in Kelvin. m is the m-value and T_m is the temperature at the midpoint of the transition from ds to ss RNA.

6.18 Supplementary Figures

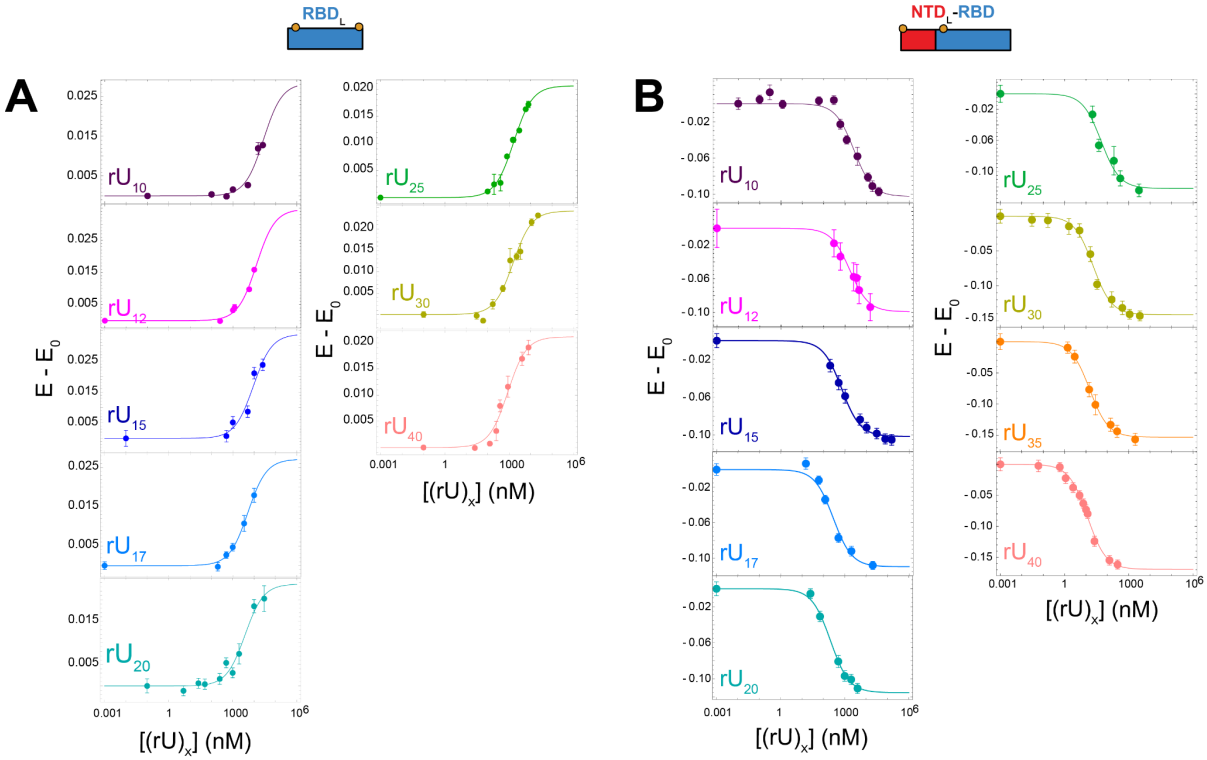


Fig. S1

Transfer efficiency variation upon binding for RBD_L (panel **A**) and NTD_L-RBD (panel **B**). Error bars represent the standard deviation of at least two independent experiments. Solid lines are fit to **Eq. 1.**

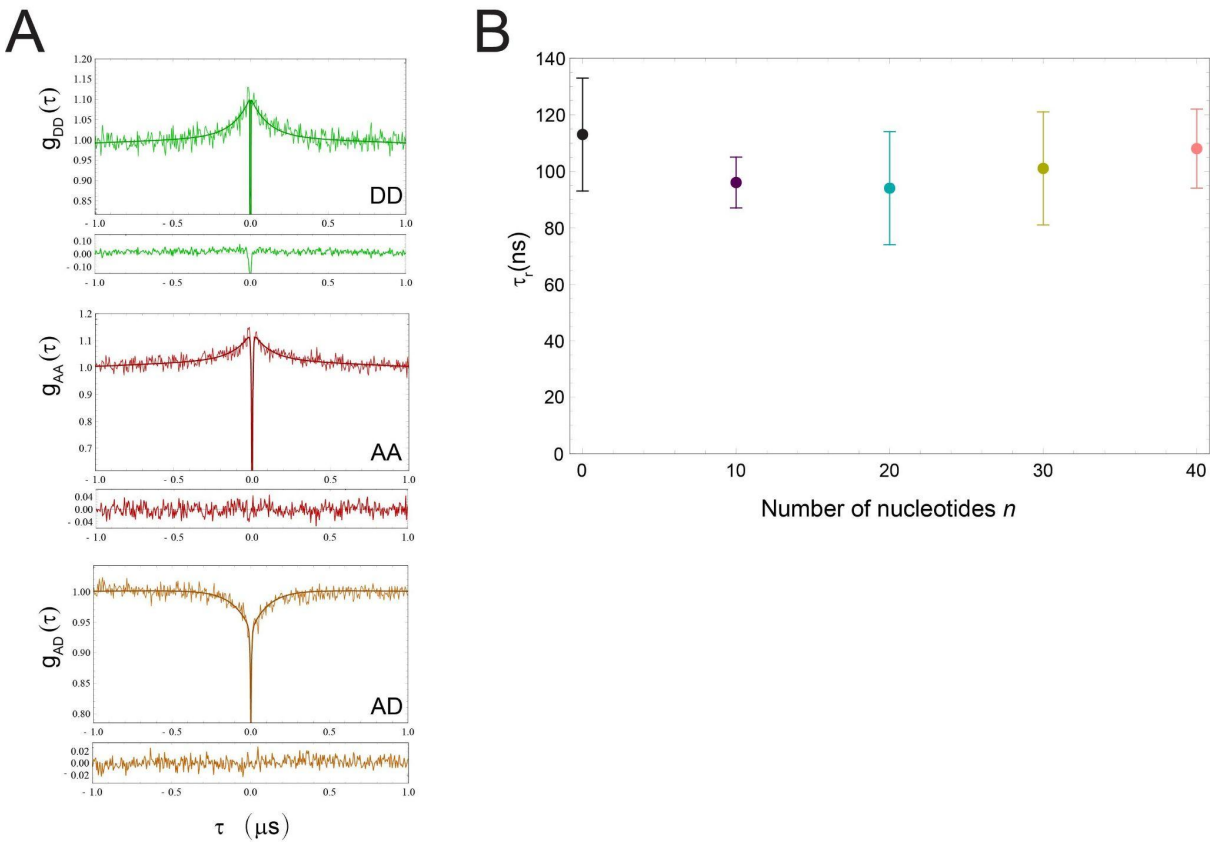


Fig. S2 Dynamics of the disordered NTD when complex with RNA

A. Example of nanosecond-FCS (nsFCS) traces of NTD_L-RBD in the presence of $(rU)_{40}$. Donor-donor, acceptor-acceptor, and donor-acceptor correlations are shown in green, red, and orange (respectively) with the fit according to **Eq. S18** and corresponding residuals. **B.** Reconfiguration times computed for the chain in the absence and in the presence of $(rU)_n$ with n = 10, 20, 30, 40.

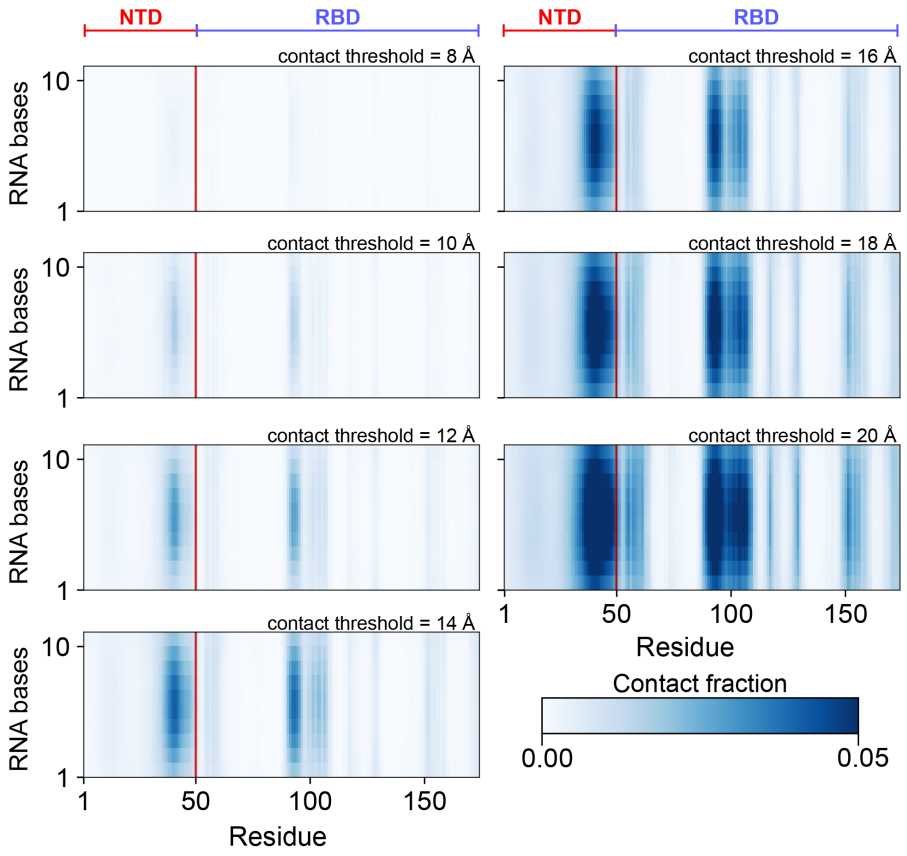


Fig. S3 NTD-RBD:$(rU)_{10}$ dependence of interacting residues on distance threshold used for contact fraction

The distance threshold used to define nucleotide:amino acid contacts was varied from 8 Å to 20 Å to assess how this altered the residues identified as RNA interacting. While, as expected, the contact fraction systematically changes as the threshold increases, the pattern of residues engaging in protein:RNA interactions remains consistent.

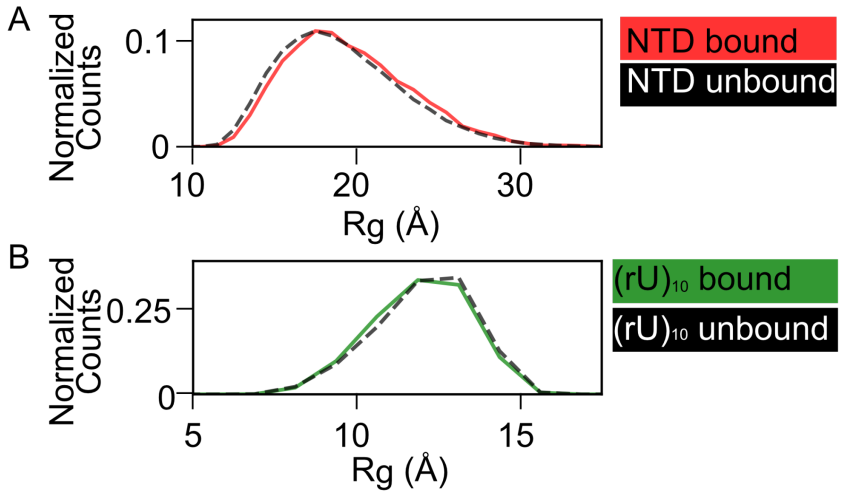


Fig. S4 The NTD and RNA remain disordered in the NTD-RBD:$(rU)_{10}$ complex

A. Histogram showing the radius of gyration (R_g) distribution for the NTD region taken from either the NTD-RBD:$(rU)_{10}$ complex (red line) or from the unbound (black line) states of NTD-RBD. The relative histogram counts have been normalized to the total number of events in the bound or unbound state. Specifically, 8,198 frames were RNA-bound in the trajectory analyzed, while 81,347 were RNA-unbound. If the NTD folded upon binding, we would expect to see a tighter distribution for the R_g at a smaller mean value, yet the R_g distribution in the bound state remains broad, with a slightly smaller mean value in the unbound state reflective of the length dependent expansion of the NTD upon binding (unbound NTD $\langle R_g \rangle$ = 19.1 Å, bound NTD $\langle R_g \rangle$ = 19.6 Å). The root-mean-square value of the end-to-end distance is reported in **Fig. 5C**. **B.** Analogous analysis from the perspective of the $(rU)_{10}$. The mean value is similar in the bound vs. unbound states (unbound $(rU)_{10}$ $\langle R_g \rangle$ = 10.6 Å, bound $(rU)_{10}$ $\langle R_g \rangle$ = 10.7 Å), but the broad distribution remains consistent with a largely disordered ensemble of conformations. See also **Supplementary Movie S1.**

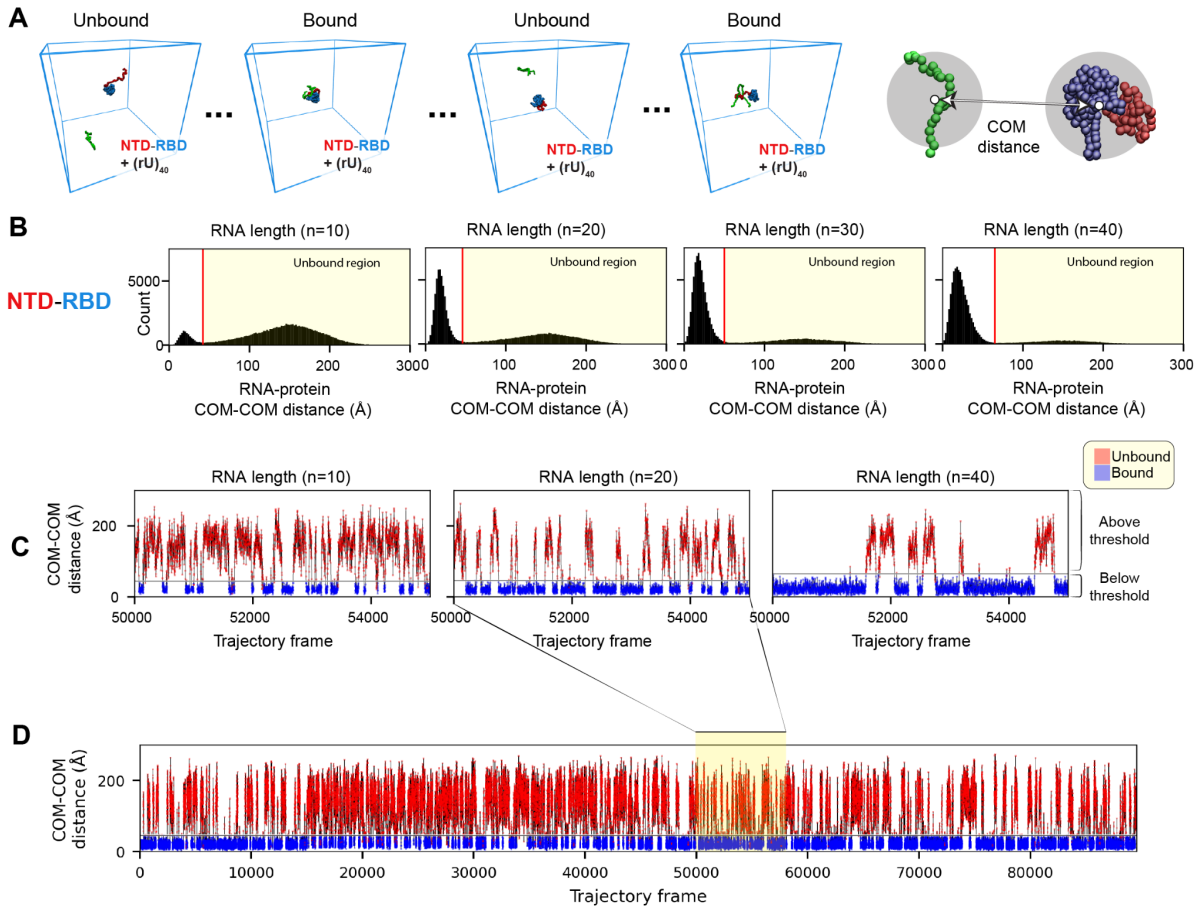


Fig. S5 Simulations of N protein construct and RNA binding

A. Example simulation snapshots from the NTD-RBD + $(rU)_{40}$ simulation showing bound and unbound configurations. On the far right a schematic of the center of mass (COM) distance is shown for NTD-RBD and $(rU)_{25}$ that are 101 Å apart. The COM for each of the two molecules is calculated using the get_center_of_mass() function in SOURSOP. **B.** Intermolecular center-of-mass (COM) distance between the protein and RNA molecules enables us to define a distance threshold that can be used to define when the two molecules are bound vs. unbound. The distance threshold for NTD-RBD binding to RNA varies between 42 Å (for RNA of length 10) and 65 Å (for RNA of length 40). Note that this distance reflects the center of mass between the two molecules, not the minimum distance. **C.** Subtrajectory taken from a simulation showing

bound and unbound states being automatically delineated based on the combination of the distance threshold introduced in panel A, alongside the requirement for five or more consecutive frames under the cutoff threshold to be used to define binding (or lack thereof). Panel C shows sub-trajectories from simulations with RNAs of length 10, 20, and 40 nucleotides. **D.** Full trajectory of simulations with RNA of length 20 showing over 200 independent binding and unbinding events for each RNA length.

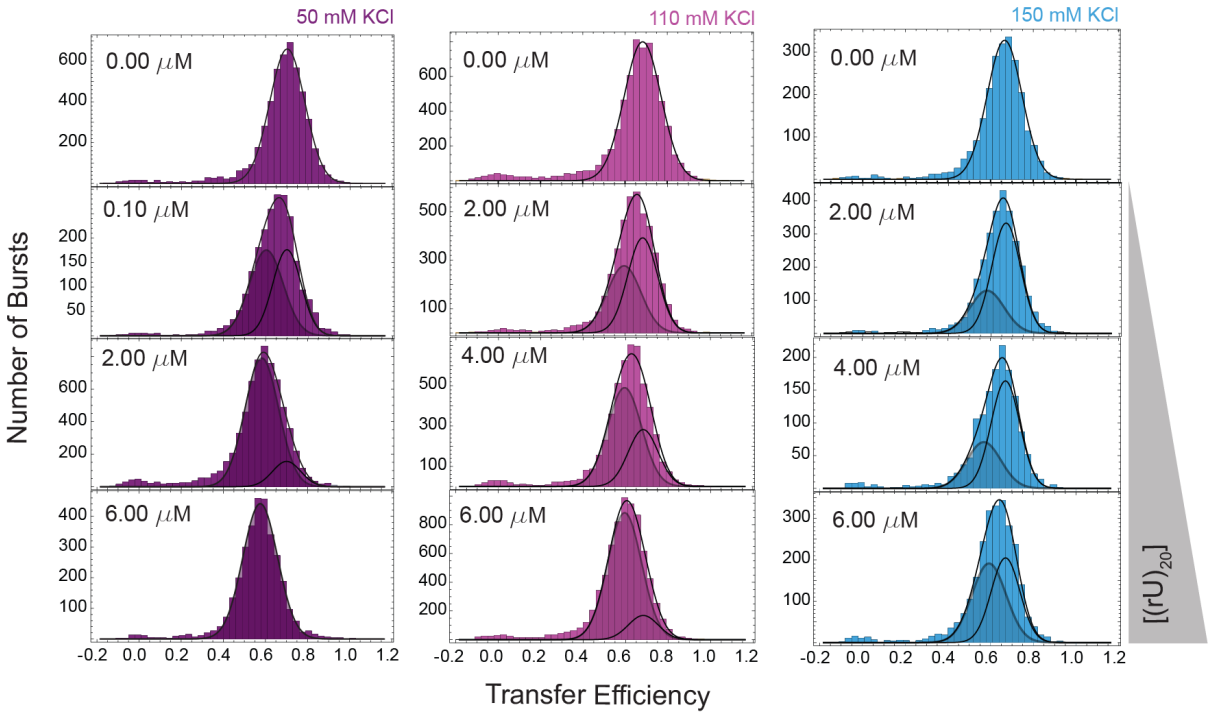


Fig S6. Representative transfer efficiency distributions of $(rU)_{20}$ as a function of salt concentration

Histograms of transfer efficiencies measured at 50 mM KCl (left, purple), 110 mM KCl (center, magenta), 150 mM KCl (right, blue) from 0 to 6 μM $(rU)_{20}$.

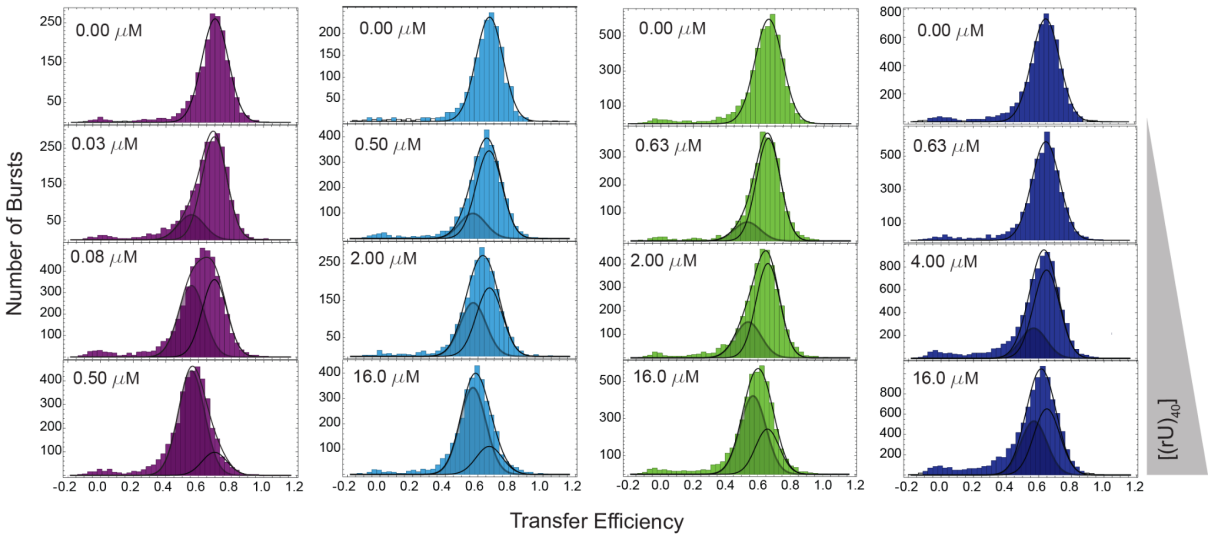


Fig. S7 Representative transfer efficiency distributions of $(rU)_{40}$ as a function of salt concentration

Histograms of transfer efficiencies measured at 50 mM KCl (left, purple), 110 mM KCl (center, magenta), 150 mM KCl (right, blue) from 0 to 16 μM $(rU)_{40}$. Distributions are fitted with up to two Gaussian distributions to quantify the fraction bound and unbound and the corresponding transfer efficiencies.

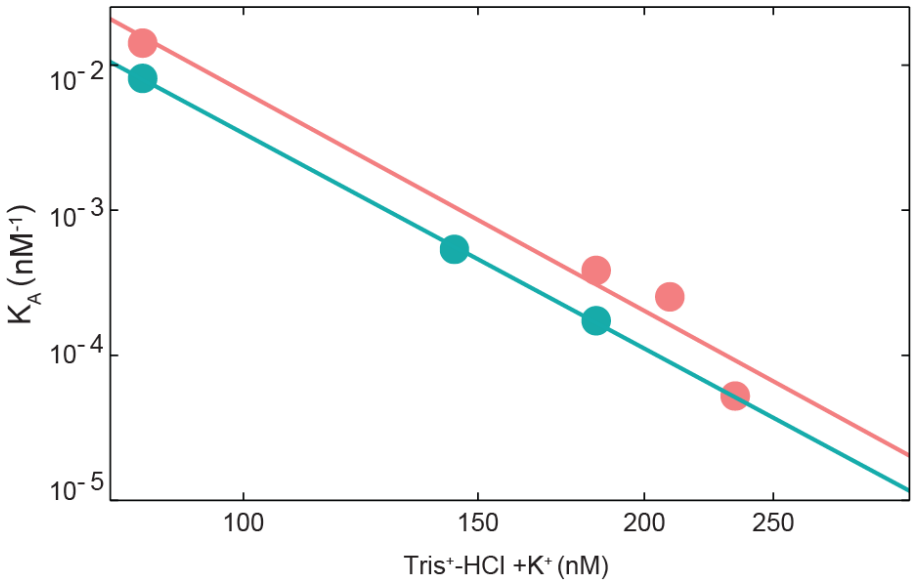


Fig. S8 Association constant as a function of the total concentration of positive ions for $(rU)_{20}$ (cyan) and $(rU)_{40}$ (pink)

Errors associated with each K_A are standard errors of the fit and are reported in **Supplementary Table 12** (not visible because smaller than the marker for the experimental point).

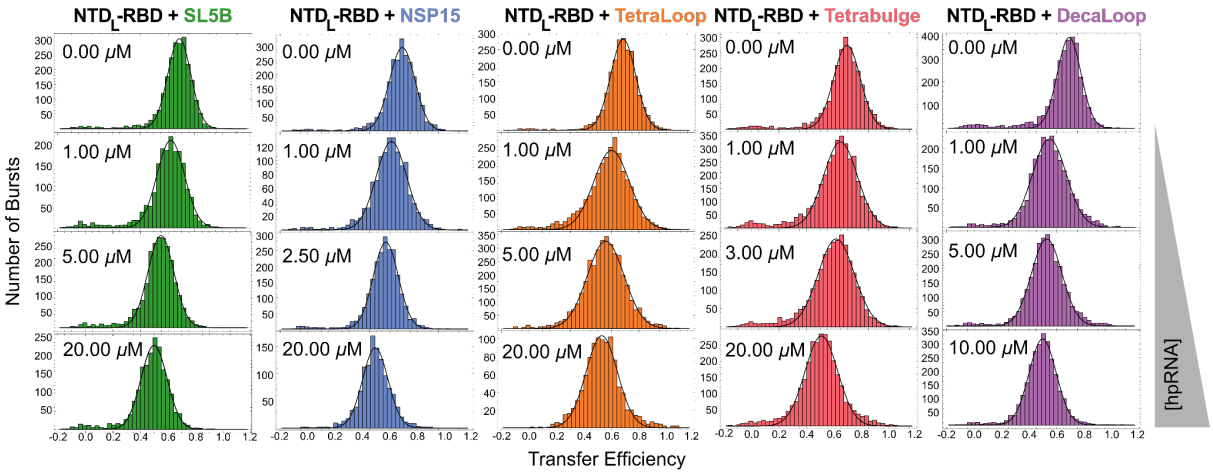


Fig. S9 Transfer efficiency distributions for NTD$_L$-RBD and RNA hairpins

Representative histograms of NTD$_L$-RBD + hairpin RNA (hpRNA) as a function of concentration .

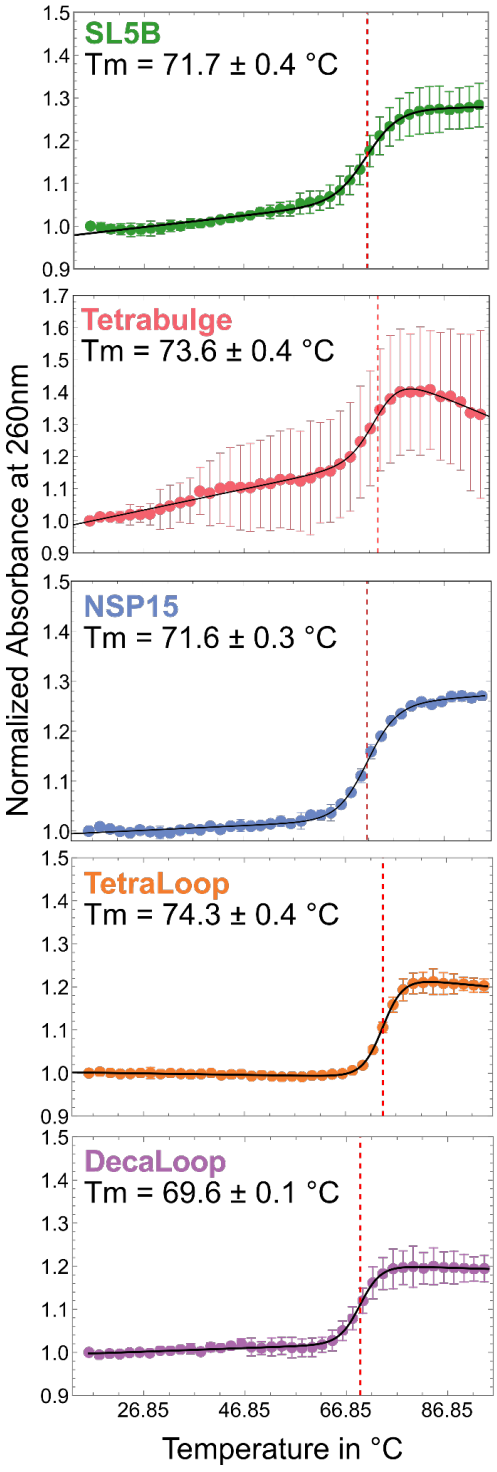

SL5B
Tm = 71.7 ± 0.4 °C
Tetrabulge
Tm = 73.6 ± 0.4 °C
NSP15
Tm = 71.6 ± 0.3 °C
TetraLoop
Tm = 74.3 ± 0.4 °C
DecaLoop
Tm = 69.6 ± 0.1 °C
Normalized Absorbance at 260nm
Temperature in °C
26.85
46.85
66.85
86.85

Fig. S10 Thermal melting curves of RNA hairpins

Absorbance at 260 nm was monitored over a temperature range of 16 °C to 95 °C in 10 mM HEPES, 50 mM KCl, 0.5 mM EDTA, pH 7.4 (23 °C). Temperature was increased in 2 °C steps at a rate of 1 °C/minute and data collected for 2 s after equilibration for 2 minutes after each step. Dots and error bars represent the mean and standard error of 2 measurements performed on different samples. Black solid lines are simulations of **Eq. S20** fitted to the data by least squares nonlinear regression. The best fit value plus/minus the standard error of the fit for the T_m is shown in the plots.

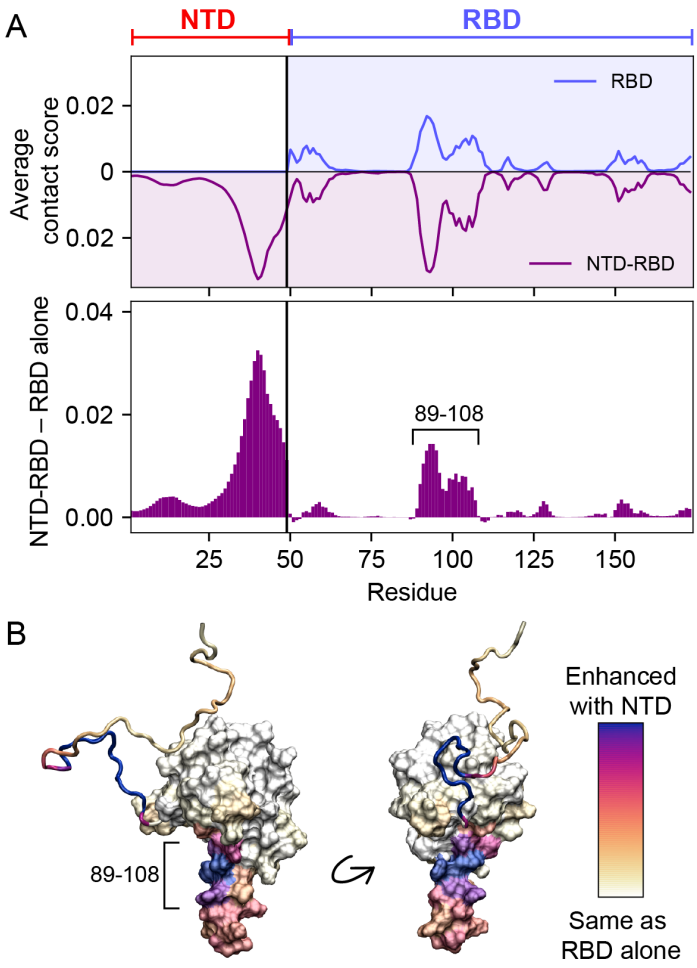


Fig. S11 The NTD does not alter the overall pattern of RBD:RNA interactions

A. To easily compare RBD:RNA interactions with and without the NTD, we calculated the average per-residue contact score for NTD-RBD + $(rU)_{10}$ and RBD + $(rU)_{10}$. Specifically, this involved averaging the per-residue contact fraction over the ten nucleotides to give a per-nucleotide interaction score (which we define as the average contact score). The scores for RBD alone vs. NTD-RBD are shown above. The profiles effectively mirror one another, even down to fine detail, supporting the notion that in our simulations, the addition of the NTD does not alter which residues on the RBD RNA interact with. However, the frequency with which specific sub-regions interact with RNA does change upon the addition of the NTD. Notably, by

comparing the difference in average scores (i.e., NTD-RBD – RBD, bottom panel), residues 89 – 108 within the RBD show an uptick in RNA contacts. **B.** We annotated a structural model of the NTD-RBD by coloring residues according to their enhanced RNA interaction in the presence of the NTD (i.e., scores shown in the bottom panel of panel A). This annotation clearly shows residues in the β-extension dominate in terms of the NTD-enhanced RNA binding.

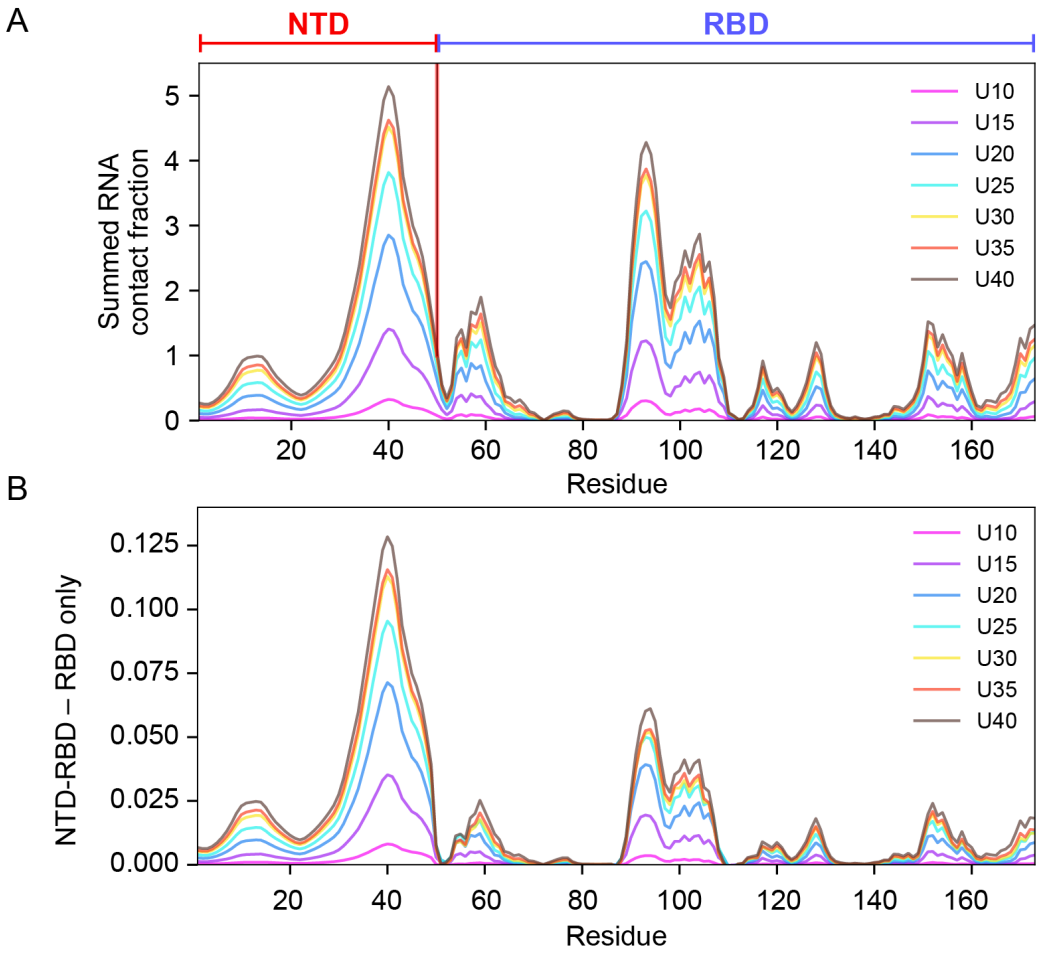


Fig. S12 RNA length tunes the magnitude of protein:RNA interactions but does not alter the overall pattern of RBD:RNA interactions

A. Following the analysis in **Supplementary Fig. 11A**, we calculated the summed contact fraction for each residue across $(rU)_{10}$, $(rU)_{15}$, $(rU)_{20}$, $(rU)_{25}$, $(rU)_{30}$, $(rU)_{35}$, and $(rU)_{40}$. By comparing these profiles, our analysis reveals that as the rU becomes longer, the regions identified in our initial analysis (residue 30–50 and residues 89 – 109) show an RNA-length-dependent enhancement in protein:RNA contacts, supporting the interpretation that these two regions are the primary determinants of protein:RNA interaction. Outside of these regions, additional loci on both the NTD and RBD also engage with RNA in an RNA-length-dependent

manner. In all cases, contacts observed in NTD-RBD:$(rU)_{10}$ simulations (**Supplementary Fig. 11A**) were enhanced as a function of RNA length, but we did not observe novel interactions appear with longer RNA molecules. **B.** We assessed how the presence of the NTD altered RBD:RNA interaction by subtracting RBD contact fractions from NTD-RBD contact fractions across the same six RNA lengths. This analysis confirmed conclusions drawn for using $(rU)_{10}$ – that the major subregion within the RBD that is influenced by the presence of the NTD is the positively-charged β-extension (specifically in residues 89–108) (**Supplementary Fig. 11B**).

6.19 Supplementary Tables

1	MSDNGPQNQR	NAPRITFGGP	SDSTGSNQNG	ERSGARSKQR
RPQGLPNNTA				
51	SWFTALTQHG	KEDLKFPRGQ	GVPINTNSSP	DDQIGYYRRA
TRRIRGGDGK				
101	MKDLSPRWYF	YYLGTGPEAG	LPYGANKDGI	IWVATEGALN
TPKDHIGTRN				
151	PANNAAIVLQ LPQGTTLPKG FYA			

Supplementary Table 1. Sequence of wild type NTD-RBD. Labeling positions are reported in red.

Name	Sequence	Start Position (WT)	End Position (WT)	Labeling Positions (WT)
NTD_L-RBD	GPCSDNGPQNQRNAPRITFGGP SDSTGSNQNGERSGARSKQRRP QGLPNNTASWFTALTQHGKED LKFPCGQGVPINTNSSPDDQIG YYRRATRRIRGGDGKMKDLSPR WYFYYLGTGPEAGLPYGANKD GIIWVATEGALNTPKDHIGTRN PANNAAIVLQLPQGTTLPKGFY A	**1**	**173**	**1, 68**

NTD$_L$-RBD Omicron (P13L, Δ31-33)	GPCSDNGPQNQRNALRITFGGP SDSTGSNQNGGARSKQRRPQG LPNNTASWFTALTQHGKEDLK FPCGQGVPINTNSSPDDQIGYY RRATRRIRGGDGKMKDLSPRW YFYYLGTGPEAGLPYGANKDGI IWVATEGALNTPKDHIGTRNPA NNAAIVLQLPQGTTLPKGFYA	**1**	**173**	**1, 68**
NTD-RBD$_L$	GPMSDNGPQNQRNAPRITFGG PSDSTGSNQNGERSGARSKQRR PQGLPNNTASWFTALTQHGKE DLKFPCGQGVPINTNSSPDDQI GYYRRATRRIRGGDGKMKDLS PRWYFYYLGTGPEAGLPYGAN KDGIIWVATEGALNTPKDHIGT RNPANNAAIVLQLPQGTTLPKG FCA	**1**	**173**	**68,172**
RBD$_L$	GPGLPNNTASWFTALTQHGKE DLKFPCGQGVPINTNSSPDDQI GYYRRATRRIRGGDGKMKDLS PRWYFYYLGTGPEAGLPYGAN KDGIIWVATEGALNTPKDHIGT	**44**	**173**	**68,172**

	RNPANNAAIVLQLPQGTTLPKG FCA			

Supplementary Table 2. Constructs used in this study

For each construct, we reported the start and end positions compared to the wild type (WT) sequence, the labeling positions, and highlighted in yellow the portion of the sequence in between the labeling positions.

RNA	**Genomic position**	**Nts**	**Sequence 5'-3'**	**Origin**
Poly(rU)	**-**	<250		Midland Certified Reagent Company

RNA	Genomic position	Nts	Sequence 5'-3'	Origin
Poly(rU)	-	10,12,15, 17,20,25, 30,35,40		IDT, Horizon Discovery
V21	127-148	21	**UAUAAUUAAUAAC UAAUUACU**	IDT, Horizon Discovery
SL5B*	228 - 252	30	**GGGCAUACCUAGG UUUCGUCCGGGU GUGCC**	*in vitro transcribed*
NSP15	19972-20000	31	**GGGCUCACUGUC UUUUUUGAUGGU AGAGUCC**	*in vitro transcribed*
Tetraloop	based on NSP15	29	**GGGCUCACUGUC UUCGGAUGGUGA GCUC**	*in vitro transcribed*
Decaloop	based on NSP15	34	**GGGCUCACUGUC UUCUUUUUUUGA UGGUGAGCUC**	*in vitro transcribed*
Tetrabulg	based	30	**GGGCUCACUGUC**	*in vitro transcribed*

RNA	Genomic position	Nts	Sequence 5'-3'	Origin
e	on NSP15		**UUCGGAUGGUAG AGUCC**	

*The SL5B sequence is taken from the SARS-CoV genome and differs for two nucleotides from the SARS-CoV-2 genome.

Supplementary Table 3. RNA sequences used in this study

Simulation components	Box size (nm^3)	Temp (K)	Steps per simulation (millions)	Number of independent replicas	Total production frames
NTD	30	298	300	30 (5x6 starting conf)	268,650
RBD	30	298	300	30 (5x6 starting conf)	268,420
NTD-RBD	30	298	300	25 (5x5 starting conf)	221,592
NTD + $(rU)_{25}$	30	298	300	30 (5x6 starting conf)	268,650
RBD + $(rU)_{10}$	30	298	300	30 (5x6 starting conf)	268,650
RBD + $(rU)_{12}$	30	298	300	30 (5x6 starting conf)	268,650
RBD + $(rU)_{15}$	30	298	300	30 (5x6 starting conf)	264,405

RBD + $(rU)_{17}$	30	298	300	30 (5x6 starting conf)	268,322
RBD + $(rU)_{20}$	30	298	300	30 (5x6 starting conf)	267,693
RBD + $(rU)_{25}$	30	298	300	30 (5x6 starting conf)	265,683
RBD + $(rU)_{30}$	30	298	300	30 (5x6 starting conf)	265,026
RBD + $(rU)_{35}$	30	298	300	30 (5x6 starting conf)	268,650
RBD + $(rU)_{40}$	30	298	300	30 (5x6 starting conf)	263,567
RBD + $(rU)_{180}$	30	298	300	30 (5x6 starting conf)	268,503
NTD-RBD + $(rU)_{10}$	30	298	300	30 (5x6 starting conf)	268,650
NTD-RBD + $(rU)_{12}$	30	298	300	30 (5x6 starting conf)	268,650
NTD-RBD + $(rU)_{15}$	30	298	300	30 (5x6 starting conf)	267,563
NTD-RBD + $(rU)_{17}$	30	298	300	30 (5x6 starting conf)	267,341
NTD-RBD + $(rU)_{20}$	30	298	300	30 (5x6 starting conf)	265,965
NTD-RBD +	30	298	300	30 (5x6 starting conf)	263,434

$(rU)_{25}$					
NTD-RBD + $(rU)_{30}$	30	298	300	30 (5x6 starting conf)	268,134
NTD-RBD + $(rU)_{35}$	30	298	300	30 (5x6 starting conf)	268,650
NTD-RBD + $(rU)_{40}$	30	298	300	30 (5x6 starting conf)	267,220
NTD-RBD + $(rU)_{180}$	30	298	300	30 (5x6 starting conf)	268,650
OmNTD-RBD (P13L,Δ^{31-33}) + $(rU)_{25}$	30	298	300	30 (5x6 starting conf)	267,333
OmNTD-RBD (P13L) + $(rU)_{25}$	30	298	300	30 (5x6 starting conf)	268,527
OmNTD-RBD (Δ^{31-33}) + $(rU)_{25}$	30	298	300	30 (5x6 starting conf)	266,023

Supplementary Table 4. Summary of simulation details

	RBD_L full-length protein	RBD_L
$c_{1/2,IF}$ (M)	1.68 ± 0.02	1.26 ± 0.03
$c_{1/2,IF}$ (M)	1.64 ± 0.02	1.21 ± 0.01
ΔG_{UI} (kcal $mol^{-1}M^{-1}$)	6.6 ± 0.5	2.8 ± 0.1
ΔG_{IF} (kcal $mol^{-1}M^{-1}$)	8.1 ± 0.5	7.6 ± 0.4

Supplementary Table 5. RBD Folding parameters

	K_{int} for poly(rU) (μM^{-1}) nucleotides
RBD_L	$(6 \pm 1) \times 10^{-2}$
NTD-RBD_L	2.0 ± 0.4
NTD_L-RBD	4.0 ± 0.3

Supplementary Table 6. Intrinsic association constants

	K_A (μM^{-1}) molecules	
n	RBD_L	NTD_L-RBD
10	$(4 \pm 3) \times 10^{-2}$	$(3.8 \pm 0.1) \times 10^{-1}$
12	$(7.9 \pm 0.4) \times 10^{-2}$	$(6.1 \pm 1.5) \times 10^{-1}$
15	$(1.1 \pm 0.6) \times 10^{-1}$	1.4 ± 0.1
17	$(1.9 \pm 0.4) \times 10^{-1}$	3.5 ± 0.8
20	$(2.6 \pm 0.7) \times 10^{-1}$	4.6 ± 0.6
25	$(6.0 \pm 0.9) \times 10^{-1}$	20 ± 5
30	$(6.6 \pm 1.3) \times 10^{-1}$	47 ± 6

35	-	66 ± 5
40	1.2 ± 0.3	85 ± 7

Supplementary Table 7. RBD_L and NTD_L-RBD association constants for $(rU)_n$ as measured by single-molecule FRET

Construct	**NTD;** K_A (μM^{-1})	**RBD;** K_A (μM^{-1})	**NTD-RBD;** K_A (μM^{-1})
$(rU)_{10}$	$(9.5 \pm 0.7) \times 10^{-5}$	$(4.6 \pm 0.4) \times 10^{-4}$	$(1.90 \pm 0.05) \times 10^{-3}$
$(rU)_{12}$	$(2.5 \pm 0.2) \times 10^{-4}$	$(8.7 \pm 0.8) \times 10^{-4}$	$(4.2 \pm 0.1) \times 10^{-3}$
$(rU)_{15}$	$(5.6 \pm 0.4) \times 10^{-4}$	$(1.7 \pm 0.1) \times 10^{-3}$	$(1.00 \pm 0.03) \times 10^{-2}$
$(rU)_{17}$	$(8.0 \pm 0.5) \times 10^{-4}$	$(2.5 \pm 0.2) \times 10^{-3}$	$(1.8 \pm 0.1) \times 10^{-2}$
$(rU)_{20}$	$(1.10 \pm 0.04) \times 10^{-3}$	$(3.5 \pm 0.1) \times 10^{-3}$	$(4.3 \pm 0.7) \times 10^{-2}$
$(rU)_{25}$	$(2.00 \pm 0.08) \times 10^{-3}$	$(5.2 \pm 0.1) \times 10^{-3}$	$(1.1 \pm 0.1) \times 10^{-1}$

$(rU)_{30}$	$(2.7 \pm 0.1) \times 10^{-3}$	$(8.1 \pm 0.8) \times 10^{-3}$	$(2.1 \pm 0.2) \times 10^{-1}$
$(rU)_{35}$	$(4.1 \pm 0.2) \times 10^{-3}$	$(1.1 \pm 0.1) \times 10^{-2}$	$(4.0 \pm 1.0) \times 10^{-1}$
$(rU)_{40}$	$(4.7 \pm 0.1) \times 10^{-3}$	$(1.30 \pm 0.06) \times 10^{-2}$	$(6.9 \pm 0.7) \times 10^{-1}$
$(rU)_{180}$	1.20 ± 0.02	3.2 ± 0.6	$(1.0 \pm 0.2) \times 10^{2}$

Supplementary Table 8. Simulation-derived association constants (K_A) in μM^{-1}

Construct	**NTD;** K_D (μM)	**RBD;** K_D (μM)	**NTD-RBD;** K_D (μM)
$(rU)_{10}$	$(1.10 \pm 0.08) \times 10^{4}$	2200 ± 200	540 ± 10
$(rU)_{12}$	$(4.0 \pm 0.3) \times 10^{3}$	1200 ± 100	240 ± 9
$(rU)_{15}$	$(1.8 \pm 0.1) \times 10^{3}$	570 ± 30	98 ± 3
$(rU)_{17}$	$(1.20 \pm 0.08) \times 10^{3}$	400 ± 30	56 ± 4

$(rU)_{20}$	$(9.0 \pm 0.4) \times 10^2$	290 ± 10	24 ± 3
$(rU)_{25}$	$(5.0 \pm 0.2) \times 10^2$	192 ± 5	9.0 ± 0.6
$(rU)_{30}$	$(4.0 \pm 0.1) \times 10^2$	120 ± 10	4.7 ± 0.5
$(rU)_{35}$	$(2.4 \pm 0.1) \times 10^2$	92 ± 9	2.7 ± 0.8
$(rU)_{40}$	$(2.10 \pm 0.05) \times 10^2$	80 ± 4	1.5 ± 0.1
$(rU)_{180}$	0.80 ± 0.02	0.30 ± 0.07	$(1.0 \pm 0.1) \times 10^{-2}$

Supplementary Table 9. Simulation-derived dissociation constants (K_D) in μM

Construct	**NTD;** K_A*	**RBD;** K_A*	**NTD-RBD;** K_A*
$(rU)_{10}$	$(8.5 \pm 0.9) \times 10^{-4}$	$(4.2 \pm 0.5) \times 10^{-3}$	$(2.0 \pm 0.1) \times 10^{-2}$
$(rU)_{12}$	$(2.3 \pm 0.2) \times 10^{-3}$	$(7.8 \pm 0.9) \times 10^{-3}$	$(4.0 \pm 0.3) \times 10^{-2}$
$(rU)_{15}$	$(5.0 \pm 0.5) \times 10^{-3}$	$(1.6 \pm 0.1) \times 10^{-2}$	$(9.0 \pm 0.7) \times 10^{-2}$

$(rU)_{17}$	$(7.2 \pm 0.7) \times 10^{-3}$	$(2.2 \pm 0.2) \times 10^{-2}$	$(1.6 \pm 0.2) \times 10^{-1}$
$(rU)_{20}$	$(1.0 \pm 0.1) \times 10^{-2}$	$(3.1 \pm 0.2) \times 10^{-2}$	$(3.9 \pm 0.7) \times 10^{-1}$
$(rU)_{25}$	$(1.8 \pm 0.1) \times 10^{-2}$	$(4.7 \pm 0.3) \times 10^{-2}$	1.0 ± 0.1
$(rU)_{30}$	$(2.4 \pm 0.2) \times 10^{-2}$	$(7.3 \pm 0.9) \times 10^{-2}$	1.9 ± 0.2
$(rU)_{35}$	$(3.7 \pm 0.3) \times 10^{-2}$	$(9.9 \pm 1.3) \times 10^{-2}$	3.6 ± 0.9
$(rU)_{40}$	$(4.2 \pm 0.3) \times 10^{-2}$	$(1.10 \pm 0.09) \times 10^{-1}$	6.2 ± 0.7
$(rU)_{180}$	10.7 ± 0.7	28 ± 6	900 ± 200

Supplementary Table 10. Simulation-derived ratio of association constants K_A* defined as (K_A of Construct + $(rU)_n$)/(K_A of NTD-RBD + $(rU)_{25}$)

	α (KCl)	**α (KCl+Tris HCl)**
$(rU)_{20}$	-3.49 ± 0.05	-5.0 ± 0.1

$(rU)_{40}$	-3.7 ± 0.5	-5.0 ± 0.7

Supplementary Table 11. Ion released upon binding of $(rU)_{20}$ and $(rU)_{40}$ (compare with Fig. 6 and Supplementary Fig. 5)

	K_A (μM^{-1})				
	50 mM KCl	**110 mM KCl**	**150 mM KCl**	**175 mM KCl**	**200 mM KCl**
$(rU)_{20}$	8 ± 2	$(5.4 \pm 0.6) \times 10^{-1}$	$(1.7 \pm 0.2) \times 10^{-1}$	-	-
$(rU)_{40}$	14 ± 2	-	$(3.8 \pm 0.4) \times 10^{-1}$	$(2.5 \pm 0.3) \times 10^{-1}$	$(5 \pm 0.9) \times 10^{-2}$
	K_D (μM)				
	50 mM KCl	**110 mM KCl**	**150 mM KCl**	**175 mM KCl**	**200 mM KCl**
$(rU)_{20}$	0.12 ± 0.03	1.9 ± 0.2	5.8 ± 0.5	-	-
$(rU)_{40}$	0.07 ± 0.01	-	2.6 ± 0.2	4.0 ± 0.4	19 ± 4

Supplementary Table 12. Association and dissociation constants of NTD_L-RBD as a function of salt concentration for $(rU)_{20}$ and $(rU)_{40}$

	K_A (μM^{-1}) molecules
K_{A1}	6.2 ± 0.3
K_{A2}	0.15 ± 0.10

Supplementary Table 13. NTD_L-RBD association constants for V21 binding

	K_A (μM^{-1}) molecules
hpRNA	NTD_L-RBD
NSP15	$(7.8 \pm 0.7) \times 10^{-1}$
Tetraloop	$(6.7 \pm 0.8) \times 10^{-1}$
Tetrabulge	$(3.4 \pm 0.7) \times 10^{-1}$
Decaloop	3.4 ± 0.5
SL5B	$(5.3 \pm 0.4) \times 10^{-1}$

Supplementary Table 14. NTD_L-RBD association constants for hairpin RNA sequences

Construct	**OmNTD-RBD (**P13L,Δ^{31-33}**);** K_A*	**NTD-RBD (**P13L**);** K_A*	**NTD-RBD (**Δ^{31-31}**);** K_A*
$(rU)_{25}$	$(7 \pm 2) \times 10^{-1}$	1.0 ± 0.2	$(7 \pm 1) \times 10^{-1}$

Supplementary Table 15. Ka* defined as (K_A of Construct + $(rU)_n$)/(K_A of NTD-RBD + $(rU)_{25}$)

Chapter 7: Conserved molecular recognition by an intrinsically disordered region in the absence of sequence conservation

This chapter was published on Bioarxiv and is submitted to a peer reviewed journal as:
Conserved molecular recognition by an intrinsically disordered region in the absence of sequence conservation. Jhullian Alston, Andrea Soranno, Alex Holehouse. *bio*R*xiv* 2023.08.06.552128; doi: https://doi.org/10.1101/2023.08.06.552128

Contributions. J.A. performed simulations and analysis. J.A. and A.S.H. developed computational tools for simulations. J.A, A.S., and A.H. wrote the paper. A.H. supervised simulations and data analysis. J.A., A.S., and A.H. conceived the experiments.

7.1 Abstract

Intrinsically disordered regions (IDRs) are critical for cellular function, yet often appear to lack sequence conservation when assessed by multiple sequence alignments. This raises the question of if and how function can be encoded and preserved in these regions despite massive sequence variation. To address this question, we have applied coarse-grained molecular dynamics simulations to investigate non-specific RNA binding of coronavirus nucleocapsid proteins. Coronavirus nucleocapsid proteins consist of multiple interspersed disordered and folded domains that bind RNA. We focussed here on the first two domains of coronavirus nucleocapsid proteins, the disordered N-terminal domain (NTD) followed by the folded RNA binding domain (RBD). While the NTD is highly variable across evolution, the RBD is structurally conserved. This combination makes the NTD-RBD a convenient model system to explore the interplay between an IDR adjacent to a folded domain, and how changes in IDR sequence can influence molecular recognition of a partner. Our results reveal a surprising degree of sequence-specificity encoded by both the composition and the precise order of the amino acids in the NTD. The presence of an NTD can – depending on the sequence – either suppress or enhance RNA binding. Despite this sensitivity, large-scale variation in NTD sequences is possible while certain sequence features are retained. Consequently, a conformationally-conserved fuzzy RNA:protein complex is found across nucleocapsid protein orthologs, despite large-scale changes in both NTD sequence and RBD surface chemistry. Taken together, these insights shed light on the ability of disordered regions to preserve functional characteristics despite their sequence variability.

7.2 Significance Statement

Intrinsically disordered regions (IDRs) are ubiquitous across the kingdoms of life, yet many fundamental questions regarding their functions remain. In particular, understanding if and how IDRs retain conserved behavior despite sequence variations is a major open question. Leveraging molecular simulations, we explored how the coronavirus nucleocapsid protein, a disordered RNA binding protein, retains its biological function despite large sequence variation in its disordered regions. We uncover a relationship between sequence composition at specific sites within nucleocapsid protein orthologs and their ability to bind to single-stranded RNA. Our findings suggest that IDRs can exhibit conserved interaction modes, despite lacking exact sequence conservation. This study reveals the need to explore beyond direct sequence alignment to understand the sequence-ensemble-function relationship in disordered proteins.

7.3 Introduction

The classical structure-function paradigm states that sequence dictates structure, and structure dictates function[779]. This understanding has driven extensive study of protein structure and dynamics. Understanding the 3D structures that proteins adopt provides insight into their normal function. It also allows us to interpret how and why mutations that disrupt those structures and/or dynamics impair function[780-782]. However, in recent years, there has been a growing focus on understanding "unstructured" or disordered protein regions[783-786]. Intrinsically disordered regions (IDRs) are poorly described by a single 3D structure; instead, they exist as a collection of structurally distinct interconverting conformations known as an ensemble[68,573]. Despite lacking a defined 3D structure, IDRs play critical roles in many aspects of cellular function. Consequently, emerging work suggests that just as folded domains follow a sequence-

structure-function relationship, IDRs can follow an analogous sequence-ensemble-function relationship[74]. Given the importance that structure-function analysis has played in understanding the molecular basis for cellular function, there is a promising and analogous opportunity to understand IDR function through the lens of ensembles[78, 80, 618, 787, 788].

A major goal of modern molecular biology is to accurately predict protein function directly from amino acid sequence. Rooted in the general assumption that similar protein sequences will exhibit similar molecular behavior, one strategy is to compare the sequence of a protein of interest to those of other known proteins[789-792]. In many cases, multiple sequence alignment of orthologous folded domains reveals high sequence conservation and, therefore, conserved protein function[789, 793,794]. This relationship enables us to predict structures of previously unsolved protein structures and infer function by aligning the sequences of an uncharacterized protein against sequences of functionally-characterized folded domains[32, 795, 796]. In sum, applying evolutionary information, directly and indirectly, is a central pillar in our modern toolkit for protein sequence analysis.

While IDR sequences can be aligned, their conservation at the residue level is typically lower than their structured counterparts[454, 797, 798]. However, even without strict sequence conservation, the presence of disorder in a given protein domain is often conserved across orthologs[29, 30, 80, 618, 798, 799]. Assuming orthologous proteins provide equivalent functions, this presents an intriguing question: "Can apparently divergent IDRs confer the same molecular functions?". For some IDRs, the only feature that matters may be the existence of Short Linear Motifs (SLiMs), such that a large IDR may appear poorly conserved, yet functional conservation is maintained as long

as a few short (5-15 residue) regions are present[800-802]. Recent work suggests that retaining specific physicochemical properties in a disordered region is sufficient to preserve function[80, 92, 414, 797, 798, 803-805]. Some of the conformations disordered proteins may adopt can be structured. This transient structure formation can underlie conservation in some IDRs, where specific interactions are needed to maintain proper folds for protein-protein and protein-nucleic acid interactions[78, 806]. Ultimately, the absence of a specific 3D structure loosens the relationship between sequence and function.

Viruses provide good test systems for exploring evolutionary conservation in IDRs. Viruses use IDRs extensively, and their rapid evolutionary rates – driven by a combination of fast replication times, massive numbers, and strong fitness selection – mean that even between serotypes of the same virus, substantial divergence in IDRs is often observed[807-812]. For viruses that infect the same host, it is reasonable to expect equivalent selective pressures and equivalent protein function. Taken together, viral IDRs offer a convenient opportunity to explore how large-scale variation in IDR sequence enables similar functional output.

In this work, we investigated the relationship between IDR sequence and RNA interaction by performing coarse-grained molecular dynamics simulations of coronavirus nucleocapsid (N) proteins[638, 813]. Coronaviruses are positive-sense single-stranded RNA viruses with relatively large (~30 Kb) genomes[632, 813-815]. They typically consist of 4 major structural proteins: spike (S), envelope (E), membrane (M), and the N protein. The N protein is the most abundant viral protein and drives genomic RNA condensation and packaging during virion assembly, but has also been implicated in the evasion of the host immune system[187, 464, 780, 816]. Given its abundance

and importance, the N protein is a tractable model system for exploring variation in sequence and function.

Coronavirus N proteins consist of five domains; two folded domains and three IDRs (Fig. 1A)[694]. Our prior work systematically characterized full-length SARS-CoV-2 N protein using a combination of all-atom simulations, single-molecule Förster Resonance Energy Transfer (smFRET) spectroscopy, and nanosecond Fluorescence Correlation Spectroscopy (ns-FCS)[187]. This work confirmed the disordered nature of the three IDRs and characterized their ensemble behavior in the context of the full-length protein. The two N-terminal domains (the N-terminal domain, NTD, and RNA-binding domain, RBD) are disordered and folded, respectively. In addition to characterization in the absence of RNA, our more recent experimental and computational work showed that these domains work together to enable high-affinity RNA binding[819]. While the RBD alone binds $(rU)_{25}$ with a binding affinity of ~0.6 μM-1, the addition of the NTD enhances this affinity around 30-fold. This enhancement in binding affinity is facilitated by a fuzzy complex that forms between the NTD-RBD and RNA, where the NTD remains fully disordered in the bound and unbound states. While we cannot exclude other potential roles for the NTD, our work to date suggests that one of its functions is to enhance N-protein:RNA interactions, presumably to facilitate genome packaging.

In this work, we focussed on the NTD and RBD as a model system for understanding the sequence constraints on molecular function. While the NTD sequence is variable across N protein orthologs, the presence of a disordered NTD is highly conserved in coronaviruses (Fig. 1B)[187]. In contrast, the RBD is highly conserved among orthologs, and, despite some variation in

sequence leading to changes in surface chemistry, it harbors a nearly identical fold across experimentally resolved and computationally predicted structures (Fig. 1C).

Given the structurally similar RBDs but differing NTDs, we wondered whether different coronavirus NTD-RBDs bind single-stranded RNA (ssRNA) in the same way, or whether they have distinct modes of interaction. Naively, given the large variation in NTD sequence, one might expect fundamentally different modes of recognition. However, recent work has shown that the conservation of IDR ensemble properties is possible despite large changes in IDR sequence[618].

To address this question, we performed coarse-grained molecular dynamics (MD) simulations of NTD-RBD constructs with poly-$(rU)_{25}$ to assess how changes in NTD sequence influence molecular function, i.e., RNA binding. Using this approach, we sought to understand how the sequence properties of an RNA binding domain and flanking disordered region enable them to cooperate to bind nucleic acids and achieve specific binding affinities. Our findings demonstrate that the ability of the SARS-CoV-2 nucleocapsid protein NTD to bind ssRNA is determined by a combination of sequence composition and the specific positioning of positively charged amino acids within its linear sequence. We identified critical 'hotspots' of protein-nucleic acid interaction in the SARS-CoV-2 NTD, where maintaining positive charge allows retention of wild-type binding affinity. These 'hotspots' result in a distinctive mode of ssRNA binding in the SARS-CoV-2 NTD-RBD complex, which we observed to be conserved across several coronavirus orthologs, despite significant variations in the NTD sequence. Our study highlights

that disordered regions can exhibit conserved interaction mechanisms, even in the absence of exact sequence conservation.

7.4 Methods

Molecular Dynamics Simulations

All simulations were performed using the LAMMPS simulation engine[817]. We performed molecular dynamics simulations in the NVT ensemble using the default parameters of the physics-driven coarse-grained force-field Mpipi developed by Joseph et al. [405] The model represents both amino acid residues and nucleotides as chemically unique singular beads and was parameterized to recapitulate the behavior of disordered proteins in isolation as well as their ability to undergo phase separation with and without RNA. Inter-bead interactions consist of a combination of short-range contributions from a Wang-Frenkel potential, which captures a combination of Van der Waals, cation-pi, and pi-pi interactions, and a long-range Coulombic potential for amino acids with net charge and RNA nucleotides. The ability of the Mpipi force field to recapitulate disordered protein dimensions has been previously shown [46,405]. Simulations were performed under an effective ionic strength of 50 mM NaCl, conditions we previously found to engender good agreement between simulation and experiment when comparing with experimentally-measured RNA binding affinities using single-molecule experiments[819].

We also assessed the ability of the Mpipi forcefield to recapitulate single-stranded RNA (ssRNA) dimensions by comparing simulations of $(rU)_{40}$ with scattering data from small-angle X-ray (SAXS) experiments for the same construct[818]. This comparison revealed excellent agreement across the full scattering curve and in terms of the scattering-derived radius of gyration; using the

Molecular Form Factor approach of Riback et al., R_g^{sim} = 30.9 ± 0.1 Å while R_g^{exp} = 30.2 ± 0.3 Å) (**Supplementary Fig. 1**) [308].

Simulations were performed in a 30 nm^3 simulation box with periodic boundary conditions. Protein and RNA are allowed to diffuse freely throughout the box. Disordered regions and ssRNA behave as dynamic flexible polymers, sampling an ensemble of conformations [405]. However, as done previously, folded domains were made rigid, and residues buried within folded domains experienced downscaled non-bonded interactions[405, 819]. Unless otherwise specified, all simulations were run for 300 million steps per replicate. The exceptions are the 'scrambled' simulations, which were run for 100 million steps per replicate. Protein and RNA configurations were saved every 10,000 steps, and the first 0.2% was removed for equilibration. Visualization of protein-RNA complexes was done with Protein Imager and VMD [820, 821]. Simulations were analyzed using SOURSOP and MDTraj [584,715]. Small angle X-ray scattering was analyzed using the Molecular Form Factor (MFF) (http://sosnick.uchicago.edu/SAXSonIDPs), while synthetic scattering data for simulations were generated using FOXS default settings[308, 822].

We performed simulations of the NTD-RBD, NTD, and RBD of six coronavirus orthologs. Specifically, we examined five coronaviruses that infect humans: SARS-CoV-2 (SCO2), Middle Eastern Respiratory Syndrome virus (MERS), Human Coronaviruses OC43, Human Coronavirus HKU1, and Human Coronavirus 229E, as well as Murine Hepatitis Virus (MHV1). Sequence alignments were compared to determine a region of the RBD that was well conserved between all orthologs to delineate the start and end positions of the NTD and RBD's of each ortholog [694, 823-825]. For simulations with ssRNA, all simulations were done using $(rU)_{25}$.

To capture conformational heterogeneity in an artificially rigid structure, we utilized Colabfold to generate five different starting structures for each coronavirus orthologous RBD[32,826]. For simulations of wild-type versions of each ortholog's NTD-RBD all five starting structures are used, to enable conclusions to be less biased by a specific starting conformation. As expected, certain RBD conformers bind RNA better than others, but in all cases where different NTDs are compared, the same sets of RBD conformers are used, such that any RBD conformation-specific biases are consistent across the set (**Supplementary Fig. 2**). For the large scrambled library, 1 conformation for the SCO2 RBD is used. All simulations were run with multiple replicates per starting RBD structure, with a minimum of five replicates per RBD conformation.

Limitations of Coarse-Grained Simulations

Our use of the Mpipi model should not be taken to imply that RNA or proteins are faithfully represented at one bead per residue/nucleotide resolution. Both proteins and RNA are complex biomolecules with many degrees of freedom, a chemically heterogeneous structure, and can engage in a variety of sequence and structure-specific interactions that are not captured by a simplified coarse-grain model. Our goal in using a simplified coarse-grain model is to enable high-throughput biophysical assessment in a system that, based on prior work, we have good reason to believe is semi-quantitative in terms of relative accuracy [405,819]. While we refer to the molecules in our simulations as protein and RNA, in reality, they are better thought of as RNA- and protein-flavored polymers. The simplicity of this model enables us to address questions that would be intractable using either higher-resolution simulation approaches or experiments.

Despite this, we are under no illusion regarding the simplifying assumptions made for a coarse-grain model.

Calculating Apparent Association Constants From Simulations

We determined apparent association constants (K_A) by using an updated version of our previous center of mass (COM) calculations that were able to qualitatively recapitulate SARS-CoV-2 NTD-RBD single-stranded RNA binding[819]. To do this, post-equilibration simulation frames were divided into bound and unbound states. This delineation was achieved by first taking the intermolecular center-of-mass distances between the protein and the RNA and plotting the distribution of distances. The histogram of intermolecular distances follows a bimodal distribution that reports on the bound and unbound states, and can be fit with two Gaussians (**Fig. 2C**). We then determined the intersection that minimizes the overlap of the two distributions to define a cutoff distance. The cutoff distance varies based on the size of the protein and RNA. Finally, as done previously, we classify frames as bound or unbound by assessing the linear intermolecular COM distance trajectory and delineating frames as bound when five or more frames are below the cutoff distance. This minimum number of consecutive frames allows us to distinguish between transient random interactions between protein and RNA vs. encounters with a reasonable "lifetime", implying direct and continuous interaction. The distributions and distance cutoffs are calculated for every set of NTD_a-RBD_b + $(rU)_n$ simulations, where *a* and *b* represent specific NTD or RBD sequences and *n* the length of the single-stranded (rU), allowing us to determine protein-RNA specific distance thresholds for each simulation.

The resultant fraction of bound frames is used to calculate an apparent K_D with the equation:

$$K_D = \frac{(1-f_{bound})^2}{N_A V f_{bound}} \quad \text{(Eq. 1)}$$

Here f_{bound} refers to the fraction of frames where the protein and RNA are determined to be in the bound state from our COM-COM distribution analysis. $N_A$refers to Avagodro's constant, and V is the simulation box volume in liters, which returns a K_D in mol/L. K_A is then calculated using the expression $K_A = 1/K_D$. While we determine if two molecules are bound or unbound in a different manner, this approach is analogous to that of Tesei *et al.* [777].

It is important to note that the K_As determined from these simulations are not meant to represent absolute values that would be comparable to those determined from experiment. Our prior work has shown that K_As calculated from Mpipi simulations for this system lack absolute agreement with experimentally measured values. Despite this, when experiment and simulation-derived K_A values are normalized by an internally consistent reference (i.e., the K_A obtained from NTD-RBD binding $(rU)_{25}$), we see good agreement between simulations and experiment, both as a function of RNA length and as a function of the presence/absence of the NTD [819]. To that end, binding affinity here is reported as K_A*, a normalized binding affinity we define as the ratio of the apparent K_A of a given protein + RNA simulation divided by the corresponding K_A for the analogous SCO2 NTD-RBD binding to $(rU)_{25}$. This enables the SCO2 NTD-RBD + $(rU)_{25}$ simulations binding affinity to be a reference point with which to understand the strength of interactions of other orthologs. All K_A* values are thus greater than 1 (stronger binding than the SCO2 NTD-RBD + $(rU)_{25}$) or less than 1 (weaker binding than the SCO2 NTD-RBD + $(rU)_{25}$).

Error is propagated for our ratio (K_A*) using:

$$\frac{R_{error}}{R} = \sqrt{(\frac{A_{error}}{A})^2 + (\frac{B_{error}}{B})^2} \quad \text{(Eq. 2)}$$

R and R_{error} here represent the ratio and the error of the ratio. *A* and *B* represent the numerator and denominator of our ratios, respectively, and A_{error} and B_{error} are their associated errors (standard error of the mean).

Calculating Charge Clustering in Disordered Regions

Charge clustering is quantified by the inverse weighted distance (IWD), a metric that has been applied to study amino acid clustering in several systems [429,827-829]. Unlike the patterning parameters $\varkappa$ ("kappa") or sequence charge decoration (SCD), which quantify the patterning of oppositely charged residues with respect to one another, here our interest is on the clustering of positive residues only [180,319]. The IWD score allows us to quantify the clustering of a specific subset of residues. When residues are clustered together, the IWD score is high, whereas when residues are evenly distributed, the IWD score is low. IWD scores were calculated using sparrow (https://github.com/idptools/sparrow).

Statistical Analysis

Every simulation has a minimum of five independent replicates, and calculated values are presented as 95% confidence intervals (box plots, with medians marked), mean and standard error of the mean, or geometric mean and geometric standard deviation (clarified in text below figures). Fitting of Gaussian distributions was done in Python using scipy.optimize.curve_fit[830].

Data Availability and Software

Analysis code and data (calculated distance distributions and contact map information) are deposited at https://github.com/holehouse-lab/supportingdata/tree/master/2023/alston_2023. For further information on the use of code, please refer to the deposited Jupyter notebooks.

7.5 Results

7.6 "Inert" Intrinsically Disordered Regions Suppress RNA Binding

Our previous work used coarse-grained MD simulations paired with smFRET-based RNA binding experiments to characterize the ability of the SCO2 NTD-RBD to bind ssRNA [819]. Simulations and experiments showed that the addition of the disordered NTD_{SCO2} to the folded RBD resulted in a 30-fold increase in the binding affinity for $(rU)_{25}$ compared to the RBD alone. We hypothesized that specific residues in the NTD_{SCO2} formed favorable interactions with RNA, driving the enhanced binding affinity observed. We further speculated that substituting the NTD_{SCO2} with an inert IDR that interacts negligibly with RNA would result in a binding affinity similar to that of the RBD alone. To our surprise, our simulations showed this was not the case.

In the Mpipi model, glycine and serine residues have negligible interactions with RNA or other amino acids. This is in good agreement with prior experimental work that suggests GS-repeat sequences behave as relatively inert Gaussian chains [311,620]. Therefore, we replaced the NTD_{SCO2} with a length-matched GS repeat – $(GS)_{25}$ – and performed simulations with this $(GS)_{25}$-RBD_{SCO2} chimera (**Fig. 2A**). [32,795,826]. Our simulations revealed repeated association and dissociation events between $(rU)_{25}$ and the $(GS)_{25}$-RBD constructs (**Fig. 2B)**, enabling us to calculate an apparent binding association constant, K_A (see Methods for details).

To our surprise, the $(GS)_{25}$-RBD suppressed RNA binding compared to the RBD alone ($(GS)_{25}$-RBD $K_A^* = (2.0 \pm 0.3) \times 10^{-2}$, whereas RBD $K_A^* = (3.7 \pm 0.4) \times 10^{-2}$)) (**Fig. 2D)**. This result is driven by an entropic effect, whereby the $(GS)_{25}$ impedes the ability of RNA molecules to interact with the RBD. The NTD, in contrast, possesses sequence features that enable direct interaction with RNA, notably residues 30-50, which enhance the macroscopic binding affinity [527,819].

To investigate the contribution of a smaller and targeted inert region, we next replaced the NTD_{SCO2} 30-50 residue region with a $(GS)_{10}$ linker. While we anticipated a decrease in binding affinity, we expected it to still be stronger than that of the RBD alone. In actuality, we again observed a suppression of RNA binding affinity, with the $(GS)_{10}$ demonstrating weaker binding affinity than the RBD alone, with a $K_A^* = (2.1 \pm 0.3) \times 10^{-2}$ (**Fig. 2D)**, but similar to the $(GS)_{25}$ replaced NTD_{SCO2}.

It is widely known that sequence composition and patterning govern the properties adopted by intrinsically disordered regions. However, for IDRs adjacent to RNA binding domains and their binding interfaces, our results suggest that sequence properties can either enhance or suppress RNA binding affinity, depending on the specific IDR sequence. Taken together, our results suggest that the sequence of the N-terminal IDR adjacent to coronavirus RBDs needs to be relatively specific and is most likely conserved, albeit not in the traditional sense of direct sequence alignment; otherwise, without specific residues, the IDR could interfere with RNA binding to the extent of suppressing binding affinity.

7.7 Coronavirus Nucleocapsid Protein NTDs have Conserved Sequence Composition

While NTD's in coronavirus nucleocapsid proteins appear to always be disordered, their absolute sequence conservation is poor (**Fig. 1B, Supplementary Fig. 3**). If NTDs exist to enhance RNA binding affinity, and disordered NTDs can suppress RNA binding if the 'wrong' sequence is present, then how do coronavirus NTDs ensure tight RNA binding is conserved despite largescale variation in sequence?

The decrease in binding affinity caused by $(GS)_{10}$ and $(GS)_{25}$ mutant NTDs indicates that any enhancement in RNA binding provided by the NTD_{SCO2} is sequence dependent. This conclusion is consistent with our prior work, in which small changes in NTD sequence had measurable effects on RNA binding affinity as measured both by single-molecule experiments and by simulations [819].

Operating under the assumption that the NTD_{SCO2} has a role in enhancing RNA binding affinity of the RBD (**Supplementary Fig. 4)**, we reasoned there may be some selective pressure towards NTD sequences that result in a consistent macroscopic RNA binding affinity for the NTD-RBD. Additionally, while RBD structures are highly conserved across coronaviruses, their charged surface residues vary (**Fig. 1D**) [770]. As such, we also wondered if there may be a co-evolutionary coupling between the NTD sequence and the RBD surface. Thus despite diverging surface charge of the RBDs, conserved interactions between the NTDs and their respective RBDs could lead to a consistent macroscopic RNA binding affinity.

To investigate this hypothesis, in addition to the NTD-RBD taken from SARS-CoV-2 (SCO2)), we examined NTD-RBD constructs from five other coronaviruses: human coronaviruses OC43, HKU1, and 229E, the Middle East Respiratory Syndrome Coronavirus (MERS), and the Mouse Hepatitis Virus (MHV1). We reasoned that focusing on coronaviruses that predominantly infect the same host would ensure host selective pressures are consistent, thereby minimizing this as a confounding factor to explain differences in RNA binding affinities.

We first examined NTD physicochemical properties that are routinely used to describe IDRs (**Supplementary Table S3-S6**). Despite the large variation in NTD length, all NTDs possess a net positive charge, with the least positive NTD possessing a net charge per residue of +0.056. Expanding this analysis to 45 different coronavirus NTDs, we found no examples in which the net charge was lower than +0.056 (**Supplementary Fig. 5**). This is consistent with RNA binding proteins typically binding RNA through positive electrostatic surfaces that interact with negatively charged RNA [831].

Next, we examined solvent-accessible residues on the RBD surface. We generated five RBD structures for each of the coronaviruses using AlphaFold2, and then took the average of our calculated properties across the five structures [826]. The net charge per residue (NCPR) of the RBD surface residues stratified into three categories: relatively positively charged (229E = 0.126, SCO2 =0.066, MERS = 0.052,) neutral (HKU1 = 0.0, MHV1 = -0.011), and negatively charged (OC43 = -0.053).

In summary, while the surface charge of the RBD domains appears more variable, our analysis suggests an extremely strong bias for coronavirus NTDs to retain a net positive charge, in line with our expectation that these IDRs facilitate enhanced RNA binding. Compositional conservation in the NTD, or the retention of specific physicochemical features (such as net charge), could enable conserved interactions, despite lack of absolute sequence conservation.

7.8 Sequence Composition Alone Does Not Determine NTD Contribution to Binding Affinity

Since NTD composition is relatively consistent across orthologs, we wondered if sequence composition alone was sufficient to dictate RNA binding affinity. To test this, we performed a tiling experiment. Here, we repositioned the previously identified 30-50 residue positive charge block region of the SCO2 NTD (NTD_{SCO2}) that we and others showed to be involved in single-stranded RNA binding [527,819]. We placed this charge block at positions 1, 6, 11, 16, 21, 26 (referred to as mutants T1, T6, T11, T16, T21, T26) and 31 (wild type) of the NTD_{SCO2} **(Fig. 3A)**. Finally, we calculated apparent binding affinities of each of these variants with $(rU)_{25}$. These sequences maintain the same sequence composition but rearrange the amino acids, which allows us to determine whether there are positional contributions to RNA binding or if sequence composition alone is sufficient to achieve RNA binding.

To our surprise, the relative position of positive tiles has a significant impact on the apparent binding affinity **(Fig. 3B)**. Two mutants showed wild type-like binding affinities, yet the others bound RNA more weakly. This suggests that the relative location of positive charge with respect to the RBD tunes RNA binding affinity. Simultaneously, this result lends credence to a model in

which mere sequence composition is not sufficient to achieve 'adequate' (wild type) binding affinity.

To further test how sequence composition impacts RNA binding, we generated 386 scrambled NTD_{SCO2} sequences in which the sequence composition is identical, yet the order of the amino acids has been changed. An initial set of 172 scrambles were generated in four ways: The first by randomly shuffling the NTD_{SCO2}; the second by shuffling the NTD_{SCO2} while also making each amino acid change be as chemically different from the wild-type sequence as possible in terms of charge and aromaticity; third, by shuffling the NTD_{SCO2} while forcing positively charged residues from falling in the 30-50 residue region; and fourth, by shuffling the NTD_{SCO2} while restricting the majority of charged residues to the 30-50 region or a region spanning residues 4-17. Using these scrambled sequences, we performed coarse-grained MD simulations to measure K_A^* with $(rU)_{25}$.

Binding affinities were calculated for each of the scrambled sequences and compared with one another (**Fig 3D, Supp. Table 7**). The dynamic range of K_A^* observed here spans five orders of magnitude, demonstrating the dramatic impact relative amino acid position can have on binding affinity. However, for the majority of the scrambled sequences, the binding affinity is fairly similar, and, importantly, this "average" binding affinity is almost an order of magnitude weaker than the wild-type NTD-RBD.

Taken together with our tiling simulations, these results suggest composition is not the sole determinant of how the NTD_{SCO2} influences RNA binding. While 172 scrambled sequences is

only a fraction of the total number of possible sequence compositions that could be generated for the $NTD_{SCO2,}$ the observation that the wild-type NTD_{SCO2} sequence is among those with the highest apparent affinity suggests that the ordering of the residues in the NTD_{SCO2} is specific.

7.9 Disordered Region Residue Sequence Positioning Dictates RNA Binding Capacity

While most scrambled sequences had similar binding affinities that were much weaker than the wild-type sequence, we identified a subset of sequences that had binding affinities equal to or greater than that of the wild-type sequence. Based on our tiling simulations, we reasoned that the relative position of positively charged residues might underlie the increased binding affinity of these select sequences, highlighting regions of the NTD that are more binding-competent.

To assess how the position of positively charged residues correlates with binding affinity, we plotted binding affinity versus the average position of all positively charged residues in each scrambled sequence that we initially tested (**Fig 3E,** blue circles are the binned means of the first 172 sequence). The average position is calculated as the mean of the location of the arginine and lysine residues in the linear sequence of the NTD_{SCO2}. This analysis revealed a correlation between strong binders and the average position of positively charged residues. When the average position of positive residues is around residues 30-40, binding affinity is drastically increased in comparison to the other regions. This same region is relatively positively charged in the wild-type NTD_{SCO2}.

The importance of the position of positively charged residues offers a 'structural' explanation for the enhanced binding affinity afforded by the wild-type NTD. Charged residues within this

region enable the formation of a 'fuzzy groove'. One half of this groove is made of the positively-charged RBD, while the other half comes from the NTD. This fuzzy groove enables simultaneous interactions between the NTD_{SCO2} and the RBD_{SCO2} with RNA and, thus, tight RNA binding (**Fig 3C**).

Curiously, we observed a relationship between the average positioning of positively charged residues and binding affinity with similarities to our tiling simulations. We binned the scrambled sequences by average positive charge positioning and compared their binding affinities. The two bins that spanned residues 5-10 and 10-15, had binding affinities on average equal to the bin that contained the wild-type sequence. Bins that spanned residues 15-20 and 20-25 were each significantly different from the wild type bin ($p = 0.00013$ and 0.016, respectively), and both were weaker on average than the wild type bin. Regions that clustered their charge between residues 30-35 and 35-40 had significantly higher binding affinities in comparison to the wild type bin. This supports our hypothesis that the relative positioning of positive residues greatly influences the binding affinity and that certain NTD regions are more binding-competent than others.

We next investigated how the arrangement of positively charged residues impacts the variability of binding affinities within each region. Despite observing variations in binding competence among different regions on average, there was still a wide range of affinities within each region. We hypothesized that the clustering of charged residues, which is not captured by averaging their linear positions, influences binding affinity.

To visualize this, we used an inverse weighted distance (IWD+) metric to calculate the charge clustering of the positive residues arginine and lysine. When plotting the IWD+ values over our binned data (**Fig 3E**), we observed relatively consistent positive clustering for most sequences we generated. However, we noticed that the bins spanning residues 5-15 exhibited higher positive clustering due to the N-terminal positioning of the average positive residues. Even within these bins, there were sequences with lower clustering and weaker binding affinities compared to highly clustered sequences. Additionally, the wild-type sequence showed higher positive clustering and had an increased binding affinity compared to sequences with less clustering within the region encompassing residues 25-30.

We noticed several sequences in different regions that exhibited significantly increased clustering of charged residues and binding affinities. We then created a second set of sequences. Our aim was to enhance the average binding affinity in each region by clustering the charged residues. Given the generally higher binding affinities for sequences with positively charged residues clustered closer to the 30-50 amino acid region, we expected to observe a response where highly clustered sequences would have increased binding affinities in comparison to lower clustered sequences. We also expected that the average positioning of positively charged amino acids, when closer to the C-terminal end of the NTD, would have higher binding affinities than sequences with charged residues positioned closer to the N-terminal portion of the NTD.

To test this hypothesis, we generated sequences by scrambling and then constraining the final sequences to have all seven positively charged residues within ±2 residues of their respective bin boundaries. This resulted in sequences with increased positive clustering, as indicated by the

IWD+ metric. We calculated the binding affinity of these sequences, using the same methodology as the initial scrambled sequences, plotted them alongside their IWD+ values, and compared them to the original sequences (**Fig 3F**). As anticipated, the highly clustered sequences exhibited, on average, increased binding affinities in each region. We classified the sequences based on either their clustering similarity to the wild-type sequence or significantly higher clustering. Sequences with clustering similar to the wild type followed the previous tiling experiments in terms of how sequence location affected binding affinity. On the other hand, sequences with increased clustered charge showed higher binding affinities, often comparable to the wild-type sequence. Further, the clustered positively charged sequences displayed an exponential relationship between the proximity to the C-terminal region and their apparent binding affinity, highlighting how the positioning of charge impacts NTD-RBD ssRNA binding.

Analysis of the bound-state trajectories revealed a dynamic or "fuzzy" complex in which specific subregions of the NTD_{SCO2} contact the RNA. From the simulations of scrambled NTD_{SCO2}, we observed that particular regions of charge are sufficient to increase binding affinity. Thus, we hypothesized that we would find such charged patches within orthologous NTD sequences that exhibited increased binding affinity. Here we propose a model for conserved disorder without conserved linear sequence. Similar to how IDRs that contain SLiMs can exhibit sequence heterogeneity as long as short motifs are maintained[832], regions that contain conserved charge clustering can also have high sequence dissimilarity but still maintain sufficiently strong binding affinity for ssRNA.

7.10 NTD-RBD:RNA Behavior in the Bound State is Conserved Across Orthologs

Our scrambles confirm that the NTD sequence has a substantial impact on NTD-RBD RNA binding affinity. We therefore asked if natural NTD sequences encode a similar "fuzzy groove" binding mode, despite seemingly large-scale variation in NTD sequence and RBD surface chemistry. In this model, specific subregions of the NTD come into closer proximity to the RBD driven by favorable NTD-RNA interactions on one side and RBD-RNA interactions on the other (**Fig. 3C**). To test this, we performed simulations of each of the six ortholog NTD-RBD constructs with $(rU)_{25}$ and assessed the bound-state conformational ensemble of the NTD.

Bound-state ensembles were visualized using scaling maps. Scaling maps capture the average inter-residue distance between all pairs of residues for RNA-bound conformers. We normalized the scaling maps by the inter-residue distance of sequence-matched NTD-RBD simulations performed in the absence of RNA (**Fig. 4A**). Shades of blue reflect distances that are closer together in the bound state, while shades of red denote regions that are further apart in the bound state. For SCO2, this analysis identified two regions in the NTD that are closer to the RBD in the bound state ensemble centered around residues 10-20 and residues 30-50, similar to our tiling simulations and as reported previously[819]. This analysis can be done selectively for one of the residues in the NTD to visualize where it increases RBD interactions when bound to RNA by mapping its distances across the entire NTD-RBD construct with RBD residues colored with respect to NTD distance (**Fig. 4B**). Doing so shows that in the bound state, the NTD moves closer to the positively charged RBD β3 extension, highlighting the formation of a fuzzy positive groove between the positive β3 extension and positive subregions in the NTD_{SCO2}.

We repeated this analysis for the remaining five orthologs to determine if these NTDs also move closer to the RBD. In line with our expectations, this analysis reveals that in all cases, two specific subregions within the NTD come closer to the RBD. Despite large-scale variation in both folded-domain surface charge and NTD sequence, the mode of RNA binding appears to be largely conserved across the six coronavirus NTD-RBD constructs examined.

7.11 Discussion & Conclusion

Intrinsically disordered proteins and protein regions are prevalent across eukaryotic, prokaryotic, and viral proteomes. They play a wide variety of essential roles yet – perhaps paradoxically – often appear to be relatively poorly conserved sequences by alignment. In this study, we sought to understand how a specific molecular function (RNA binding) could be conserved despite large-scale changes in amino acid sequence. We utilized two domains of various coronavirus nucleocapsid protein orthologs as a convenient model that contains both a disordered region (NTD) and a folded domain (RBD) that binds RNA. Despite poor sequence conservation assessed by alignment across NTDs, we found that the orthologs were compositionally conserved. That is, the orthologs have similar charge properties in both the NTD and portions of the RBD. Specifically, NTDs harbor a net positive charge, while RBDs retain specific positively charged regions on their surface. Despite this conservation, the length and sequence of N protein NTDs vary dramatically, and while RBDs maintain the same 3D structure, orthologous RBDs showed a diverse set of surface properties, from highly negatively charged to highly positively charged.

To assess how the sequence composition of the disordered NTDs influences interactions with the RBDs and impacts RNA binding, we performed various coarse-grained molecular dynamics

simulations of coronavirus nucleocapsid proteins with single-stranded RNA. These simulations enabled us to interrogate the role of sequence composition and residue positioning in coronavirus NTDs ability to increase binding affinity of the NTD-RBD. We first showed, that replacing the NTD with a glycine-serine repeat sequence suppresses RNA binding, illustrating the impact that IDR sequence can have on intermolecular interactions. By testing hundreds of different sequences with the same overall composition, we determined that composition alone does not dictate RNA binding affinity. Instead, our simulations highlight the importance of clusters of positively charged residues, and that the relative position of positive clusters along the NTD also matter. Taken together, our use of rationally designed synthetic sequences illustrates that, at least in simulations, the absolute linear sequence can have a profound impact on IDR-mediated molecular interactions, even for simple systems using simple physics-based models.

Finally, we performed simulations of five orthologous NTD-RBD constructs, noting that despite dramatic changes in both the charge properties of the RBDs and the NTD sequence, the bound-state ensemble conformational properties were conserved, with the NTD wrapping against the positively charged beta-extension, forming a so-called "fuzzy groove" that can accommodate RNA. Despite differing sequences, we uncovered a similar mode of interaction between the NTDs and their RBDs, with each NTD having two 'hotspots' of interaction that coordinate to interact more often in the bound state with their RBD's positively charged β3 extension. Curiously, the ortholog for which these hotspots are least prominent (229E) also has the most positively charged RBD, pointing to a potential mechanism to compensate for a 'weaker' (less positively charged) NTD. Both our tiling simulations and scramble simulations of the NTD_{SCO2}, showed positional contributions to RNA binding affinity, which corroborated the role of the

'hotspots' in increasing binding affinity and supported a model where proper positioning of specific residues, in this case positively charged, is the important factor for a flanking disordered NTDs ability to increase RNA binding affinity. While they lack absolute sequence conservation, the conserved nature of these hotspots and their interactions with the RBD β3 extension opens up the possibility of developing inhibitors that can interact preferentially with the β3 extension to modulate the nucleocapsid proteins ability to bind ssRNA. The conservation of ensemble conformational properties in the absence of sequence conservation highlights how IDRs can simultaneously facilitate functional conservation despite supporting highly variable sequences.

While this study focused on the NTD-RBD from coronavirus nucleocapsid proteins, we expect that the information learned here will be widely applicable to a range of disordered nucleic acid-binding proteins. While absolute sequence conservation may not be present, there is still the possibility of conserved behavior encoded into diverging sequences. Rather than solely focusing on sequence alignments to provide information on conservation and important residues, quantitatively describing the ensemble that a disordered region takes on and assessing how it behaves with and without its ligand(s) may provide better insight into the residues that are important and sequence features that need to be maintained to ensure proper biological function.

7.12 Acknowledgements

We thank members of the Soranno lab and Holehouse lab for many useful discussions over the years. We particularly thank Dan Griffith, who has provided useful insights for data visualization with Python. We thank Dr. Emery Usher for help in performing Guinier analysis of $(rU)_{40}$ scattering data. We would like to thank Dr. Jerelle Joseph for the parameterization of the Mpipi

force field, which enabled us to do this work. We thank Dr. Lois Pollack and Dr. Steve Meisburger for sharing scattering data for $(rU)_{40}$. Funding for this work was provided by the National Institute of Allergy and Infectious Diseases with R01AI163142 to A.S.H. and A.S., by the Human Frontiers in Science Program (HFSP RGP0015/2022) to A.S.H, and by the National Cancer Institute with an F99CA264413 to J.J.A.

7.13 Competing Interests

No authors have any competing interests.

7.14 Figures

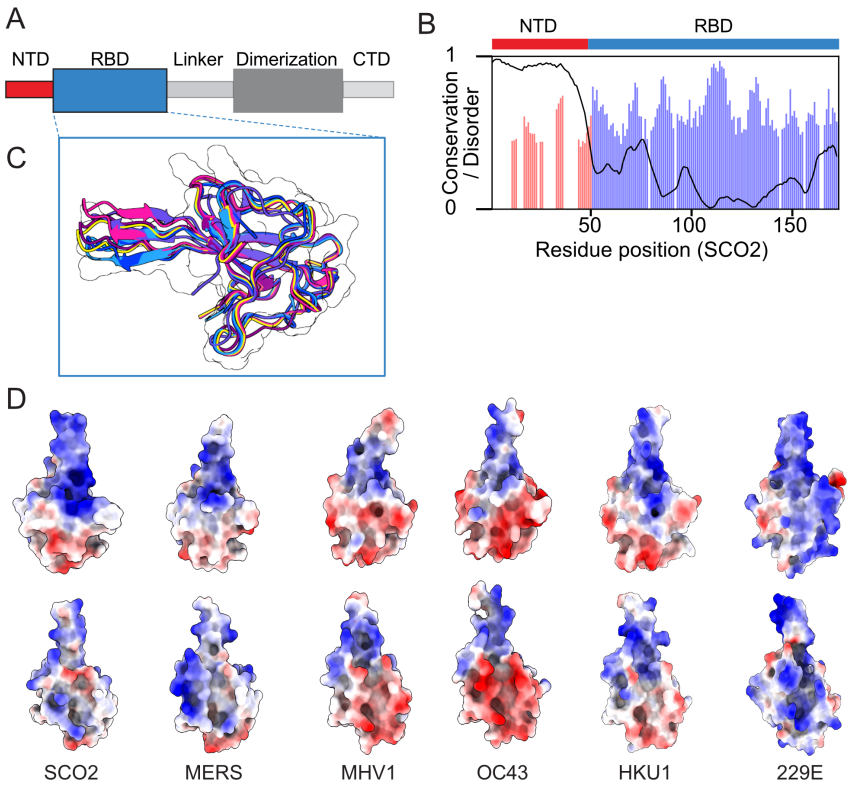


Figure 1. Coronavirus nucleocapsid proteins possess a disordered, poorly-conserved N-terminal domain (NTD) and a more well-conserved folded RNA binding domain (RBD)

A. Schematic showing full-length nucleocapsid protein architectures from coronaviruses. The nucleocapsid protein contains three IDRs (NTD, Linker, CTD) and two folded domains (RBD, and Dimerization domains). **B.** Per-residue conservation calculated over 45 orthologous NTR-RBD constructs, including SCO2, MERS, OC43, HKU1, 229E, and MHV1. Conservation is calculated based on the positional Shannon entropy, with values shown only for residues where 80% or more of orthologous possess a residue. The NTD contains many gaps in a relatively poor alignment, while the RBD is almost uniformly populated with relatively highly conserved residues. **C.** Overlay of RBD structures for SCO2, MERS, OC43, HKU1, 229E, and MHV1, revealing a high degree of structural conservation in the RBD fold. **D.** Surface charge properties of the six RBD structures overlaid in panel C, highlighting differences in surface charge properties despite the conservation of the overall fold.

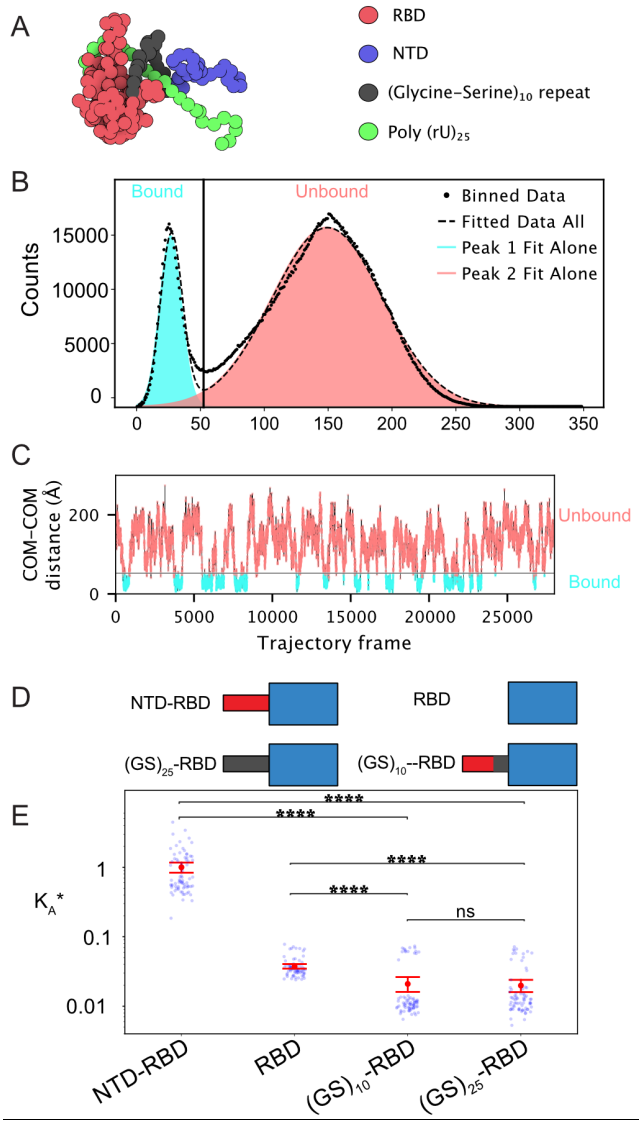


Figure 2. An inert disordered region can suppress a folded domain's RNA binding ability

A. A snapshot of the bound state from a $(GS)_{25}$-RBD + $(rU)_{25}$ simulation trajectory. Simulations utilize the Mpipi forcefield [405]. The model represents both amino acids and nucleotides as single beads with specific amino acid-amino acid and amino acid-nucleotide interactions. Folded domains are rigid, and both disordered regions and nucleic acids are dynamic. **B.** The distances between the COM of the $(GS)_{25}$-RBD and $(rU)_{25}$ are plotted over the course of the simulation. A distance threshold (black line) is determined in C (see also Methods) and plotted to delineate the

bound and unbound frames. **C.** COM-COM distances from B are plotted as a histogram and show a bimodal distribution that correlates with the bound and unbound states of the protein. The distributions are fitted with dual Gaussians. A distance threshold, which separates bound and unbound frames, is determined by minimizing the overlap of the two populations**. D.** Schematic of the four constructs shown in current "D" + $(rU)_{25}$. **E.** An apparent binding affinity (K_A) is calculated by utilizing the fraction of bound and unbound frames and Eq. 1. This is then converted to a relative apparent binding affinity (K_A*) by normalizing all values by dividing by the K_A calculated from the SCO2 NTD-RBD + $(rU)_{25}$ simulations. Blue points represent each individual simulation K_A*, while the red point is the mean of all of the replicate simulations for a given construct. The error bars are the ratio propagated standard error of the mean calculated using Eq. 2. Significance is determined by a Mann-Whitney-Wilcoxon test two-sided with Bonferroni correction. p-value annotation legend: (ns: 5.00e-02 < p <= 1.00e+00), (*: 1.00e-02 < p <= 5.00e-02), (**: 1.00e-03 < p <= 1.00e-02), (***: 1.00e-04 < p <= 1.00e-03), (****: p <= 1.00e-04)

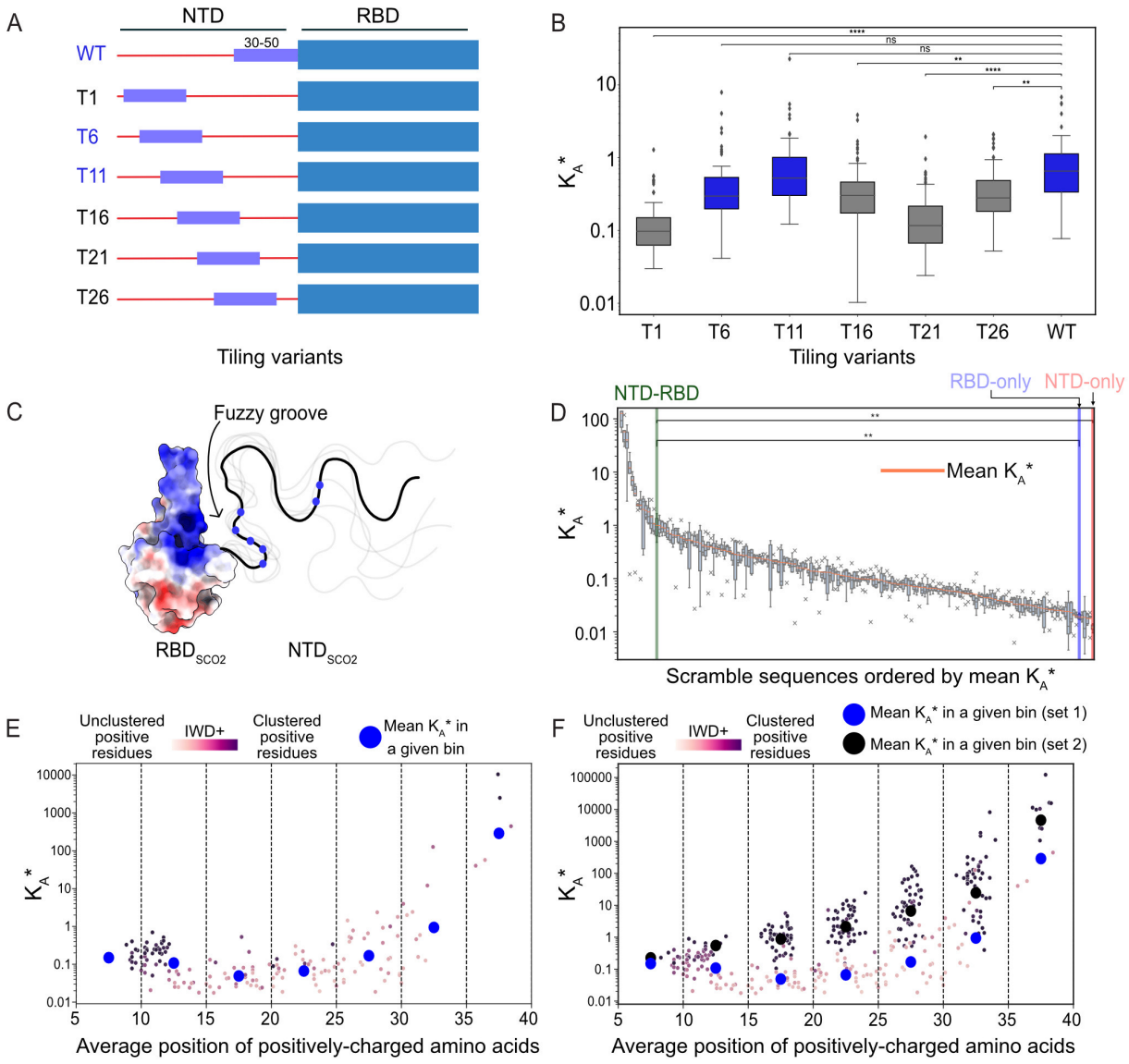


Figure 3. Clusters of positively charged residues determine the affinity enhancement provided by the NTD on RNA binding

A. Schematic showing the wild type and tiling mutants that systematically reposition residues 30-50 from the wild-type sequence. **B.** Binding affinity for tiling mutants schematized in panel A. Tiling mutant T6 and T11 show wildtype-like binding affinity, whereas all other variants show binding affinity less than the wild type. **C.** Graphical schematic highlighting the positively-

charged "fuzzy groove" that can form upon RNA binding between the positively-charged beta extension on the RBD and the cluster of positively charged residues on the NTD. In the RBD positively charged surfaces are colored blue, negatively charged surfaces are colored red, and neutral surfaces are colored white. A representative NTD is drawn with the blue circles representing the relative positions of the positively charged residues. **D.** Binding affinities for 172 scramble variants. Each variant reports on the binding affinity for an NTD-RBD construct, where for each variant the NTD sequence was randomly scrambled. Despite having an identical amino acid composition, sequence order enables a four-order-of-magnitude change in binding affinity, highlighting the importance of sequence in dictating binding affinity. **E.** Scramble sequences plotted with binding affinity vs. the average position of positively charged residues distributed across the sequence. For positional bins, average binding affinity is shown as a blue circle. Individual points are colored based on the IWD+ score, which reports on the clustering of positively charged residues (darker colors = more highly clustered). **F.** Same data as shown in E, with an additional set of scrambles designed to cluster positively charged residues. The average binding affinity of this second set is shown as black circles.

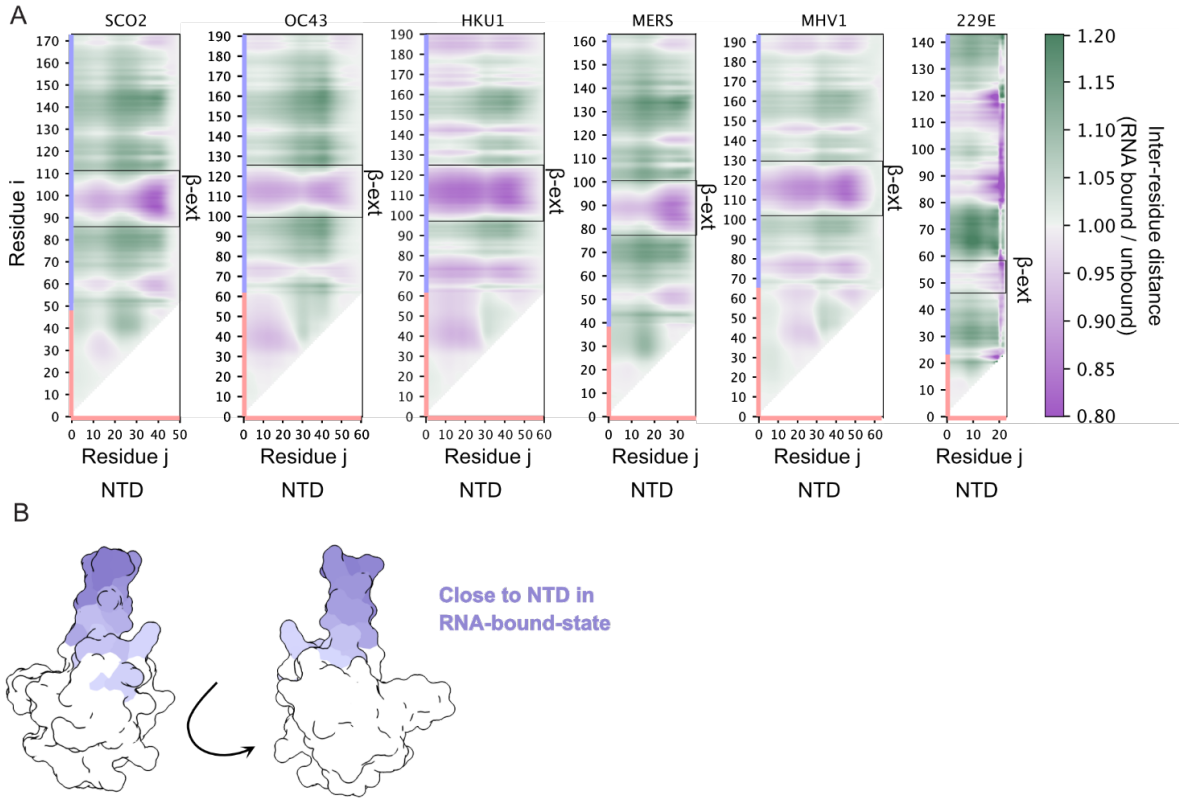


Figure 4. Orthologous nucleocapsid proteins show similar bound-state ensembles despite variations in RBD surface charge residues and NTD sequence

A. Scaling maps quantify the average inter-residue distance between NTD residues (X-axis, colored pink) and NTD or RBD residues (Y-axis, colored pink and light blue respectively) in the bound state. Heatmap values are calculated by calculating the average inter-residue distance in the RNA-bound state and dividing that distance by the average inter-residue distance in the RNA-unbound state. Purple colors report on inter-residue distances that are closer together in the bound state while green colors report on inter-residue distances that are further apart in the unbound state. In all six orthologs, the NTD is closer to the β-extension in the bound state, reporting on the formation of a positive fuzzy groove in the bound state. **B.** Regions closer to

the NTD in the RNA-bound state are highlighted on the SCO2 RBD structure in shades of purple with more intense purple signifying closer on average.

7.15 Supplementary Figures

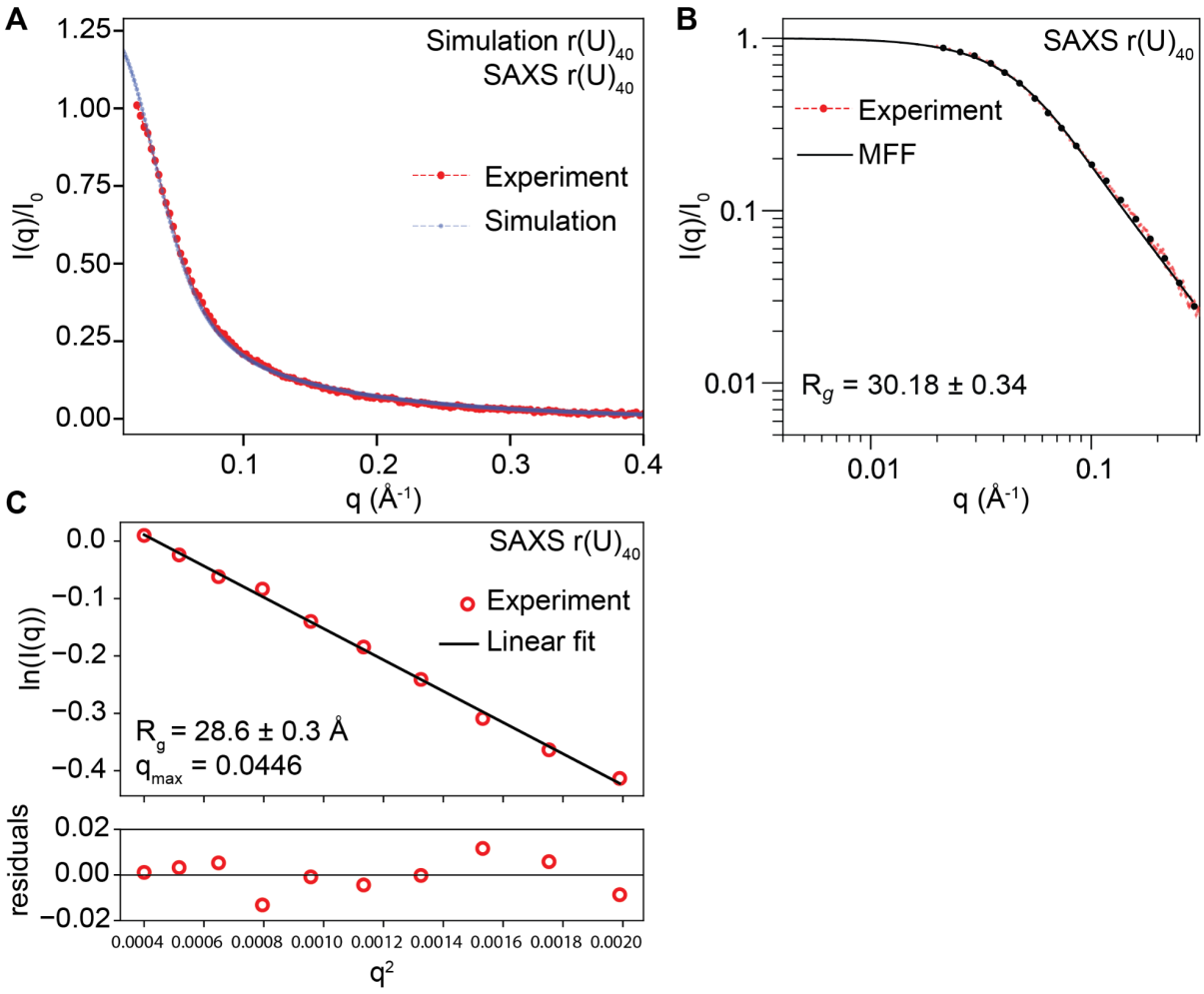
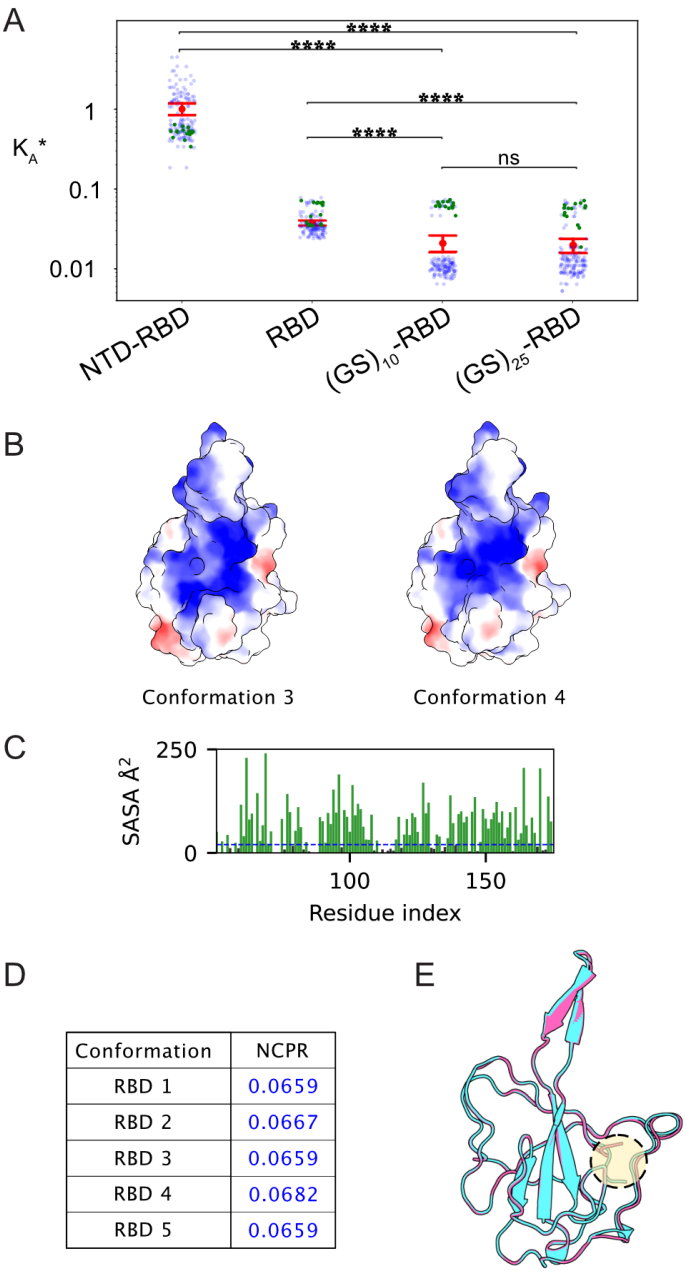


Fig. S1 The Mpipi forcefield captures the experimental dimensions of ssRNA

A. To assess how well rU homopolymeric RNA molecules behave in Mpipi, we compared small-angle X-ray (SAXS) scattering profiles obtained for $(rU)_{40}$ with scattering profiles generated from simulations of $(rU)_{40}$ using FOXS[822]. The agreement is extremely good, as shown by the tight overlay of the simulated and experimental scattering curves. **B.** We estimated the $(rU)_{40}$ radius of gyration (R_g) using the Molecular Form Factor (MFF) approach of Riback et al. [308]. This approach yielded an R_g of 30.2 ± 0.3 Å. Analyzing synthetic scattering data from simulations in

the same way yields an R_g of 30.9 ± 0.1 Å, while calculating the Rg directly from simulations gives an Rg of 32.2 Å. **C.** As a complementary analysis we also analyzed the SAXS data using Guinier analysis, fitting up to $qR_g < 1.3$. Based on this analysis we calculated a slightly smaller R_g of 28.9 ± 0.3, although this value is in good agreement with both simulations and the R_g obtained from fitting to the MFF.



Conformation	NCPR
RBD 1	0.0659
RBD 2	0.0667
RBD 3	0.0659
RBD 4	0.0682
RBD 5	0.0659

Fig. S2 Structural heterogeneity in the RBD impacts the relative apparent binding affinity

A. The relative apparent binding affinity (K_A*) plot from Figure 2E is plotted with RBD_{SCO2} conformation four plotted in green, highlighting its average binding affinity differs from the other conformations. For the wildtype NTD-RBD conformation, four behaves similarly to other

conformations. However, for the RBD alone and GS mutants, it is a better binder of non-specific RNA than the other conformations. This can be explained by reasoning that for constructs where the binding affinity is dominated by the RBD alone, the properties of the RBD will greatly affect binding affinity. However, where binding is a combination of NTD and RBD interactions, the affinity will be affected by how the two domains cooperate to bind RNA. To test this hypothesis we examine the charge properties of the different conformations. **B.** Structures of conformation 3 and 4 of RBD_{SCO2} with charge patterning determined by ChimeraX Coulombic electrostatic potential. While small, charge distribution differs around the β-extension that is involved in RNA binding. **C.** Surface-accessible residues are calculated for all RBD_{SCO2} conformations (conformation 4 is shown as an example). Green bars represent residues that are surface accessible, while black bars show residues that are buried. **D.** Net charge per residue (NCPR) is calculated for all surface-accessible residues for each conformation. Conformation 4 has a higher surface-accessible NCPR than the other constructs. **E.** Overlay of conformation 3 (pink) and 4 (teal). Conformation 4 has a shift in its N-terminal residues that alters its accessible charge patterning as highlighted by the beige circle.

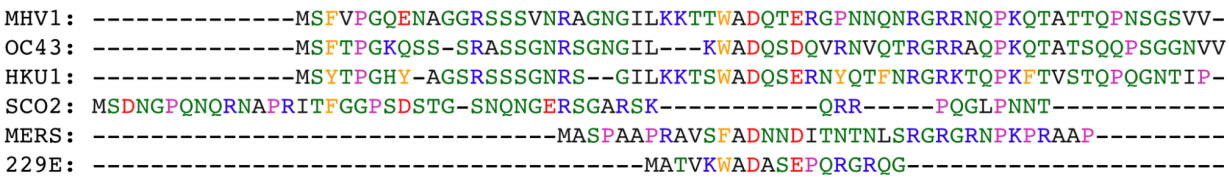


Fig. S3 Multiple Sequence Alignment of Coronavirus N-Terminal Domains

A

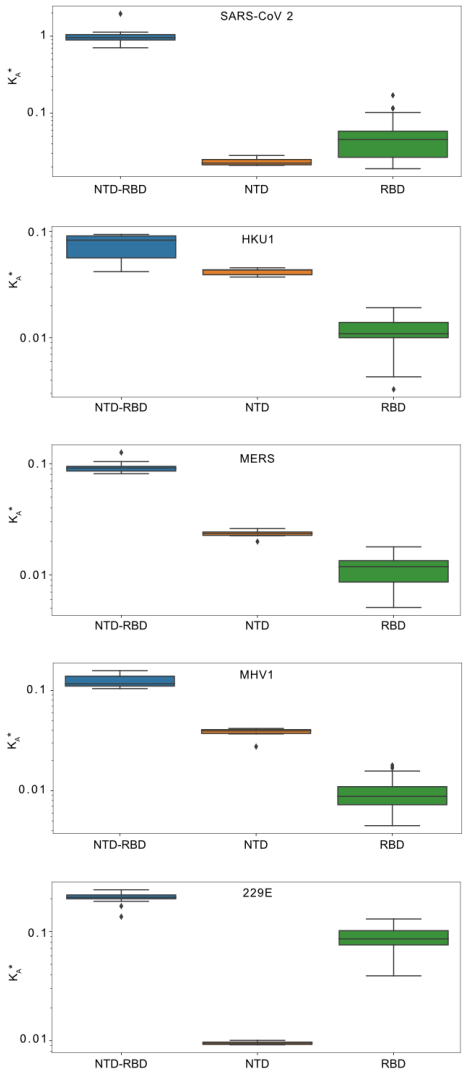


Fig. S4 For all NTD-RBD orthologs, the combination of NTD and RBD has an increased binding affinity than RBD alone

A. Five conformations of each orthologous RBD were generated by Colabfold and simulated with their NTD and $(rU)_{25}$. Relative binding affinities were calculated as stated in the methods for the NTD, RBD, and NTD-RBD simulations. OC43 is not shown due to NTD and RBD binding alone being too weak to fit to a double Gaussian distribution.

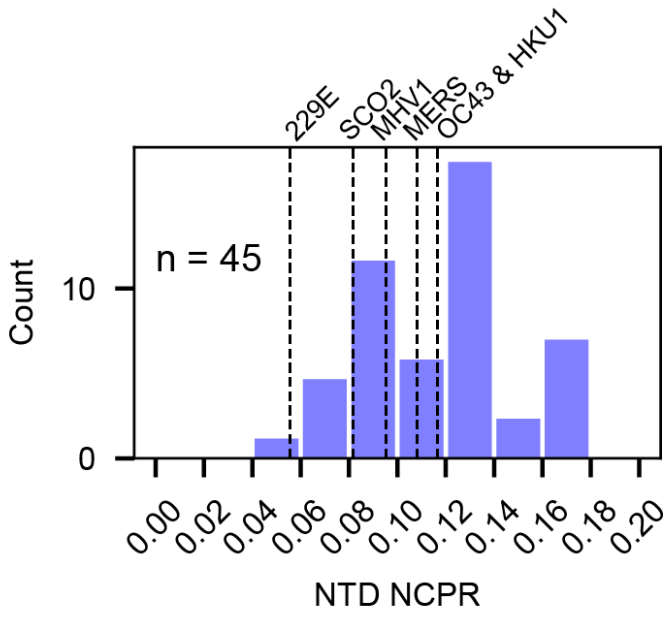


Fig. S5 Distribution of net-charge per residue (NCPR) for 45 different coronavirus N-terminal IDRs.

7.16 Supplementary Tables

Homolog	NTD Sequence
SCO2	MSDNGPQNQR NAPRITFGGP SDSTGSNQNG ERSGARSKQR RPQGLPNNT
MERS	MASPAAPRAV SFADNNDITN TNLSRGRGRN PKPRAAP
MHV1	MSFVPGQENA GGRSSSVNRA GNGILKKTTW ADQTERGPNN QNRGRRNQPK QTATTQPNSG SVV
OC43	MSFTPGKQSS SRASSGNRSG NGILKWADQS DQVRNVQTRG RRAQPKQTAT SQQPSGGNVV
HKU1	MSYTPGHYAG SRSSSGNRSG ILKKTSWADQ SERNYQTFNR GRKTQPKFTV STQPQGNTIP
229E	MATVKWADAS EPQRGRQG

Supplementary Table 1. Coronavirus orthologs NTD

Homolog	Sequence (NTD and RBD separated out)
SCO2	MSDNGPQNQR NAPRITFGGP SDSTGSNQNG ERSGARSKQR RPQGLPNNT ASWFTALTQHGKEDLKFPRGQGVPINTNSSPDDQIGYYRRATRRIRGG DGKMKDLSPRWYFYYLGTGPEAGLPYGANKDGIIWVATEGALNTPK DHIGTRNPANNAAIVLQLPQGTTLPKGFYA
MERS	MASPAAPRAV SFADNNDITN TNLSRGRGRN PKPRAAP NNTVSWYTGLTQHGKVPLTFPPGQGVPLNANSTPAQNAGYWRRQDR KINTGNGIKQLAPRWYFYYTGTGPEAALPFRAVKDGIVWVHEDGATD APSTFGTRNPNNDSAIVTQFAPGTKLPKNFHIE
MHV1	MSFVPGQENA GGRSSSVNRA GNGILKKTTW ADQTERGPNN QNRGRRNQPK QTATTQPNSG SVV PHYSWFSGITQFQKGKEFQFAEGQGVPIANGIPASEQKGYWYRHNRRS FKTPDGQQKQLLPRWYFYYLGTGPHAGASYGDSIEGVFWVANSQADT NTRSDIVERDPSSHEAIPTRFAPGTVLPQGFYVEGS
OC43	MSFTPGKQSS SRASSGNRSG NGILKWADQS DQVRNVQTRG RRAQPKQTAT SQQPSGGNVV PYYSWFSGITQFQKGKEFEFVEGQGPPIAPGVPATEAKGYWYRHNRGS FKTADGNQRQLLPRWYFYYLGTGPHAKDQYGTDIDGVYWVASNQAD VNTPADIVDRDPSSDEAIPTRFPPGTVLPQGYYIEGS
NL63	MASVNWADDR AARKKFPPP SFYMPLLVSSDKAPYRVIPRNLVPIGKGNKDEQIGYWNVQERWRMRRG QRVDLPPKVHFYYLGTGPHKDLKFRQRSDGVVWVAKEGAKTVNTSL GNRKRNQKPLEPKFSIALPPELSVVEF
HKU1	MSYTPGHYAG SRSSSGNRSG ILKKTSWADQ SERNYQTFNR GRKTQPKFTVSTQPQGNTIP HYSWFSGITQFQKGRDFKFSDGQGVPIAFGVPPSEAKGYWYRHSRRSF KTADGQQKQLLPRWYFYYLGTGPYANASYGESLEGVFWVANHQADT STPSDVSSRDPTTQEAIPTRFPPGTILPQGYYVEGS
229E	MATVKWADAS EPQRGRQG

	RIPYSLYSPLLVDSEQPWKVIPRNLVPINKKDKNKLIGYWNVQKRFRTR KGKRVDLSPKLHFYYLGTGPHKDAKFRERVEGVVWVAVDGAKTEPT GYGVRRKNSEPEIPHFNQKLPNGVTVVEEP

Supplementary Table 2. Full Length Sequence of NTD-RBDs from each ortholog

Ortholog	Kappa	
	NTD	RBD
SC2	0.364	0.191
MERS	0.448	0.204
MHV1	0.287	0.185
OC43	0.254	0.213
HKU1	0.278	0.191
229E	0.392	0.206

Supplementary Table 3. NTD-RBD orthologs Kappa values

Ortholog	Fraction Charged Residues		
	NTD	RBD	Solvent Accessible RBD
SCO2	0.204	0.202	0.242
MERS	0.216	0.204	0.198
MHV1	0.190	0.185	0.245
OC43	0.183	0.213	0.263
HKU1	0.183	0.177	0.237
229E	0.278	0.296	0.425

Supplementary Table 4. NTD-RBD orthologs fraction charged residues.

Solvent accesible RBD are calculated from the average of their 5 AlphaFold2 generated structures

Ortholog	Net Charge Per Residue		
	NTD	RBD	Solvent Accessible RBD
SCO2	0.082	0.040	0.066
MERS	0.108	0.032	0.052
MHV1	0.095	0.000	-0.011
OC43	0.117	-0.031	-.053
HKU1	0.117	0.008	0.0
229E	0.056	0.088	0.126

Supplementary Table 5. NTD-RBD orthologs net charge per residue

Ortholog	Hydropathy	
	NTD	RBD
SCO2	2.776	3.856
MERS	3.589	3.841
MHV1	3.219	3.769
OC43	3.333	3.854
HKU1	3.210	3.790
229E	3.378	3.713

Supplementary Table 6. NTD-RBD orthologs hydropathy

Ortholog	Fraction of Disorder Promoting Residues	
	NTD	RBD
SCO2	0.776	0.669
MERS	0.730	0.651
MHV1	0.730	0.687
OC43	0.737	0.664
HKU1	0.750	0.692
229E	0.833	0.616

Supplementary Table 7. NTD-RBD orthologs fraction of disorder promoter residues

Supplementary Table 8. List of scrambled sequences and their binding affinities

Rep = Scramble #

Rep	Sequences	Ka* average	Ka* std
0	SRGTSQGPNPDQKPIQQGSSGFNENRDPSMGTRAQGANNPRSNRRGNLTA	2.363	1.540
1	TRGGNLRQNNGTQSFAENTPPQIPDSGGGRNPKQDMGSSSAPRRNQSRNA	0.922	0.248
2	RNPNGGTQGRDANPRMGSRPPNTSSTLNSQQQINASNDPSKQGGFGRREA	0.036	0.010
3	SNSTFPNGAQGNGDNISRRSEARLQDGKMNSPQNQPPGPRQNTTGSRRGA	2.417	0.351
4	PNGSGNQNNNRSTLGSNSSGPRGPRMGTRSQATGAQRQDDNIEFPRKQPA	0.032	0.005
5	SEQANNSPTGGPPPPSRTRISQSNNMRQNDFNGKNTQGSGRRDQGLRAGA	0.592	0.387
6	SNGRGNAMTPNDSNQRRGNTFGLTSNKPPDARIGQGRGPPQSSQRNSEQA	0.066	0.025
7	GSDPTPGNKNNSAGNLSQGQSETGRIFNQSQNGRQDPMRARRGRPTPNSA	126.308	88.768
8	PQNLDSINFPADSSQNGNQRQPGTSSGTTMPRGNGGPGNRQAKESRRRNA	445.878	144.100
9	MNPTNGRRIRPDGNSNPFKASGQRSTNQDRSPGSGNEPSGQNAQTLQRGA	0.099	0.025
10	GPSGSGAGNNPSSQQRRSTQGRPNRDFTTMNNANQQIRRPGEPSLDNGKA	0.074	0.029
11	SPRPRQSNGGRGTNQMLNNQEPGISTGGQDSTANKFNPSRNRPGRQDASA	0.085	0.032
12	GRSNLMTAIKQSGGGNRTPPRQRGNANPSDQQESQTSFDGGSNRPNRNPA	0.133	0.062

13	QSTRTDFGQPRKDNQPRTRNPIGAQNSPGPRSNGSGSSNQNGENLARMGA	0.023	0.012
14	SSQQGSQPDTPRMIQRNEAPNRNRGSGNRPGSGAFTNQNRNGKLTPDGSA	0.033	0.016
15	NQQNRDTGIRGPKLNSPNNQNFRARSDRTMQEGPSNSPGPTASSGRQGGA	0.096	0.042
16	SGSTDGQPRRIGFAGTLPPSSGSNANSNQRNMKGNNRNDGTPREPQQRQA	0.137	0.029
17	NNRNGPPSQRQGQMRTPANNGSNSQNTPTQDPGSGRSEIRADFSKLRGGA	0.097	0.017
18	PMNSGNSRKPNNLNTGEFQSIPQPPTNQTADRGQDGGRGQRNSSGSARRA	0.457	0.420
19	GGNRRIPAKTTSSANGSNGNFPQSTQMSNQQGRDGRQRPLNRSPNGEPDA	0.097	0.049
20	QNRTGFNRMATGPPSRPSQSGASNILNRSSGDQEDQNNKGQGPGPRNRTA	1.438	0.929
21	FTAGPPSNSTQNSKTNRRPSGGGDLEAQNMQPGDSQRGNQRGPNRRNISA	3.965	2.316
22	PINTGRSPSQRGNNTDSQNPNQSNAFRGRPGKPDNAMLSSRTGGQRGEQA	0.063	0.011
23	QDTGPRSSNSPPSQNADANLTNRGGQTRRPGINRSKNQPGRGMFGNEQSA	0.482	0.121
24	QNSGEPGGDRTQRNDTGPSPRMFSRQQGRIQTSKASPGNLARPNNSNGNA	0.061	0.010
25	QDPPNPQMNKNASNEADSPNRGSTGGGRFSGRNTQRNQGSGILRTSQPRA	1.995	1.549
26	ERDTQDGNSFSRTQTLNPGSPNMPRNQSGPNKRGNGRAQSASGPGQNRIA	0.528	0.433
27	STRRGSDAPIQSKLSGPQRNSGDNTNPNANPSQQGPGQFRNRETNGGMRA	0.053	0.012
28	PSQPQRTQMNRGIDSTPDNPNSRNNGSQQRGPGSEAGRGTKGSAFRLNNA	0.113	0.023
29	GTMRQKDPSGTSTPGPNAGNSNRFNGRSLRNGDGR	0.028	0.010

	QPPQSNSNEARIQQA		
30	GQGTGNNSQKRNSRMINSQTASPRGPGSQPRNNPLDTPFRDEQAGRGSNA	0.068	0.027
31	QNPARSQFGNRIRNTLRDSGNRGEAGDNPKPQGNPRSSMTNSGTGQQPSA	0.033	0.015
32	QIPNRTDSRKTPEGLRMRQGNQANGNSSSGTQPNSFAPDGPNSRGQNGRA	0.065	0.006
33	NTRQSGGFNMRGTAETRSRPQANNDGNPPRPRQQQSSSSGKGNDNLPGIA	0.019	0.003
34	GNNARIQRNPLFRERNNQTSDPPSSRQQTRGTSGSGQDAGGKMPNPGSNA	0.027	0.017
35	NRSNFPNPQSQLGQGAPRRPDGMNSRIAGNETDTSSTNSRKGNRQQGGPA	0.493	0.289
36	NPLNMDRQPSSTNNDNRQQSTRGGGIRGGSKPGRQARSNENSAPGPTFQA	0.132	0.049
37	NGGNRQNDNRQPITNFRNRRSPTPSKQASRQSPTGSNGQEGGLDMPASGA	0.081	0.009
38	KPTPNRLQPGRGPSSTNGRTDSQNSGNGAQPRGAFRGNQENQRSSMNIDA	0.026	0.005
39	QSIPNPNTLRNRPFRQSMDGTQSGRNSGTGAGRGNGDEPANQRQNPKSSA	0.069	0.022
40	NRIQQTGSPRQDMGQANNGQTPASGPNRTGRGRKGSESLPNPNRSNSDFA	0.027	0.018
41	SGRSSSSNRLQPGREDDGTGRPNPPNGQNNANRKQTNPIQQMFTARGSGA	0.188	0.116
42	RKGQFRPGIQGPSSASDNNNTAMDNLQSERSSPRPNNPNGTGQRTGRQGA	0.220	0.233
43	QSAIANTTQGSSPSNPRRRQQGNRSDLRKRNGFSTNPQPGGGEPGNNMDA	0.104	0.056
44	SNLNGGQGPTGPIGRAMRPFAGNDSSTRNDQPRNQRETKSPGSNQNSQRA	0.078	0.010
45	RSPPNQPDRNGTAGGTNSAPRDFQNNGSSNSNRIGQMGRQKPLGRQETSA	0.395	0.233

46	SGRGQPPSSSMARSQNGRIRNGTPQDDPNRQEGFNQ ANLTGRTKNSPGNA	0.041	0.010
47	GNSGSDQANIPKAGNFNPPSQMQGSQGPTNGSLRER GDRPSTTNRNRQRA	57.134	35.759
48	GSGSKDQSGNQPFTSRGNGAGDANRNTLPNNPTQM PPRGSNQQERISRRA	0.156	0.037
49	GPQSGRNNIGASRANNDNFSQPGPESQRPNMTPSRS QRNTGGQGLDRTKA	0.053	0.017
50	FSSRRRNGPGRSGNRRSNNTDKSIATPGPQTAGSMQE QLQQDPGPNNGNA	0.095	0.024
51	GNQRAPRTRDQDGPRRQGNSQSSFQRSNNAIPNGNL GPSGMNGEKSTPTA	0.057	0.012
52	SMTGARAENGKNQFTGSPRQSQGQNTRRRRLPGQS PDPISNPGNNNGDSA	0.138	0.081
53	QMERSLSQRQNGTSRPNNQGAGGRRNGSDAQTSTR NSPPGDFNPGIPKNA	0.302	0.139
54	SNTNNGSRAQGNRTRGNQQRGDRRPISFPKGEPPMP ANSGSQQGSNTLDA	0.046	0.008
55	RRRPGSGNGGGDQQTSRRPNRNSATADQNKQMSGL ITFSNQGNEPNPPSA	0.066	0.011
56	SPNRSGDQQMGRNTPAPGNERPRSRGGNPQNTISFQ LQNRANKDTSGGSA	0.120	0.072
57	QPNNQSNGNPRTSPARRQNERQGKGRRGQDSSPGSS IFMLANGDNTPTGA	0.024	0.008
58	RFTGNTLPDEGNRASQNNNSINRRGRTQDQGGPAQ QPPRGKSGSSNPMSA	0.101	0.024
59	SQANRNDRSQTSRLESNQGRKANGGGRPRPSNNGSF GPMIDQPTPTQNGA	0.042	0.014
60	GRNSSNNTGNGQRKSIRDRTGNGSQRPREPDAQTG MPSSNFLAPGQNQPA	0.025	0.003
61	TRRRQQGNNGIRSQPMSGPARPPNSKLRSTNDGGQG NAFGTSPQNDSENA	0.063	0.012
62	RNRNSGPSGTGDSRRPKNGQNQRGSARITSPQDSNP	0.033	0.012

	LQGNPQFMETNGAA		
63	NPESSGAISLRSARFNQRRDDQRSRGKTNQNPGSPTMQNNGQPTNGPGGA	0.070	0.024
64	NNRMPGSPNGARERRKPNDRGTGGRQNGSNSSSSNPQTIQTGQAQPDLFA	0.033	0.009
65	NARPQGQRTRNSTNRSPFQSNGGTNMARERDGPSGISSDQNLKPNGGQPA	0.059	0.020
66	NNQRISRSGGRKNGETLAMSQRSRQAFRTGPNQQTSPPGPGSDNNGPNDA	0.018	0.004
67	FRQLSIGSAMDSRGPESNSQPNARRTRKRGDNGPPTQSNNPNGGQTQGNA	0.089	0.027
68	GGRRFSDPAQPGPKTSRQTGNNRMLRRNSNEIANQDSTPGNPGSNGSQQA	0.131	0.065
69	QGGRSNSPSRIKTDNRGMNGLENTDRRGRPQSASNPQNTQAQPGPGSNFA	0.043	0.012
70	TPSTQMKRPPIASNNESRNRNSGRSRQDRNSQQGQGDPLANGGFPNGGTA	0.081	0.024
71	NRPNKGNEDSPGTDQQPTRRRPRSGSRIQGNSAQPSQGNTNGNMSAGLFA	0.111	0.049
72	NGPNPNGPLDQRIGKMRSSQGRESGQRRTRSAGQDNFQSTTGPANPNNSA	0.078	0.017
73	RQKRSDDQGNGEGRGPNTSQRRGPMFNGQRSATNPTLSGNPANNPSSIQA	0.032	0.011
74	DRRGKSGPPPTRDSQMNTANRRTSPFESLRQNQQNSQNSAGPGINNGGGA	0.025	0.007
75	RNGSSPRQNGRMNPNGQDKRNQSRPIPSQRSNAQNGGDGGATFESTLPTA	0.030	0.014
76	QGTLKTNNRRNDNRQPPMGRGGNAPRNGPRQNQDTIGSGFEASSSQPSSA	0.026	0.005
77	RNNRREQRQGSTTKQSNSSGDSGRPRPFAPNGNLQSGGDNQPNMAGPTIA	0.021	0.007
78	STKNRRQMGIPPSNRENDPRQQANGRPRNSGGTTSDGLGFNQSPNQSGAA	0.024	0.022

79	RGDRAPNRSSLRGPPNSQSQQRSRKQEFTATNPGTQ GGDGGNPNNMISNA	0.066	0.015
80	PPRGMDTSSSQPGNRPNGRTRRQRAQEGKDTNSNAS QPINFNNGSGGLQA	0.052	0.010
81	GRQREDIQQQATNPPPGFSTGKRRSRAQSRNTGGN NPLSMNNNSPGSDGA	0.528	0.128
82	PRRMDGTQGRPRLGGPPQNNKSNANSRQRSENQGS STGDNQTPIANGFSA	0.031	0.009
83	SSERQMQIGPRNNGNGSARGTKRQGGSRNRPFDNT TGQLDPAQPSPNSNA	0.037	0.007
84	TEQPFQTGQKRNPQNIGSRRGRTNGRRPDAMNGNS SSDAPPSNQSNLGGA	0.212	0.046
85	NGKRPGTGSNSRQNTSPRSQNSRATRSRGPIDQMEG QNGQNPLDAPFGNA	0.047	0.006
86	RPRQASSGIRGKNFGTTRGNQDRGNQQGSRPSTQAP GNMLSEDNPNPSNA	0.025	0.004
87	DFAPTDNSQEGPMSQRGRGPKARRRLGRPSTNPSST QGNSQGNNNIGQNA	7.102	2.195
88	QGRRRANMTGNIQQNRQPRGNREGANPKLTSSPQD FSTSGNSPDNSPGGA	0.054	0.006
89	SKGGRNPMNQPDNNLQPPTRRRSTENARGRQGFSG QQSTNGASDSNPIGA	0.030	0.005
90	RRRSSGTGPNQNQLRPGQKNGGPSRQRDDTITANP MGPQSNSNAGENSFA	0.039	0.005
92	NRQQNNSQETDRAGFSRSGMGLRTSTIPNPSNGQPA DGSPGRKPNGRNQA	0.136	0.044
93	GNNENQDPQTSRSFTNKNSRIDNSPGGGAAGQNLP QPRTGGMSQPSRRRA	12.046	8.509
94	PPPARSKRDTGGNQRNGLSQTDRGMARENPRQNGQ PSTGGSNSQNFNSIA	0.018	0.006
95	EGQRSPGTSQTTGDNNRQSNSGRQSARRNQSFMPA DIGPNKGRGPNPLNA	0.103	0.029
96	PFGLNNQPGNGRPMGRRGANEPGPKSQSNNSQSTT	0.019	0.017

	DSSIRGQNATRRDQA		
97	TTTRNQPNRIGQPNNPKPQNQNARSMEGGRSQSND GSAGRGSDSFLGPRA	0.019	0.004
98	NDDNSTPGPGPRENFTGSTNIRNQKRGGNQSQGAP RQRGPANSSQLMRSA	0.368	0.097
99	STGEKRTSSGLIGDQGQSNGQGPMRPNGNNSQATP QRANNNPDPRSFRRA	1.784	0.889
100	SPGQSSQTNKSRGGNGLGTNDPNPGFTARRRSRPRQ PEDQNAGMNSINQA	0.339	0.099
101	DGIFQGSNPGRGLQPSQGNGESNDAQNTKSSTARPQ TGMSNPRRNNRPRA	40.181	27.957
102	NRPLDTRPSQGGNFQSQIRAPENNGRGGNRRNTMK PSGGSNQATPSSDQA	0.057	0.017
103	GTDQSNGTRQPEQSNPNIGNRGRLRNNPSGNQRASF PGMGKSDPTSRQAA	0.658	0.269
104	RGPQGSTFDNLRSSGTQQNRPNMGIRPPKRNQQNPR GSDGSNSANTEGAA	0.070	0.015
105	KRNFSANTDGNGTQDRGSQNQNGPAGTPMESNQR PLSPSGQIPRRGNSRA	0.784	0.259
106	NRRFANNRSGATKRSQQQSPTNTPEDSSLQQPGNIPP GGMGSGDNNGRRA	0.084	0.025
107	RGAGSNMATQTGNSRSGGQPPQKETRQPLNSQFNN RGNNSRPSRDIDPGA	0.039	0.013
108	GGTKNSNSRRSQDNATRTASRNPPNDGPIPPSNSGQL MREGQQFQNGRGA	0.044	0.009
109	GNRGRASRPRNSTNFLPKRMSTDPQTDPGAPNQGSG NNGRQQSSQEIGNA	0.072	0.033
110	STGSGPRDIQSKLTRRQTMRSDRNNQNANQGGNPN NGFERQPAGPPGSSA	0.052	0.006
111	SRRGGNPRPGGQSTSNPTGNRNRERSGQTIDSDQPF QANQPMLSNKGANA	0.026	0.010
112	AGPSGSNGTQSPNDNKQGRNRPSQDAQGILEPRPST GNRFRSTMGRQNNA	0.562	0.125

113	NRNSIDNNRQRNPSMTERGQATQTSGGPRKNGSGGPGASRPQPLNQSFDA	0.038	0.009
114	SLAERFQITQAGNGRRRPGQPGSMNTSQPGNNNGRSTPGRNKNDQSSPDA	0.078	0.009
115	GNRPMGLNQSSGTAINQKGSRNTDFDGNQSSPGQRQPTGANRRPPESNRA	0.148	0.045
116	PPQAGAPQQSTMGSRRFNQINPNETGTKNQDPGSRRDNRGRNGGSSLSNA	0.058	0.008
117	TNQESPFRRGNSGPPLRDTQARPQGKQGARNPSIGNGMTGRNSSSNDNQA	0.044	0.019
118	SPNQQARSNNNTPPFPERGTIQQGRGLRMSGSNSPDNTGKGSDRGRQNAA	0.037	0.011
119	STNGPSSNRQGSGDRPNGGFPLNTQGNDGRSISQAQMAQKRPEPTRNNRA	0.126	0.050
120	SKNGEQQGRRGNPGSDMAPFRINQRDTGRTPNQPNNTNSRQPALGSSGSA	0.028	0.007
121	SQTQRNSGSKARRNRERRPFGGNNMTNNSQITPLGPAGPPSNQDGDSQGA	0.099	0.026
122	QGTTSRRARRDRGNKRAGGGGQPTSSNNNPNSPQEPSSNFGIMNLQPQDA	0.232	0.049
123	MSAPRNRRERFPSDKRARGNQNSSTSGTNGNSNQQPQTDPNGGPLGIQGA	0.060	0.008
124	SNEPRQRGRRSKQPARNRMQSGNGSNQDNDTNSPSPGTGIAQLPGNFTGA	0.154	0.067
125	GGAQRRRRDGQNRRFSGKIGNQNQNGSPPALNSGMQTNPPSNTESTSPDA	0.131	0.081
126	SEGNFRRQMRKSDRRRNQPGNPSSSDPNTIQNGSQNTLGQNGATPGPAGA	0.278	0.127
127	LNNERRQTRPRPPRKGGRGFSGQSIMNNNTTGQSPQSDGADPGSNASNQA	0.094	0.048
128	DNTSPRPQRNGRRNKRRPLGTAEQQSGPFNMNGNDGIGSSGQPTASNQSA	0.393	0.161
129	NFPSGGSGRQNRKSRRRRNSSNEMPNGTADQLPDQ	1.065	0.550

	AINGGNGTPQSPTQA		
130	GNFGSGRQRRRNKQRPARQPSPSTGDIGPTQDSGQL AGSNNNPESMNNTA	0.194	0.030
131	PSNQRDRRARRGNRGNKDQTSPNGFMATLGQPNSN EITGSQGSGPNQPSA	0.077	0.037
132	LGNDRRRPGRRGSNFRKNPQANQSTMPNDGTSTSE NNPIPSGAQGSQQGA	0.200	0.048
133	QLAQRSRRRNRNIKPGFRQMSGGNSEGDGDTTPQA TNGNQPPGNSSNPSA	0.402	0.211
134	ENQNSPRRQLKRRRRSIGPNGTSMNQQQADSTSNPP NGATGDGGNSPGFA	0.889	0.201
135	PSNDNSRRQGRTKTIRRRNGAFSANPGGPGPGNQPQ MSQTSGSLNEQNDA	0.211	0.017
136	AGQNTRRRSRNRNGKRSPTQQSMDFGNSDPEQPNT SSIAPLNNGGGQGPA	0.246	0.033
137	QADSSPGRANRRRRRKQQNTPIFLNGGSPTDNGSPQ ESQTNGPGMSGNNA	0.645	0.342
138	NNGPLRRDSQRSMRKPRRGPNPITNDTGGFGEAATN GQGQPQSNSSNQSA	0.189	0.071
139	TSGQENSRPNRRAQRRKRNMSSQGGLFPGPSTNNGD DQAPGSQTINPNGA	0.638	0.171
140	TQSSFRRMERKQNNGRRRGPDSQGQPINTNTPGGN GANPLPSSQSNGDAA	0.099	0.029
141	DSAAKLIQRPRNRRRRFTDENGSSNGQGGSPGQSTN NPGMNTSPGQNPQA	0.698	0.331
142	TSEGRFTRRNGRGRRPKTNGQQQAQNSNSGLMPQP NIPASNPGDGNDSSA	0.252	0.079
143	NENGGNPRLRPNKRRRRQSGQGMNPTQNDSATSSQ NGSFGSIGDAQPPTA	0.286	0.220
144	NFDERRMNRRARNKPRPSQNNLPTTGQNTDPGSGQ SNGSSQSGGAQPIGA	0.115	0.034
145	PAFNQPRGERRRSRRGKPASDDSNQNQNGGQMTSN LGNSPISQTGPTGNA	0.271	0.158

146	SDTQRANRRLRPTKRGMRQGPQPNPQDENTGISSNANQGFSSPNNGGGSA	0.379	0.202
147	NENDRRAGGRRRGTGGRKNGQSSNSDNAIQTSPQPFPPNSMGTSQNQPLA	0.160	0.038
148	NQGNRRGGNTRGRNKRRDTSMINEDQSPPFQSSGASALQPPGQNGSPNTA	0.050	0.042
149	SSNQSGRRTRRLRRESKTAQNGGQINGPMNPDGFQNSAPGDPQPSTNNGA	0.232	0.031
150	NAGSGRPNKRTQRGRGRRFNQSTPPENISGSLNNTDGNDQSSMAPGQPQA	0.360	0.052
151	NNASQNKRRRSPRRGQSRPNNDTPGQADIGGTPLQSNMSQPSFTGGNEGA	0.204	0.096
152	TGDNKGPRRSRQRSTRRSGSMQNGGDIQNPPNTAGPLPNFSANSQQENGA	0.196	0.067
153	DAPQRSGPRGSKNRSRRRDNQGTNTQSGGNGAEPPTSLNSPNGNMFIQQA	0.277	0.114
154	QSPGRRRSTNRQDRRKGNSSTSAQGGALDNNFMPNGETGPNPIGPNSQQA	0.109	0.017
155	NPGSKRRQQFRIRNGDRRQPNGNGNPSSSEGLSGTNNADPMSGQPTATQA	0.069	0.022
156	GSNTGNDKNRRNRRTRREGIQSQGGFPNPMSLQPSTNQPQGSSGPADNAA	0.514	0.082
157	QGTNSRRRGRRKAITFNRPSENPPSMGAPNNGNQSNQTGQSDGGPLDSQA	0.156	0.051
158	DGGQRKGSNNIRTRRLRRMPNTPNSPASGNDQPQSNANSFPSGEGTQGQA	0.201	0.174
159	NPGNRNRRKQRRRGINESGTSFQPNALQSGPGPPASTQNSTMSDGQNDGA	0.229	0.045
160	SNFGGRQPSNRKRARARRNDNPQETSGPSQMGLQNINNTTDPGSPGGQSA	0.343	0.187
161	QADTGRLTKGRGRPRTRRGNPSSNQNMPSEANQGINSGPSSGNNFQDQPA	0.272	0.136
162	QPATQRRRRKQERSNRPNPNFAMTPQSSLPNGIQGT	0.328	0.140

	DGNGDGGSSSNNGA		
163	QTSPGKNRNFRRPARQRRSPNPTSGEGQDQSGANLG QSTNNGDNSGIMPA	0.245	0.033
164	NEGAFSRRTPGKRGRRQRSSPPQNTNDTSNGNDGS MQPSGNQIQANGLPA	0.418	0.140
165	GQDPPRGKRNSNRRMTRRFNAQPGNTTSNGDLGQE GQPSANSINSQPGSA	0.322	0.037
166	NQNIGPRRSNKRRDPRRLPSTMNGQGGQEQGDQNS AGNSFSGNPTAPTSA	0.249	0.119
167	PNGSRRNERQPNKRRRTQPAPGDIGTNFPGATQGQ NNLDGQMSSSGSNSA	0.096	0.024
168	TNSTRELRQKRNRSRPARFPAQSGNGQGGNPGISQD GNPSSTDNQNPMGA	0.130	0.034
169	FQQTRNSTAKRRRNIRGRSNPGESGAQSSPNMSNQG PGGLTPGDQPNNDA	0.242	0.035
170	QANDRNFRRGGGMKSRRRPNSTPITLQATNNQNEQ PSSGPQDGNSGPSGA	0.121	0.034
171	ANNPNGNGEPFQQRRRSRRRGKSSTQDNPSTGNGG SSNAQTLIMGPQDPA	0.992	0.385
172	QDTQSNPPIGNQNFQKRRRRQRRSEGGGTASSGGA GLDMTPSPNNSNNPA	1.904	0.655
173	GPGNSSEQPFNGTRKRSRRSRRLAMGIDATNTPNNQ GQNSSQPDNQGPGA	2.148	0.185
174	GQASGGNPQPNGNKRPRRRRRGDISDPTGSSNQMA NQQPTNEFNGLSTSA	0.892	0.311
175	GTPITGSQSPNPMRRKRNRRNQRGASPGSANQLENG FQGDSPQDTGSNNA	1.607	0.559
176	NGSNSPIPLQDGGRKSRRRDRRQSNANNNPNFQAGT GGQSTPQMSGTPEA	0.187	0.028
177	DGAGPTGNPGNQMRQRSRSRRKRSNPNTGPFQNGT SQLDGQISEANPNSA	0.654	0.073
178	AQINDTPNNQSQGRPRPRKMRRRAQNDGNGPGTNS PNFSSESLGQTGGSA	1.040	0.384

179	DANPQGGSSANQFRKRRRRRQPTESTLTNGDMSPPGGNNSNSPQQGNGIA	1.054	0.161
180	QGQPQPPLFAGNQRRKRSRRGNRNGDMEGSQSNDSTGNSTNANTGISPPA	1.327	0.151
181	PNNQNDSSTPGDGRRFARRQKRRTENPGSNSSQLQMGGPNIGTASNQPGA	5.569	3.532
182	PTTQNGPSNQSLSRKRRPRRQRSNQDPIDNNAGGENTGPNMQSGFGASGA	2.442	1.277
183	NTQFSGSQMGGNPRRKARRTNRRPAPPISGDQNQGPNLNGSQGSTNSDEA	0.373	0.060
184	TNNDPQTSSSGGARRRRKNRRGFANTQLGGPGMQPESNPQSINPSGQNDA	1.668	0.320
185	SNSQGPNTQPNNPIQRKRRRRRGSTPSNAPNGSQGGMLGETDAFQDGNSA	2.218	0.560
186	GDNQSNSNGNPQTRRDRKSRGRRPTGSQPMASPGTEIQNLGFNNQPASGA	0.770	0.265
187	PPSSNEGTGQTNPRKFRRRRNRSQSLNAPANGPQNSSGNIQMDQTDGGGA	0.838	0.124
188	GAGSEGNIGDTLPRRRRQQRSKRNFMTAPSQGTQNNQPSNSNPDGSNGPA	1.204	0.223
189	PGIGMPGSNQPNQGDRRRRRTKRQASTNSGANQGSNESTDNLSGQPPNFA	1.120	0.874
190	GFISMSSDTGNGGRRNSKPRRRRANPQDNNPQTGNLAQESGQPGNTSQPA	2.370	1.163
191	IGFQGPALESTGDRRQRRQRKRQGSDNNGAPNSSSGPNQGSNTTMPNPNA	0.105	0.077
192	PQTPQSGDGSQTTRRGIRRKRPRAFNGSSNNPGNQGDSQMNLESNPNGAA	2.097	1.167
193	STFQNNDGISAGGPRKRRRRNQRGSQPQGLANSPPGDNSPETMQSGNTNA	0.946	0.408
194	AQEGQSGNPTSPGRGRNRKSRRRFGQLPDAPNNTSGSNPTSGNQNDIQMA	2.731	1.086
195	NQAMSGLPTAGFDRIRGSRRKRRDNSQNGTGPSSNP	0.271	0.090

	QPSNQPGNTGNEQA		
196	SPPEPGGNSGQQGTRKPRRRNRRFIGNNDQGSSSSPAALTTQGQNNMDNA	2.561	0.737
197	SNINQQGMGNPNARRRPKQPRRRNTSETSGSDQGGTDGPPNGSQFALSNA	0.265	0.093
198	PNSTSQPNPTLSGGIRKRRRRRQNPFGNQDSNASSGTEQANGPNMQGDGA	2.935	1.443
199	NQNGMFNGTLNAPKRINRRRRRSTQQEDGSDTGSQSPGNPSAGNPPSGQA	1.213	0.784
200	TPMGPGGFSNNGIRRQNKRRDRRQTSEDNPLPSGQGAQSSGNPSQNATNA	0.268	0.084
201	PQQQIDQSLENTSRDRGRRPRKRMQTSGSFAPGPSNNPGNGASNGTNNGA	0.768	0.291
202	NGSLGPGPEPNPGNRRRTRQKRRTQSDNINGGGQMSSDNSQAQPNASTFA	0.757	0.475
203	NGDGSPGSFNNGSRRQRKRRIRGGDNSPQQGTPLQPPTAMQTSSNNNAEA	0.481	0.377
204	SNTGPGMQSTGDSKRRRRRRPGAGNSNAQGNNDQINQSFPPPSQNGETLA	1.291	0.568
205	GQSQQGSGNLAINRSRRTGKRRRNFTGGNPAQPDPPPNSMTQNSESDGNA	0.782	0.253
206	APPQGQPGQLDGSRKTRRRGRRNNNGASSQMPFGNGQNSIDTENSSTPNA	1.216	0.335
207	PQNMGLGSQTSNDRQRNRKRRNRTQGIPPNAATNEGQFGSPDSGGPSNSA	1.273	0.488
208	GQQFNLDSNGQAPRRSRNKIRRRGPPNASDQMTSGGNTNPEGSNTQPSGA	1.210	0.503
209	NGPQTSANGPNAQGSRRRKRRQRELIGGQPNSGQNDNMTTPSNFDPGSSA	1.831	0.897
210	PGSSGSQQSPQQPKDRRRRTRNRFGQPEANDTGISLGNMAGGTNNPNNSA	1.036	0.913
211	NETQPNNSDPPPNRLFRRRKGRRGGNSNDSMGSQSQGTAGGQQIATPSNA	1.139	0.535

212	QQQSFTSDPTPNSRRRKRRPNSRGPGPIGNNSDNGAGQATEGQMNLSNGA	0.701	0.117
213	SSGISQTPPGGNSRARRPRSRKRGNNQQTDLAQNGSPGNEGFPMNNQTDA	0.310	0.112
214	TGLSQGMPQGGPSRPRNRRRNKRSDTQANSTGSQGQNNADNSEGNPFPIA	0.895	0.480
215	QFMATNGGPSNGNNRRPRSRRKRTGNGSQEDPDIQPSPNQTGGANQSSLA	0.878	0.545
216	GQGFNIEMGGQNDRPPRRNRRRKTNDSGPAQLQNQNASPTTNSGSSPGSA	0.950	0.797
217	FGPAANSNNPQDQRKQRRERRRGGIPNGQPTSGPQLMSGDNTGNNSSTSA	0.225	0.088
218	GANQNEIQNGNGSPDRRRRRRRKSTPPSLAPQNFDTQSGTGNSQGPMNGSA	4.596	1.743
219	ALPSNMQNPADTGRDRRGRRRTKQNNSGINQPPSEQGSTPSNGQGNSGFA	0.769	0.437
220	GDFPSANNLNQNSSRNKRGRRRRGQASPPPMQNTGNSQSTGPTEGDQGIA	0.451	0.099
221	SEGPNTQGINSSSGMPSARSRARRRKGRDNNTQGTPPGNQQPNGFNQDLA	0.620	0.603
222	PPFDNMTIQGGGGPNASTSRRRQRSRRKSNPQNNGDNQEASLTQPGSNGA	9.138	2.387
223	QSPGETSINAQQPMFGSNRRRRGDRSKRPNPGGANTGNQSQTDNPNGSLA	1.428	1.199
224	TPGAGNGNNENGNAGSSSRRKRQNRTRRPSFGPPSDPQLDNQMQGQISTA	1.043	0.521
225	SGTLFQPNNIGEMASSDTQRRRRRSRKQPAGPGQPDNGNNSNQTGNSPGA	10.939	5.654
226	MGAPLNNDEGNNSSNNSTRASRRRRKQRPQGGSQNPTPGQSTGPDQFGIA	4.205	1.600
227	SPGNPGTGISQQGANGTPKRRRRRNRPTPESGLQDGNFMQSQNNASSNDA	3.707	1.265
228	FSGAPQNSPSSIGQMGQNRPRQRKTRRRGPENNGDS	4.854	1.766

	NLQNDANPGGTTSA		
229	GQSNSSNDGFNQSGGPDPRRRPRKRRSTQAGGIGSN NTAQNMNPPTQLEA	2.171	0.765
230	NEGPTSPSPGNAGMFGQPRRLRTKDRRRGATPQQIS GGDNNNQSQSNNSA	1.083	0.747
231	QGGPESQATQSNMPPSNGRRRRLRSPRKGNNNQNP DADGSGFNTGSTIQA	2.122	1.629
232	NQNPGFPGNSSPSISGQARRTRQRKRGRQPTPTDNQ AGNGLDENGSSNMA	2.315	0.774
233	FNAGGAGTQDIPSSNTQNGRRSKRRLRRPNSGNGM DNNQPGEPPSQTQSA	1.786	1.214
234	TPSGGEDNGINPDAQFPNRGNKRRRRRQASGMPSSQ NSPGNTGQSLQNTA	3.996	3.993
235	QIAESANSSQDSQSNTPGNRKRRGRRPRNFGTNGGP SPDMGQGNLPQNTA	13.916	4.699
236	MNQNPFESNISSSGTGNARRRRPRPRQKPGNTALND GNDSQGGPGQSTQA	3.803	2.379
237	PDDIGSGMTSQPNQTQNNRSKRRGRRRSAQNGGNP TEFPSANPLGNSGQA	1.406	2.077
238	MQISPTSQNGQGNLPQPESNRRKNRRRRGPGQSFNT PGGTAGDSDASNNA	1.294	0.324
239	LNGGGPNNSNSAPSSPSQRRRRRDQIRKNGAPTQFN TSEGNGDGTPQMQA	2.200	0.907
240	PDANAQISMTFSNNNNPLRKGRRRQRRNGSTESGNG GPGSGSQQQDPPTA	0.669	0.298
241	PGSENNGTGQGNLPQQSSRRRRKSRFMRGGTAIGSP NDANQDPNPTSQNA	1.671	0.310
242	GPNNPGQPADSGNQGNPANTLRRKRRRREGNSQQF DSNSSTPQIMGSTGA	24.669	9.803
243	QTNPPDGEMNANGLNFNGRRIQQRRRRRKSQTTPAQ SDSPPGNGSGNSSGA	1.205	0.530
244	NNQPGDQGNASPTMNGTIRSRKRRRDNRSGPGSLQ GSAPQNTSEQPNFGA	1.012	0.937

245	MGQSEIGGPLQNQNNAQSQFTRRRRRKRTSNSNTG NSDAGPPNPDPSGGA	1.388	0.339
246	NTNNEGDSSDGNSPGQPNRRQKRRRTRGFPPQQNI AGSQSSMPGNAGLTA	6.060	1.443
247	GPTFNSASPDPNGGPDSQKRRRSTRRGRTGGMNSNN ISPNAQLQEQQGNA	1.341	0.413
248	LPSAQQNGGTTNGGPSPQRRRRKRIRFDPNGDPGNT MNSNQAGESQSNSA	0.709	0.327
249	NPSNSTQTTGLPGPDQAPRKRRGGRRREQPGNSMNI FDQSASSGNNGNQA	3.881	2.487
250	SNSSGNGGTPTADGNGEPRQRRRITRKRPNSDPNGQ AQMFNLQPGQSNSA	1.366	0.469
251	STGAGIPNQQEQQPGFPDGRRRRRRDTKSANSNGL MSNGQSPGPNTNSNA	3.216	2.039
252	STATSNSGNGQNGQPQGARRPTRRNKRRSSQDEFD GINNMGGQPLNSPPA	1.760	0.617
253	GASSNFAGDNGQENSPISRRRKRNNRDRQGPSNLST NMGQPQTTGGPPQA	0.701	0.306
254	SQMQPGPTLDPENGIDFQRRARRSRKTRGSGGSNNS NGAQQPNPGSTNNA	3.154	1.344
255	IQSEMGGNQSDQPGFTGTRRGKRRQRDRSNPGANG SSSQPNPLNTNAPNA	0.390	0.104
256	NNESGGANTQTDQQSLSDRKRSPRPRRRPNGQMAG SNSQGFGGINTNPPA	3.684	1.473
257	EQPGDPDLGNIFATQGNSKRRRRRPSSRNASGSGPTN PGNTQGSNMQQNA	1.097	0.714
258	PPGGPSNSSSQADQFNNGRNRARQRRKRNGNLETN PQSQSIGTGPMGDTA	0.748	0.289
259	DNTNNDGPTGSSASNMQNRRPRKRRRQGLEQIGGQ NQAFPGPNPSSTSGA	6.686	3.235
260	NPGNGNGPPASEMTPGQDGPRRRRKRSRTIFLNQG NSGQSQTNNADQSSA	13.981	3.187
261	QNPGAGSFPSDQSQPNNSQRGRRRRRKGNTTMEDN	7.407	2.099

	AGLPNTQSGGNSPIA		
262	PANNGFGIGQTESDSPGQRRKPPRRDRRNQTQLTSS APQNNSMNGSGGNA	0.149	0.054
263	NDQQPNANSIDNSGEGGPRKPRRTRRSRASSQGQPG MFQPNSGTTGNLNA	7.696	8.015
264	TGPQQQGEANISPLGSSDQRKRNGRRRRGTSNNPNN PSMGSFGPNATQDA	8.431	1.907
265	NGQGSPGSDPSNMASQNGRGPRRRRRKDNTQTSEP GTAQFPILSNNNQGA	4.790	2.640
266	NGGQDMGAPNQSGNSPPIPRGRRRQRKRSAPLGQTT DGSETSNNNFQNSA	3.219	1.561
267	QAPGTGDNNPQPNGNQPDKRRRRSSRRFGSQITAG NENNGPMLSTGSSQA	3.725	1.624
268	TAGGEQNSDGPLQSPNNTNRRRKGRGRRQSPNSAD SSNQFQNPPGTGMIA	9.709	5.751
269	ANDNSPQIGQSPTPGNSTRQRSRGRRKRGNSDPGPN ASTNQMGNEFGLQA	6.552	1.044
270	GNNQGGNMSQTQGSGTNFRDRLRRRTRKSPPSPNIP NSPAASDQEQGNGA	0.751	0.368
271	TQGGGTQNNSDNAGNGQDSPGQFKTRRRRRPNRIS MSNQPAPSESLNPGA	28.752	12.245
272	GQQEGSNSPNPDNTLGMNINTPSKRNGRRRRGRGG FSPSSADQPNQQATA	78.551	63.623
273	FTQQNNGSSGAINGQSDTQPTSERRKRRARGGRSLPS DPNNGGNPMPQNA	0.787	0.454
274	GSNAFAGGDTMTTQGGQNQNNGPRKRSRERRRPSN QPLNSSDGNPSIPQA	0.535	0.166
275	QNESMTQGPNGPNNDDQNSGNLFRRRSRRPRSKSSG AAIGQPGGPTQTNA	9.241	2.491
276	PSNSQNFSPNGQGNGPNNQDTPPRRRRRQKTRAQN EGGDTSMSGSGILAA	8.602	2.851
277	ELQNIPTSPQNFGTSQDSPQNSGGSRKRRRRGRAQSP DPAMNGGNGTNNA	27.887	19.434

278	PSQSNNGQGQSPGQGQNSNLSNPPRSRRRRKRENDF GTGMDIPTNGATAA	1.115	0.732
279	NNQLSDNGNPGSQAFDPPGGNNSRNRRRPRKRQTG STMQGIQTAESGSPA	3.456	1.927
280	TGFGDSNGQTSTIGQQPLPSQPPRKGRQRRRNRPNN SDANGSANEMSGNA	7.735	2.074
281	LQQNSPSSPDNGITSGGETGSNSRAKRRRRRFGPNMA PTGQDNPQGNNQA	51.045	33.968
282	SNSGSPSANQQETNFSPANMPINRRKRDRRQGRTQP GLNDTQGPGSGNGA	4.495	2.161
283	GSDQDNGSNTGAINETGQLPQTMRRRGRGAKRRPS NSNPSSGFPQNQNPA	36.676	32.758
284	QGNLGINSPTQQFPTGSMDESPQRKRRRAGRRAGGS PDGNNNNTNSQSPA	1.812	0.437
285	NGSNNPGTNINDNQAGQDMQSGFRRSPPRRRRKQP LPTGSNGSQAGSETA	29.174	23.579
286	NNPQQNGPQAGLNGQDNQSDIGPRRKRRRTSRGAG GNSTSNETSMFSPPA	4.381	1.613
287	SSQPSENGPFGITTGGQQPMNSPNRLRKRGRRRQTN NPGDSDNQNASAGA	29.874	7.870
288	GMDNQSSGSNNAGPDGEANSNQNRRIRKRLGRRTQ PQTGNPFGPPSQSTA	66.032	46.026
289	GNGQGNSNSGPDPSQQDMNTENLKRARSRRRGRG GNPPAFSNTQSTQPIA	13.359	12.420
290	TGSTNANNAGGNGLGIPSNGQSMNKQRRRPRRRSSD QDPQQPNPTEGFSA	17.741	12.614
291	TQPSNNSPPMQGQANGNTAGLNFRTRRRKRPRGSGS DQNSNSGPDEGIQA	3.614	5.693
292	STNDFGQINGNNSSPTNAGQDEGSKRGRRRRPRQSP MNGPGQQLNPTASA	117.597	81.866
293	SQNFNPPNLSSTPGSADQGQTGNRGRKPRRRERNM DQIQGTANPSGGNSA	0.531	0.138
294	DDIQTFQNNQSNGMGPPGLGGNQKRRRRSPSRRSAS	6.551	8.074

	SPTTANPEQNGNGA		
295	TQAQDNQNSPGDTPSGESANSNTRGLRRRKRIRPGP GGSNNGPNFMQQSA	21.862	11.567
296	AASNTNDSQQQGIGSQNLGNGGGRNRKRRRESRPN FPQPPSNMPGDTSTA	0.848	1.419
297	NDSTQGNGPGSNQPSGESAFNQMQRRKRRPRGRNL GTQNSGITDNSAPPA	2.778	0.616
298	MPNTNGPQQEISFSNGSSNGNPTGRRRRRAKGRSGN NQAPTSQGPQDDLA	21.380	3.110
299	SGINDGTSQNLEPPMNDSQGGASRKNRRTGRRRNQ FAQNQGPPTPSSGNA	5.451	2.221
300	NNGNGQTFSGLQSGITDNNNMTQKRPRRRDGRRGP ESQGNPPAPSAQSSA	2.757	2.553
301	GSINPDNGPPSTGGGAFDNGLTQRGRRKRRNQRNQ SSPQNPSETQMNSAA	10.007	7.579
302	MNSPTQPSNDQDEAGLQNGNQSTRGRNRRKRRGPS GGPNQPSAIGNFTSA	163.026	57.051
303	DPPNNTSQNGLGEQNGQMIPTNFRRRDGTRRKRSS NSSGQPAASPGGQNA	13.254	2.623
304	SSQPNMQNNGQNLGPPNPDNNPSRIRKGRRRRGSD QTAASGFEGQTSGTA	2.480	0.570
305	LNDGTSGDQMNPPGSQPSTFAGNKQRRGNRRRRSG TQNIEGSPNANSQPA	42.929	35.001
306	QGSNSSAQNSPDSSNQPTNFPLNQRNRRRRIKRMGT GDNPPEAGQGGGTA	10.991	9.071
307	PQSDNIGQTQQLNTSNSGFAENPRKRARRGRNRPNG QSTPDPMGGSGSNA	64.173	24.330
308	TSGQNGQPFLSQNGPGTSESDPNRRRRKISRMRQNP NPASTDNNQGAGGA	1.780	1.125
309	GQDLSEGGNQQQSSPSNNSPFTDRKRRRRARGGMG GNAPNPSIQNNTPTA	5.307	0.960
310	GTSGGQNTGQSTMNSAGGLAQNNKNRRRRRNRQD NPSFSQIDEGPSPPPA	11.873	10.659

311	PSGNSPNTGSQEMLQISPDQGAGSRRRGKARRRPNF DNNQNTPTGGNQSA	87.348	31.554
312	DGFQQNISEGNGSNGQGPANDTTMRKGRRRRRQT NSPGLASNSQPNSPPA	136.269	77.374
313	LQAGFPGNSQTNQDTNTGPAPDMRKRRSGRRRSSPE SIQQGNNGGNSPNA	12.417	4.498
314	NAFSGNQGPPASQSDILQGNNGNKRPRRGDRRRTG MNTNEPSSPTGQQSA	12.087	6.863
315	QQGQGSTGNPSQLPMNTQNADTAKRNRRRIRRNFP GGPPGSSDNSSNGEA	11.890	7.626
316	SPNLNNSQSIQSSSMQFNPPDPGARRRRQKRNRNQG TGDGGTGTGNEPAA	2.617	0.395
317	QSNGPQTSNSSEIGFQNTSTSGDNKPRRNRRRRAGP NPNGMADLPQQGGA	1.037	0.576
318	ASSSNQQFSNNTQDLGIPNQPEGKRRRRRQTTRNGS PPSAMPGGNGDNGA	8.539	4.873
319	NSQNMPGGPAQGNINTEGTSGSSRDRDRRRRNKPT GNPSQQGFANPSQLA	0.694	0.518
320	ENSDGNGNAPQFQNGDPMQPQNGRGRRKSQRRRT SGGSNTTPALSSNIPA	9.605	7.051
321	TTPDANNAQGNGFLQPGQSSIGGPQGSNRNKDTRR RRRQSNGMPPENSSA	16.724	30.411
322	QMGIPNTSGPSQFGQEPNGQNANGSGTSGKNRARR RRRQNTDDSNPLPSA	71.885	49.463
323	NSGTGSQNPPGQSIAGPDNNMAGTFNGSRRDQRRQ RKRSGLPSPNQNETA	4.311	2.556
324	PNSNQGTDSPGNASNLEGNQQPQDPGSMRRKFTRR RGRASNNIPTGSGQA	275.400	349.123
325	QGGAGNQMNTSPSDSTTISPQPLGPNDPKSSRRNRRR RNQNQNEGGFGAA	327.639	664.680
326	NPQNNTFGASGGLGQSMNENNQPTSQIGKARRPRS RRRQTGSDSDNPGPA	227.587	110.054
327	NSTQIGEPASSNSTGQGGADFGMSNPNPRKRTRRQR	111.354	79.094

	DRPPGNNGSNLQQA		
328	PNQGDSESPDGSTGNSGNNNTSPSQQPNRIGRKRRA RRTAMPFLGQGNQA	100.893	94.087
329	GGSGILGNQFNPGSATNNQDQNPQASPTNRRKRPR DRRESNPMGTGSSQA	1.319	0.619
330	GNQDDQQNSSQESINQFGPPNGSNPSASRRNRRGRT KRLTNGGPMTPAGA	40.485	56.131
331	GGGSGSDQTNAQSFGPNPEQQTSGNAIPRRMRSNKR RRDGSLTNNNQPPA	227.283	161.009
332	NMQDPQPGGQNGNLTGTSSAGPNPSFQITRRRKRER GRQSSDAGSNPNNA	39.687	32.402
333	DLNINGMNESFGPTNSGGSSQNTPQQQGRRRRNKQ RSRGPPATNDPAGSA	69.786	41.415
334	PPSGTSPSNGFGNGMQINNTQGNDNGNTRAGRRRR SKRPSQDQLSPEAQA	1.939	0.654
335	GNFSNSLQQASGPINGGTAPNDQPNNNDRTSRRRPR KRQQGTPGGSSEMA	0.397	0.055
336	QSDQGNGNPTQELNAIAPQSSSFNDSGNGPRRKTRR RRGMTSNPPGQNGA	8167.374	4679.335
337	DGANALNIGGSQNSSPGQSSEQTNPGMGKPRRRTPR RRPTFSDGQNNNQA	176.150	42.024
338	LGGANMTEGGPNTSIGDNQSNQAPSNPFRGRRKSQ RRRQSNPDTSGPQNA	122.630	108.895
339	PPESSNNFNGQTTNPTISAAQGGGQNMSRKRRRRRQ PGGNDGSDNPQLSA	86.825	79.913
340	SQPSGPGSGGDTNSPPQPDGNSNGNQFEKRRNRRR NARMLTNTSQGAIQA	55.761	37.267
341	QTQGPNSSSANAGSGDTQQNSTPGNFSDIRRRRGKR RMPNQGEPGNPNLA	73.533	61.964
342	PNGSNMDDQPAGGPGNSPQQTSTEFANQRSRRKRR RNTGGQPNNLSSGIA	49.750	77.367
343	TQTPSPAGNGQFILTGEMNSPDADNQQSRRPRKRRS RGNSNQGGSPNGNA	322.527	210.639

344	SNSQTGAAPPPNFQNGGGGSTEGNGMPNRRRKQRR SRSDQNQILSNPDTA	7.093	1.663
345	TGMPPNQQNDSPNGSNFSGPGGNILGQGKRNQRQR RRRATSNSPASETDA	19.733	16.108
346	GPSNNQSSPFGDGMLISNNGTTQNPQPERTRKRRAR RGNQDGGSNASQPA	54.745	72.541
347	GSPIQGQNNSANDTNGQTSFSGSNPMAGLKRRRQRR PRSGDPTPNGEQNA	33.377	37.549
348	FNAGMNETGGSADQNSNQGGTQQSDGSLPNNKRR RRRRPPQGNSPITSPA	1108.910	555.489
349	GNSDNGQNLEGTSQPGNAPNPNPGSGTQRRQSFKR RRRSMNQSAGITPDA	1.365	1.000
350	SFPNAINGQQPQNASQDSTGPNEGNMGLRKGNRRT RRRGSNTPQSGSPDA	6.352	1.696
351	QSSMGSGNNNDGDNQPQITPGQNTLENPGARKRRR SRRGNFSQPASTGPA	16.531	6.090
352	NNGMPNQPINGQNQTGSTSQLGSFGASPRRRNKPR GRRSGPQANDESDTA	4.504	2.523
353	GPSDTTQSNQGNQPSNAILQESGNGGPGRRRSMRD KRRPTSNGPNNAQFA	1.043	0.582
354	NNQSNPSQQGAQDGNGNAPMTPGPPNNSSRFRGKR RRRGEDSILGTQSTA	17.098	10.976
355	PQQQNDDGNPSSGTTTGQNFSQMAPESNRRKPRRG RRNGNAPGNLSIGSA	197.226	127.581
356	NSSQLEPTISGNNDNGPTPAMSPNGDGNKRRSRRRA QRQPGNTGGQSQFA	47.644	56.124
357	DNQINQGNAPMPPEAGQDGGFTSSSGGSNRKRRSG RRRNNQNLSPQTPTA	1493.647	317.584
358	QNQQSNSLNNGTASNMPPGPETQGNISPGRRRRGKS RRQTDPGFGDSNAA	2.870	1.526
359	LGPQPSSQQPNSNSGIGDMNNNGFPSPTRRARGRRD RKGGNQNEASTQTA	6.818	14.119
360	GTNNEGSGNSPDGPNGITNSNMQPQPAQRPRRSKSR	1848.318	844.283

	RRNAFGDQQSTGLA		
361	QSQTTGSGPPQGGPATGNNQGDSNMNPDRRRRRESRLKNQSPNSIGNAFA	1.299	0.853
362	PAQSNSPPGQGSNQPNTQEDGGTGNSTLRRRMRDKNRRSPSNAIGQFGNA	1.065	0.503
363	LQGSAPNNNSGMSGPPNQGDQNGDPGQNKRIRTPRRRRNSGTTASFSEQA	67.831	46.433
364	NNTPLNNIGQPTGSGSASQENPGGFPMQQRRRRRGKRDSSSANGPNTDQA	128.774	36.707
365	TNIGLNADSPNNSGSDSGQFSENGTNGSKRRGRRTRRPQQPGNPAMQPQA	97.866	68.059
366	TLPSMSGNTSAINNGPNESSNGPNPGSQRRKRRGRDPRGQTGQDFNAQQA	70.854	47.715
367	AQTDMNSSGLPSESTPGQNGQNQDNGINRRRNFRRRAKQPSGGPTPNGSA	95.266	189.611
368	NSGNGPASSSPQPPTTNNNDQGEDQGISRKRRRGQMRRGTSLNGNQPAFA	82.836	57.344
369	NDGSGFGPSPQMQQQGNQNISPPGTGSGTRKRRERNRRPSNTAALSNNDA	163.476	97.261
370	SNQPMSESSNGQNAITNGDDASNPGNQPRKRTQRRRQRGPSGTLPNFGGA	3.273	2.949
371	GQDNDPGFSGNNPQGEQPTNTSIQSNTASLSMNSAKRRRRRGRNGGPPQA	16174.409	11539.709
372	IFPNAQDTDPGSTSGTNNNNNMEGGGLSAQGSQRNQKRPRRRRPPSSGQA	15707.192	8795.577
373	NFPNMGQASENSGIDNGDPSTTNSPPGSQLGQNRQKRRRRPARQSGGNTA	9848.032	5233.482
374	NNPNMAGANGFQGTQSNSGPGSISDESQQLPGQDRRKRRPRNRPTGTSNA	1063.322	755.303
375	MTNFETSNSSQGQPGNGGPQGNSQDSNTPLAGNRSARKRRRQRNPDPIGA	121820.409	77731.007
377	SGSFTESGNSNGQSQPMNLIGAQDNGGNQPDPARRRPTRRNKRTPGNSQA	10492.849	6629.336

378	GPQNPNEGTSGGFLAQNNNAITSPPSSDQNQGTRRS NKRRRRGDPMSGQA	2616.687	1836.891
379	SGAMQPTEDNGNAPGISGGTLPQQNPNSNSNTNRR RKRPSFRRQGDGSQA	11358.190	7967.531
380	GGNSEMNPQFTADIPQNTNGNSSGSPPQNPGTLRRD NKRRRGRQGSQASA	2501.247	2181.342
RBD	ASWFTALTQHGKEDLKFPRGQGVPINTNSSPDDQIG YYRRATRRIRGGDG	0.019	0.002
NTD	MSDNGPQNQRNAPRITFGGPSDSTGSNQNGERSGAR SKQRRPQGLPNNT	0.012	0.002
WT	MSDNGPQNQRNAPRITFGGPSDSTGSNQNGERSGAR SKQRRPQGLPNNTA	1.000	0.491

Chapter 8: Conclusions and Future Directions

In this thesis, I employed a combination of coarse-grained molecular dynamics, Monte Carlo simulations, and single-molecule fluorescence spectroscopy to investigate the behavior of the SARS-CoV-2 nucleocapsid protein and its interactions with RNA. By integrating simulations and polymer physics theory, I proposed mechanisms to understand elucidate how the nucleocapsid protein behaves in the presence of specific and non-specific RNA. To further explore these interactions, I conducted experiments to examine the binding capability of the first two domains of the nucleocapsid protein (NTD-RBD) with double-stranded RNA, specifically targeting two proposed packaging signal RNAs. These experiments revealed that the first two domains of the nucleocapsid protein bind to double stranded -RNA hairpins with low micromolar affinity. This is an order of magnitude weaker than single-stranded RNAs of similar length, and suggest that the NTD-RBD interacts non-specifically with these putative packaging signals.

To better understand the molecular basis for NTD-RBD:RNA interactions, I deployed coarse-grained molecular dynamics simulations to determine how the NTD-RBD interacts with RNA. In agreement with complementary single-molecule experiments, this work revealed that the disordered NTD forms a fuzzy complex with RNA, potentiating the binding affinity of the RBD by providing an additional multivalent surface for distributed RNA interactions. Finally, I expanded this analysis across five additional coronavirus NTD-RBD constructs with differing NTD sequences and RBD surface chemistry to ask how variable IDR sequences interact in analogous structural contexts. This work showed that despite massive variation in IDR sequence, similar modes of interaction emerged, illustrating how IDRs enable the conservation of molecular function despite substantial sequence variation.

Finally, I created a polymer reference model that can serve as a valuable tool for comparing the conformational behavior of disordered proteins. In this summary, I will provide an overview of the broader implications of my findings and suggest potential future research directions stemming from this work.

8.1 A Reference Model for Comparing Conformational Behavior of Disordered Proteins

Disordered proteins exhibit extensive conformational heterogeneity, making it challenging to describe them using characteristics of folded proteins. As a result, the use of polymer scaling laws to understand the chain dimensions of unfolded and disordered proteins has become increasingly popular. While unfolded proteins generally behave as expected for polymers in a good solvent, the behavior of disordered proteins in native conditions is more variable[572,578,579]. This variability is influenced by the underlying sequence, which introduces biases in chain dimensions due to physicochemical interactions between residues. Consequently, comparing polymer scaling characteristics of proteins under native-like conditions becomes difficult when their scaling deviates from the expected behavior.

To address this issue, we developed a sequence-specific reference model that allows for the comparison of protein chain dimensions obtained from computational or experimental metrics. I performed simulations of homopolymers of varying lengths, each composed of a single residue, and adjusted their behavior to mimic that of an ideal chain. Through these simulations, I confirmed that each protein homopolymer behaves as a Gaussian chain. Subsequently, we parameterized a model that incorporates the sequence-specific contributions of each residue to polymer scaling. This model provides a simple reference state specific to the protein under investigation, enabling comparisons to be made between the actual behavior of the protein's

dimensions and the reference model. Our findings demonstrate that sequence-specific effects in diverse disordered proteins can result in more compact or expanded conformations compared to the reference model. Moreover, this reference state can be utilized to normalize simulated disordered and folded proteins, revealing the relative compactness or expansiveness of specific subregions in relation to the reference model.

In the broader scope of polymer physics based descriptions of protein ensembles this work allows an easily implemented and interpretable description of chain dimensions to be used that can be readily understood by a general audience.

8.2 The role of Specific and Non-specific Interactions in Mediating Nucleic Acid Compaction and Phase Separation.

SARS-CoV-2, like other coronaviruses, must package its large 30kb positive sense single-stranded RNA genome. This involves compacting down a large negatively charged polymer. To that end the nucleocapsid protein, a relatively positively charged protein, serves as an effective counterion to screen repulsive RNA-RNA interactions and facilitate its compaction for packaging. The ability of the nucleocapsid protein to condense RNA is at odds with its propensity to undergo phase separation with nucleic acids. I utilized ultra coarse-grained Monte Carlo simulations and concepts from polymer physics to investigate how the nucleocapsid protein could undergo phase separation with nucleic acids or induce nucleic acid compaction dependent on the concentration regime of the macromolecules and the strength of the interactions involved. While this concept had been previously introduced by Post and Zimm[440], here I utilized simulations to apply the theory to studying RNA packaging in coronaviruses. While the model was ulta-coarse-grained such that we represented RNAs as single bead per

residue polymers and the nucleocapsid protein as two bead dimers, we were still able to generate a system that recapitulated the concentration-dependent condensation of polymers vs. phase separation put forwarded by Post and Zimm.

This topic is becoming increasingly important as disordered proteins are found to form biomolecular condensates with nucleic acids. Understanding how changes in the concentration of constituent components of these systems effects the behavior of individual molecules as well as macromolecular assemblies will be paramount for understanding the normal and dysregulated behavior of proteins.

In a broader context the molecular mechanisms underlying SARS-CoV-2 genome packaging, as well as coronaviruses at large, still remain to be elucidated. The invocation of phase separation as the mechanism of packaging because of the disordered regions the nucleocapsid protein possess overlooks the other functions that can be mediated by a disordered protein[443,635]. Phase separation also has the conundrum, mentioned previously, of needing to package single genomes in the midst of high concentrations of RNA in nucleocapsid protein driven biomolecular condensates. Here my coarse grained modeling shows that while multivalent interactions can drive phase separation, they can also drive polymer condensation.

8.3 Exploring the Ability of the Nucleocapsid Protein to Interact with howSpecific dsRNA

While our coarse-grained modeling approach used a high-affinity binding site to represent a packaging signal that interacts with the nucleocapsid protein, I have performed the initial investigations into the ability of the nucleocapsid protein to bind double-stranded RNA

(dsRNA) based on putative packaging signals. I discovered that the first two domains have relatively nonspecific interactions with dsRNA hairpins and bind with low affinity, ranging from 1 to 3 micromolar. This affinity is an order of magnitude weaker than single-stranded RNAs (ssRNA) of the same length. This was counter to our initial hypothesis that packaging signals would mediate high affinity interactions with the nucleocapsid protein. There are several potential explanations for this finding, the first of which are in regards to where and how the packaging signal can interact with the nucleocapsid protein. On one hand we have tested the first two domains which do not undergo dimerization and oligomerization like the full length protein. It is possible that the packaging signal is recognized by a dimer of the full length protein. Additionally, a different domain may be responsible for recognition as evidenced by other coronavirus packaging signal recognition motifs being localized to the C-terminal disordered domain[833].

To fully comprehend how specific RNAs influence the behavior of the nucleocapsid protein, a detailed molecular dissection of each remaining domain's capability to interact with different RNAs in isolation, along with their adjacent domains and in the context of the full-length protein, will be necessary.

8.4 Modeling Nucleocapsid Protein Single-Stranded RNA Binding with Simulations

The nucleocapsid protein is a multidomain protein with interspersed disordered and folded regions that binds RNA. An open question for the nucleocapsid protein, as well as proteins with similar folded+disordered architectures, is how their disordered and folded regions interact and modulate each others behavior. This work combines single-molecule fluorescence spectroscopy and simulations to interogate how the disorderd N-terminal domain modulates single-stranded

RNA interactions with the folded RNA binding domain. While single molecule experiments provide a high resolution characterization of protein-RNA behavior, it is challenging to dissect sequence dependent contributions to binding. To assess how the NTD modulates NTDRBD binding and what amino acids are involved I have used coarse grained simulations to model this system.

The Mpipi model used here was developed specifically for modeling disordered proteins and their interactions with RNA and has made a significant contribution to this work, allowing us to develop new pipelines for evaluating the conformational behavior of both folded and disordered proteins and their interactions with single-stranded RNA (ssRNA). As a result, I have been able to capture qualitatively the ssRNA length-dependent expansion of the first two domains of the nucleocapsid protein, the increase in binding affinity dependent on ssRNA length, and the disordered N-terminal domain (NTD)-dependent enhancement of ssRNA binding affinity.

Furthermore, I successfully replicated other experimental observations, such as the RNA binding profile of the RNA binding domain (RBD) based on NMR chemical shift perturbation experiments. Additionally, I confirmed the existence of an interaction "hotspot" in the disordered N-terminal region. In parallel, we have developed a qualitative method for assessing binding affinities in simulations, which has been able to qualitatively reproduce experimental results, and opened the possibilities of doing high throughput molecular dissections of disordered and folded protein interactions with and without RNA.

8.5 Conserved Interactions in a Disordered Region Without Sequence Conservation

While our primary focus was on understanding the behavior of the nucleocapsid protein in SARS-CoV-2, we also placed our studies within the broader context of investigating the interplay between disordered regions and adjacent folded domains. It has become increasingly apparent that disordered regions can modulate the behavior of adjacent folded domains, for example in transcription factors[834,835].

A key question that remains unanswered is how disordered regions, which often exhibit lower sequence conservation compared to their folded counterparts, can still maintain similar biological functions. To explore this, I utilized the pipeline I developed for assessing protein-protein and protein-nucleic acid interactions to compare coronavirus orthologs. This analysis revealed a conserved behavior in the disordered N-terminal domains (NTDs) despite the absence of sequence conservation. This finding is significant, as it aligns with previous observations of maintained experimental outcomes in disordered regions despite the lack of absolute sequence conservation. It demonstrates that conserved behavior can be encoded within disordered regions even in the absence of conserved sequences.

Moreover, this opens up possibilities for developing new tools that can capture the behavior of disordered regions, focusing on conserved behavior and interactions rather than solely relying on conserved sequences and physicochemical properties. Just like there is more than one way to skin a cat, there are multiple approaches to maintaining conformational behavior... and we certainly prefer the protein-friendly methods

8.6 Future Directions: Single-Molecule Characterization of Nucleocapsid dsRNA Binding

This study focused on characterizing the first two domains, NTD and RBD, of coronavirus nucleocapsid proteins. However, there are three other domains that still need to be investigated regarding their intra-protein interactions, dynamics in the presence and absence of RNA, and their ability to bind specific and nonspecific single and double-stranded RNA. Of particular interest is understanding double-stranded RNA interactions that may serve as the packaging signal for SARS-CoV-2.

Coronaviruses possess long positive-sense single-stranded genomes of approximately 30 kb. Within the cell, the genomic RNA exists in various conformations. The virus must manipulate these conformations to achieve a compact state for packaging single genomes into virions. This process needs to be relatively specific to avoid packaging host cell and subgenomic RNA. Previous studies have identified regions in the dimerization domain and C-terminal disordered region as crucial for recognizing the packaging signal. However, the specificity of this region for the packaging signal at the single-molecule level remains unexplored.

Investigating the dynamics, using nano-second fluorescence correlation spectroscopy, of the double-stranded RNA-bound state of the CTD could provide insights into whether specific sequences induce folding of the disordered tail upon binding, potentially facilitating a conformation conducive to genome-specific packaging. Simultaneously, the CTD may exhibit strong binding to specific dsRNAs, which could be characterized using single-molecule fluorescence spectroscopy. On the other hand, while it has been observed that the NTD-RBD binds single-stranded RNA with high affinity (nanomolar) and double-stranded RNA with lower

affinity (micromolar), the behavior of the full-length protein in RNA binding at the single molecule level remains open. It is possible that there are competing RNA binding domains that influence the ability of other domains to bind RNA.

The proper compaction of the single-stranded coronavirus genome is essential for packaging into the virion. Assessing the nucleocapsid protein's capacity to condense RNAs, both nonspecific and those containing potential packaging signals, would enhance our understanding of the effects of RNA sequence and structure on packaging and could provide therapeutic intervention strategies targeting specific RNA sequences. Utilizing force-based micromanipulation techniques to measure the forces exerted by the nucleocapsid protein on RNAs could be beneficial for assessing condensation. A dual optical trap could be employed to trap RNA with DNA tethers, and both full-length and truncated nucleocapsid proteins could be introduced to assess their ability to condense the RNA.

8.7 Future Directions: Computational Characterization of Structured Molecules

Generating models of dsRNA interaction would enable the interrogation of how structure impacts the mechanism of binding of the nucleocapsid protein. While an accurate description of double-stranded RNA dynamics would not be the initial objective for this type of coarse-graining, a model that utilizes elastic potentials or implements distance restraints based on NMR Nuclear Overhauser Enhancement (NOE) measurements could be used to perform simulations of dsRNA. An initial comparison with single-molecule fluorescence spectroscopy and NMR measurements could confirm if binding affinities and involved residues seen in simulations recapitulate experiment observables. This would open the possibility of high-throughput modeling of dsRNA interactions with proteins. Modeling folded domains could also benefit

from these same implementations - i.e., including a network of elastic bonds to facilitate flexibility in the folded state structure. The application of elastic networks to encode conformational flexibility in folded domains has fallen out of favor in recent years, despite substantial work in this space in the 2000s[836,837]. However, given improved accuracy in transferrable coarse-grained models, this approach could enable better descriptions of folded domain interactions with disordered regions and with nucleic acids.

8.8 Future Directions: Computational Modeling of Nucleocapsid Protein Phase Separation

The nucleocapsid protein has been shown to interact extensively with nucleic acids and has the ability to form biomolecular condensates in vitro and in cells. While the full extent of the involvement of condensates in the coronavirus lifecycle has yet to be established, a description of the residues that drive and interact in condensates could enable the ability to target condensates as an antiviral strategy, as has been done for other viruses[459,838]. Having computationally characterized the multivalent sequence-dependent contributions of binding of the first two domains of the nucleocapsid protein to RNA and confirmed these recapitulate experimental observables, there is an opportunity to investigate if and how these interactions influence condensate formation. While the importance of condensate formation for SARS-Cov-2 remains unclear, it is unambiguously true that N protein must interact with RNA for viral packaging. As such, condensate formation (or lack thereof) as a function of mutations, environmental perturbations, and small molecules offers one route to screen and systematically assess protein:RNA interaction in high throughput by visualizing perturbations that influence RNA binding via a readout that is amenable to high-content imaging.

8.9 Summary

While we have characterized the ability of the first two domains to interact with RNA, much work still remains to be done to assess how the rest of the nucleocapsid protein behaves in isolation and with specific and non-specific RNA. On a broader scale, investigating how disordered proteins can interact with various ligands in highly dynamic complexes and how they can mediate similar biological function with diverse sequences will enable a better fundamental understanding of biology as it relates to health and disease.

References

1. Mirsky, A. E. & Pauling, L. On the Structure of Native, Denatured, and Coagulated Proteins. *Proc. Natl. Acad. Sci. U. S. A.* **22**, 439–447 (1936).
2. Chen, S.-J. *et al.* Protein folds vs. protein folding: Differing questions, different challenges. *Proceedings of the National Academy of Sciences* **120**, e2214423119 (2023).
3. Juretić, D., Zucić, D., Lucić, B. & Trinajstić, N. Preference functions for prediction of membrane-buried helices in integral membrane proteins. *Comput. Chem.* **22**, 279–294 (1998).
4. Klevanik, A. V. Hydrophobicity and prediction of the secondary structure of membrane proteins and peptides. *Membr. Cell Biol.* **14**, 673–697 (2001).
5. Lomize, A. L., Pogozheva, I. D., Lomize, M. A. & Mosberg, H. I. The role of hydrophobic interactions in positioning of peripheral proteins in membranes. *BMC Struct. Biol.* **7**, 44 (2007).
6. De Marothy, M. T. & Elofsson, A. Marginally hydrophobic transmembrane α-helices shaping

membrane protein folding. *Protein Sci.* **24**, 1057–1074 (2015).
7. Dill, K. A., Alonso, D. O. & Hutchinson, K. Thermal stabilities of globular proteins. *Biochemistry* **28**, 5439–5449 (1989).
8. Wolynes, P., Luthey-Schulten, Z. & Onuchic, J. Fast-folding experiments and the topography of protein folding energy landscapes. *Chem. Biol.* **3**, 425–432 (1996).
9. Leopold, P. E., Montal, M. & Onuchic, J. N. Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 8721–8725 (1992).
10. Smeller, L. Folding superfunnel to describe cooperative folding of interacting proteins. *Proteins* **84**, 1009–1016 (2016).
11. Vendruscolo, M., Paci, E., Karplus, M. & Dobson, C. M. Structures and relative free energies of partially folded states of proteins. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 14817–14821 (2003).
12. Levinthal, C. How to fold graciously. *Mossbauer spectroscopy in*.
13. Zwanzig, R., Szabo, A. & Bagchi, B. Levinthal's paradox. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 20–22 (1992).
14. Lindorff-Larsen, K., Piana, S., Dror, R. O. & Shaw, D. E. How fast-folding proteins fold. *Science* **334**, 517–520 (2011).
15. Szczepaniak, M. *et al.* Ultrafast folding kinetics of WW domains reveal how the amino acid sequence determines the speed limit to protein folding. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 8137–8142 (2019).
16. Glyakina, A. V. & Galzitskaya, O. V. How Quickly Do Proteins Fold and Unfold, and What Structural Parameters Correlate with These Values? *Biomolecules* **10**, (2020).
17. Rumbley, J., Hoang, L., Mayne, L. & Englander, S. W. An amino acid code for protein folding. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 105–112 (2001).
18. Dill, K. A. & Chan, H. S. From Levinthal to pathways to funnels. *Nat. Struct. Biol.* **4**, 10–19 (1997).
19. Durup, J. On 'Levinthal paradox' and the theory of protein folding. *Journal of Molecular Structure: THEOCHEM* **424**, 157–169 (1998).
20. Sali, A., Shakhnovich, E. & Karplus, M. How does a protein fold? *Nature* **369**, 248–251 (1994).
21. Naganathan, A. N. Modulation of allosteric coupling by mutations: from protein dynamics and packing to altered native ensembles and function. *Curr. Opin. Struct. Biol.* **54**, 1–9 (2019).
22. Shishido, H., Yoon, J. S., Yang, Z. & Skach, W. R. CFTR trafficking mutations disrupt cotranslational protein folding by targeting biosynthetic intermediates. *Nat. Commun.* **11**, 4258 (2020).
23. Shindyalov, I. N. & Bourne, P. E. An alternative view of protein fold space. *Proteins* **38**, 247–260 (2000).
24. Yang, A. S. & Honig, B. An integrated approach to the analysis and modeling of protein sequences and structures. II. On the relationship between sequence and structural similarity for proteins that are not obviously related in sequence. *J. Mol. Biol.* **301**, 679–689 (2000).
25. Krissinel, E. On the relationship between sequence and structure similarities in proteomics. *Bioinformatics* **23**, 717–723 (2007).
26. Lau, C. K. Y. *et al.* Structural conservation despite huge sequence diversity allows EPCR binding by

the PfEMP1 family implicated in severe childhood malaria. *Cell Host Microbe* **17**, 118–129 (2015).
27. Perutz, M. F., Kendrew, J. C. & Watson, H. C. Structure and function of haemoglobin. *J. Mol. Biol.* **13**, 669–678 (1965).
28. Lesk, A. M. & Chothia, C. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.* **136**, 225–270 (1980).
29. Hill, E. E., Morea, V. & Chothia, C. Sequence conservation in families whose members have little or no sequence similarity: the four-helical cytokines and cytochromes. *J. Mol. Biol.* **322**, 205–233 (2002).
30. Sousounis, K., Haney, C. E., Cao, J., Sunchu, B. & Tsonis, P. A. Conservation of the three-dimensional structure in non-homologous or unrelated proteins. *Hum. Genomics* **6**, 10 (2012).
31. Chao, T.-H., Rekhi, S., Mittal, J. & Tabor, D. P. Data-Driven Models for Predicting Intrinsically Disordered Protein Polymer Physics Directly from Composition or Sequence. *ChemRxiv* (2023) doi:10.26434/chemrxiv-2023-wrnq1.
32. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
33. Moesa, H. A., Wakabayashi, S., Nakai, K. & Patil, A. Chemical composition is maintained in poorly conserved intrinsically disordered regions and suggests a means for their classification. *Mol. Biosyst.* **8**, 3262–3273 (2012).
34. Burger, V. M., Gurry, T. & Stultz, C. M. Intrinsically Disordered Proteins: Where Computation Meets Experiment. *Polymers* **6**, 2684–2719 (2014).
35. Hofmann, H. *et al.* Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 16155–16160 (2012).
36. Schuler, B., Soranno, A., Hofmann, H. & Nettels, D. Single-Molecule FRET Spectroscopy and the Polymer Physics of Unfolded and Intrinsically Disordered Proteins. *Annu. Rev. Biophys.* **45**, 207–231 (2016).
37. Holehouse, A. S. & Pappu, R. V. Collapse Transitions of Proteins and the Interplay Among Backbone, Sidechain, and Solvent Interactions. *Annu. Rev. Biophys.* **47**, 19–39 (2018).
38. Zhao, Y., Cortes-Huerto, R., Kremer, K. & Rudzinski, J. F. Investigating the Conformational Ensembles of Intrinsically Disordered Proteins with a Simple Physics-Based Model. *J. Phys. Chem. B* **124**, 4097–4113 (2020).
39. Shea, J.-E., Best, R. B. & Mittal, J. Physics-based computational and theoretical approaches to intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **67**, 219–225 (2021).
40. Flory, P. J. *Principles of Polymer Chemistry*. (Cornell University Press, 1953).
41. Dyson, H. J. & Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **6**, 197–208 (2005).
42. Sormanni, P. *et al.* Simultaneous quantification of protein order and disorder. *Nat. Chem. Biol.* **13**, 339–342 (2017).
43. Hsu, C. C., Buehler, M. J. & Tarakanova, A. The Order-Disorder Continuum: Linking Predictions of Protein Structure and Disorder through Molecular Simulation. *Sci. Rep.* **10**, 2068 (2020).

44. Zandi, R., Rudnick, J. & Golestanian, R. Radial distribution function of rod-like polyelectrolytes. *Eur. Phys. J. E Soft Matter* **9**, 41–46 (2002).
45. Marsh, J. A. & Forman-Kay, J. D. Sequence Determinants of Compaction in Intrinsically Disordered Proteins. *Biophys. J.* **98**, 2383–2390 (2010).
46. Lotthammer, J. M., Ginell, G. M., Griffith, D., Emenecker, R. J. & Holehouse, A. S. Direct Prediction of Intrinsically Disordered Protein Conformational Properties From Sequence. *bioRxiv* 2023.05.08.539824 (2023) doi:10.1101/2023.05.08.539824.
47. Bianchi, G., Longhi, S., Grandori, R. & Brocca, S. Relevance of Electrostatic Charges in Compactness, Aggregation, and Phase Separation of Intrinsically Disordered Proteins. *Int. J. Mol. Sci.* **21**, (2020).
48. van der Lee, R. *et al.* Classification of intrinsically disordered regions and proteins. *Chem. Rev.* **114**, 6589–6631 (2014).
49. Sacquin-Mora, S. & Prévost, C. When Order Meets Disorder: Modeling and Function of the Protein Interface in Fuzzy Complexes. *Biomolecules* **11**, (2021).
50. Munshi, S. *et al.* Tunable order-disorder continuum in protein-DNA interactions. *Nucleic Acids Res.* **46**, 8700–8709 (2018).
51. Sugase, K., Dyson, H. J. & Wright, P. E. Mechanism of coupled folding and binding of an intrinsically disordered protein. *Nature* **447**, 1021–1025 (2007).
52. Lowe, E. D. *et al.* Specificity determinants of recruitment peptides bound to phospho-CDK2/cyclin A. *Biochemistry* **41**, 15625–15634 (2002).
53. Avalos, J. L. *et al.* Structure of a Sir2 enzyme bound to an acetylated p53 peptide. *Mol. Cell* **10**, 523–535 (2002).
54. Mujtaba, S. *et al.* Structural mechanism of the bromodomain of the coactivator CBP in p53 transcriptional activation. *Mol. Cell* **13**, 251–263 (2004).
55. Wu, H. *et al.* Solution structure of a dynein motor domain associated light chain. *Nat. Struct. Biol.* **7**, 575–579 (2000).
56. Borgia, A. *et al.* Extreme disorder in an ultrahigh-affinity protein complex. *Nature* **555**, 61–66 (2018).
57. Holmbeck, S. M., Dyson, H. J. & Wright, P. E. DNA-induced conformational changes are the basis for cooperative dimerization by the DNA binding domain of the retinoid X receptor. *J. Mol. Biol.* **284**, 533–539 (1998).
58. Gearhart, M. D., Holmbeck, S. M. A., Evans, R. M., Dyson, H. J. & Wright, P. E. Monomeric complex of human orphan estrogen related receptor-2 with DNA: a pseudo-dimer interface mediates extended half-site recognition. *J. Mol. Biol.* **327**, 819–832 (2003).
59. Bjarnason, S. *et al.* DNA binding redistributes activation domain ensemble and accessibility in pioneer factor Sox2. *bioRxiv* 2023.06.16.545083 (2023) doi:10.1101/2023.06.16.545083.
60. Lunde, B. M., Moore, C. & Varani, G. RNA-binding proteins: modular design for efficient function. *Nat. Rev. Mol. Cell Biol.* **8**, 479–490 (2007).
61. Kendrew, J. C. *et al.* A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* **181**, 662–666 (1958).
62. Pauling, L., Corey, R. B. & Branson, H. R. The structure of proteins; two hydrogen-bonded helical

configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. U. S. A.* **37**, 205–211 (1951).
63. Pauling, L. & Corey, R. B. Configurations of polypeptide chains with favored orientations around single bonds: two new pleated sheets. *Proc. Natl. Acad. Sci.* **37**, 729 (1951).
64. Perutz, M. F. *et al.* Structure of hæmoglobin: a three-dimensional Fourier synthesis at 5.5-Å. resolution, obtained by X-ray analysis. *Nature* **185**, 416–422 (1960).
65. Wright, P. E. & Dyson, H. J. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **293**, 321–331 (1999).
66. Dunker, A. K., Obradovic, Z., Romero, P., Garner, E. C. & Brown, C. J. Intrinsic protein disorder in complete genomes. *Genome Inform. Ser. Workshop Genome Inform.* **11**, 161–171 (2000).
67. Uversky, V. N. & Gillespie, J. R. Why are 'natively unfolded' proteins unstructured under physiologic conditions? *Proteins: Struct. Funct. Bioinf.* (2000).
68. Forman-Kay, J. D. & Mittag, T. From sequence and forces to structure, function, and evolution of intrinsically disordered proteins. *Structure* **21**, 1492–1499 (2013).
69. Davey, N. E. The functional importance of structure in unstructured protein regions. *Curr. Opin. Struct. Biol.* **56**, 155–163 (2019).
70. Babu, M. M., Kriwacki, R. W. & Pappu, R. V. Structural biology. Versatility from protein disorder. *Science* **337**, 1460–1461 (2012).
71. Wright, P. E. & Dyson, H. J. Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* **16**, 18–29 (2014).
72. Sadowski, M. I. & Jones, D. T. The sequence–structure relationship and protein function prediction. *Curr. Opin. Struct. Biol.* **19**, 357–362 (2009).
73. Guzzo, A. V. The influence of amino-acid sequence on protein structure. *Biophys. J.* **5**, 809–822 (1965).
74. Das, R. K., Ruff, K. M. & Pappu, R. V. Relating sequence encoded information to form and function of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **32**, 102–112 (2015).
75. Mittag, T. & Forman-Kay, J. D. Atomic-level characterization of disordered protein ensembles. *Curr. Opin. Struct. Biol.* **17**, 3–14 (2007).
76. Vancraenenbroeck, R., Harel, Y. S., Zheng, W. & Hofmann, H. Polymer effects modulate binding affinities in disordered proteins. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 19506–19512 (2019).
77. Dahal, L., Kwan, T. O. C., Hollins, J. J. & Clarke, J. Promiscuous and Selective: How Intrinsically Disordered BH3 Proteins Interact with Their Pro-survival Partner MCL-1. *J. Mol. Biol.* **430**, 2468–2477 (2018).
78. Borcherds, W. *et al.* Disorder and residual helicity alter p53-Mdm2 binding affinity and signaling in cells. *Nat. Chem. Biol.* **10**, 1000–1002 (2014).
79. Zosel, F., Mercadante, D., Nettels, D. & Schuler, B. A proline switch explains kinetic heterogeneity in a coupled folding and binding reaction. *Nat. Commun.* **9**, 3332 (2018).
80. Martin, E. W. *et al.* Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science* **367**, 694–699 (2020).
81. Brucale, M., Schuler, B. & Samorì, B. Single-molecule studies of intrinsically disordered proteins. *Chem. Rev.* **114**, 3281–3317 (2014).

82. Zheng, W. & Chung, H. S. Chapter 4 - Single-molecule fluorescence studies of IDPs and IDRs. in *Intrinsically Disordered Proteins* (ed. Salvi, N.) 93–136 (Academic Press, 2019).
83. LeBlanc, S. J., Kulkarni, P. & Weninger, K. R. Single Molecule FRET: A Powerful Tool to Study Intrinsically Disordered Proteins. *Biomolecules* **8**, (2018).
84. Metskas, L. A. & Rhoades, E. Single-Molecule FRET of Intrinsically Disordered Proteins. *Annu. Rev. Phys. Chem.* (2020) doi:10.1146/annurev-physchem-012420-104917.
85. Huang, J. & MacKerell, A. D., Jr. Force field development and simulations of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **48**, 40–48 (2018).
86. Best, R. B. Computational and theoretical advances in studies of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **42**, 147–154 (2017).
87. Baker, C. M. & Best, R. B. Insights into the binding of intrinsically disordered proteins from molecular dynamics simulation. *WIREs Comput Mol Sci* **4**, 182–198 (2014).
88. Levine, Z. A. & Shea, J.-E. Simulations of disordered proteins and systems with conformational heterogeneity. *Curr. Opin. Struct. Biol.* **43**, 95–103 (2017).
89. Borgia, A. *et al.* Consistent View of Polypeptide Chain Expansion in Chemical Denaturants from Multiple Experimental Methods. *J. Am. Chem. Soc.* **138**, 11714–11726 (2016).
90. Fuertes, G. *et al.* Decoupling of size and shape fluctuations in heteropolymeric sequences reconciles discrepancies in SAXS vs. FRET measurements. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E6342–E6351 (2017).
91. Holehouse, A. S. & Sukenik, S. Controlling Structural Bias in Intrinsically Disordered Proteins Using Solution Space Scanning. *J. Chem. Theory Comput.* **16**, 1794–1805 (2020).
92. Ignacio Gutiérrez, J. *et al.* SWI/SNF senses carbon starvation with a pH-sensitive low complexity sequence. *bioRxiv* 2021.03.03.433592 (2021) doi:10.1101/2021.03.03.433592.
93. Martin, E. W. *et al.* Sequence Determinants of the Conformational Properties of an Intrinsically Disordered Protein Prior to and upon Multisite Phosphorylation. *J. Am. Chem. Soc.* **138**, 15323–15335 (2016).
94. Murthy, A. C. *et al.* Molecular interactions underlying liquid-liquid phase separation of the FUS low-complexity domain. *Nat. Struct. Mol. Biol.* **26**, 637–648 (2019).
95. Zheng, W. *et al.* Molecular Details of Protein Condensates Probed by Microsecond Long Atomistic Simulations. *J. Phys. Chem. B* (2020).
96. Leach, A. R. *Molecular modelling: principles and applications*. (Pearson education, 2001).
97. Best, R. B. Atomistic Force Fields for Proteins. *Methods Mol. Biol.* **2022**, 3–19 (2019).
98. Huang, J. *et al.* CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **14**, 71–73 (2017).
99. Robustelli, P., Piana, S. & Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E4758–E4766 (2018).
100. Piana, S., Robustelli, P., Tan, D., Chen, S. & Shaw, D. E. Development of a Force Field for the Simulation of Single-Chain Proteins and Protein-Protein Complexes. *J. Chem. Theory Comput.* **16**, 2494–2507 (2020).
101. Best, R. B. & Mittal, J. Protein simulations with an optimized water model: cooperative helix

formation and temperature-induced unfolded state collapse. *J. Phys. Chem. B* **114**, 14916–14923 (2010).
102. Best, R. B., Zheng, W. & Mittal, J. Balanced Protein-Water Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association. *J. Chem. Theory Comput.* **10**, 5113–5124 (2014).
103. Tang, W. S., Fawzi, N. L. & Mittal, J. Refining All-Atom Protein Force Fields for Polar-Rich, Prion-like, Low-Complexity Intrinsically Disordered Proteins. *J. Phys. Chem. B* (2020) doi:10.1021/acs.jpcb.0c07545.
104. Vitalis, A. & Pappu, R. V. ABSINTH: A new continuum solvation model for simulations of polypeptides in aqueous solutions. *J. Comput. Chem.* **30**, 673–699 (2009).
105. Mercadante, D. *et al.* Kirkwood-Buff Approach Rescues Overcollapse of a Disordered Protein in Canonical Protein Force Fields. *J. Phys. Chem. B* **119**, 7975–7984 (2015).
106. Ploetz, E. A., Bentenitis, N. & Smith, P. E. Developing Force Fields from the Microscopic Structure of Solutions. *Fluid Phase Equilib.* **290**, 43 (2010).
107. Saunders, M. G. & Voth, G. A. Coarse-graining methods for computational biology. *Annu. Rev. Biophys.* **42**, 73–93 (2013).
108. Ruff, K. M., Harmon, T. S. & Pappu, R. V. CAMELOT: A machine learning approach for coarse-grained simulations of aggregation of block-copolymeric protein sequences. *J. Chem. Phys.* **143**, 243123 (2015).
109. Wu, H., Wolynes, P. G. & Papoian, G. A. AWSEM-IDP: A Coarse-Grained Force Field for Intrinsically Disordered Proteins. *J. Phys. Chem. B* **122**, 11115–11125 (2018).
110. Dignon, G. L., Zheng, W., Kim, Y. C., Best, R. B. & Mittal, J. Sequence determinants of protein phase behavior from a coarse-grained model. *PLoS Comput. Biol.* **14**, e1005941 (2018).
111. Ruff, K. M., Khan, S. J. & Pappu, R. V. A coarse-grained model for polyglutamine aggregation modulated by amphipathic flanking sequences. *Biophys. J.* **107**, 1226–1235 (2014).
112. Perdikari, T. M., Jovic, N., Dignon, G. L., Kim, Y. C. & Fawzi, N. L. A coarse-grained model for position-specific effects of post-translational modifications on disordered protein phase separation. *Biophys. J.* (2021).
113. Dignon, G. L., Zheng, W., Kim, Y. C. & Mittal, J. Temperature-Controlled Liquid–Liquid Phase Separation of Disordered Proteins. *ACS Cent. Sci.* **5**, 821–830 (2019).
114. Baul, U., Chakraborty, D., Mugnai, M. L., Straub, J. E. & Thirumalai, D. Sequence Effects on Size, Shape, and Structural Heterogeneity in Intrinsically Disordered Proteins. *J. Phys. Chem. B* **123**, 3462–3474 (2019).
115. Cragnell, C., Rieloff, E. & Skepö, M. Utilizing Coarse-Grained Modeling and Monte Carlo Simulations to Evaluate the Conformational Ensemble of Intrinsically Disordered Proteins and Regions. *J. Mol. Biol.* **430**, 2478–2492 (2018).
116. Klein, F., Barrera, E. E. & Pantano, S. Assessing SIRAH's Capability to Simulate Intrinsically Disordered Proteins and Peptides. *J. Chem. Theory Comput.* **17**, 599–604 (2021).
117. Ramis, R. *et al.* A Coarse-Grained Molecular Dynamics Approach to the Study of the Intrinsically Disordered Protein α-Synuclein. *J. Chem. Inf. Model.* **59**, 1458–1471 (2019).

118. Latham, A. P. & Zhang, B. Maximum Entropy Optimized Force Field for Intrinsically Disordered Proteins. *J. Chem. Theory Comput.* **16**, 773–781 (2020).
119. Nath, A. *et al.* The conformational ensembles of α-synuclein and tau: combining single-molecule FRET and simulations. *Biophys. J.* **103**, 1940–1949 (2012).
120. O'Brien, E. P., Morrison, G., Brooks, B. R. & Thirumalai, D. How accurate are polymer models in the analysis of Forster resonance energy transfer experiments on proteins? *J. Chem. Phys.* **130**, 124903 (2009).
121. O'Brien, E. P., Ziv, G., Haran, G., Brooks, B. R. & Thirumalai, D. Effects of denaturants and osmolytes on proteins are accurately predicted by the molecular transfer model. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 13403–13408 (2008).
122. Mazouchi, A. *et al.* Conformations of a Metastable SH3 Domain Characterized by smFRET and an Excluded-Volume Polymer Model. *Biophys. J.* **110**, 1510–1522 (2016).
123. Song, J., Gomes, G.-N., Shi, T., Gradinaru, C. C. & Chan, H. S. Conformational heterogeneity and FRET data interpretation for dimensions of unfolded proteins. *Biophys. J.* **113**, (2017).
124. Song, J., Gomes, G.-N., Gradinaru, C. C. & Chan, H. S. An Adequate Account of Excluded Volume Is Necessary To Infer Compactness and Asphericity of Disordered Proteins by Forster Resonance Energy Transfer. *J. Phys. Chem. B* **119**, 15191–15202 (2015).
125. Onufriev, A. V. & Izadi, S. Water models for biomolecular simulations. *WIREs Comput Mol Sci* **8**, e1347 (2018).
126. Piana, S., Donchev, A. G., Robustelli, P. & Shaw, D. E. Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *J. Phys. Chem. B* **119**, 5113–5123 (2015).
127. Zerze, G. H., Zheng, W., Best, R. B. & Mittal, J. Evolution of All-Atom Protein Force Fields to Improve Local and Global Properties. *J. Phys. Chem. Lett.* **10**, 2227–2234 (2019).
128. Braun, E. *et al.* Best Practices for Foundations in Molecular Simulations [Article v1.0]. *Living J Comput Mol Sci* **1**, 5957 (2019).
129. Karplus, M. & McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **9**, 646–652 (2002).
130. Conicella, A. E. *et al.* TDP-43 α-helical structure tunes liquid–liquid phase separation and function. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 5883–5894 (2020).
131. Zheng, W., Borgia, A., Borgia, M. B., Schuler, B. & Best, R. B. Empirical Optimization of Interactions between Proteins and Chemical Denaturants in Molecular Simulations. *J. Chem. Theory Comput.* **11**, 5543–5553 (2015).
132. Vitalis, A. & Pappu, R. V. Chapter 3 Methods for Monte Carlo Simulations of Biomacromolecules. in *Annual Reports in Computational Chemistry* (ed. Wheeler, R. A.) vol. 5 49–76 (Elsevier, 2009).
133. Ulmschneider, J. P. & Jorgensen, W. L. Monte Carlo backbone sampling for polypeptides with variable bond angles and dihedral angles using concerted rotations and a Gaussian bias. *J. Chem. Phys.* **118**, 4261–4271 (2003).
134. Dodd, L. R., Boone, T. D. & Theodorou, D. N. A concerted rotation algorithm for atomistic Monte Carlo simulation of polymer melts and glasses. *Mol. Phys.* **78**, 961–996 (1993).

135. Favrin, G., Irbäck, A. & Sjunnesson, F. Monte Carlo update for chain molecules: Biased Gaussian steps in torsional space. *J. Chem. Phys.* **114**, 8154–8158 (2001).
136. Whitelam, S. & Geissler, P. L. Avoiding unphysical kinetic traps in Monte Carlo simulations of strongly attractive particles. *J. Chem. Phys.* **127**, 154101 (2007).
137. Gelb, L. D. Monte Carlo simulations using sampling from an approximate potential. *J. Chem. Phys.* **118**, 7747–7750 (2003).
138. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **21**, 1087–1092 (1953).
139. Hastings, W. K. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* **57**, 97–109 (1970).
140. Warner, J. B., 4th *et al.* Monomeric Huntingtin Exon 1 Has Similar Overall Structural Features for Wild-Type and Pathological Polyglutamine Lengths. *J. Am. Chem. Soc.* **139**, 14456–14469 (2017).
141. Rauscher, S. *et al.* Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment. *J. Chem. Theory Comput.* **11**, 5513–5524 (2015).
142. Henriques, J., Cragnell, C. & Skepö, M. Molecular Dynamics Simulations of Intrinsically Disordered Proteins: Force Field Evaluation and Comparison with Experiment. *J. Chem. Theory Comput.* **11**, 3420–3431 (2015).
143. Nerenberg, P. S., Jo, B., So, C., Tripathy, A. & Head-Gordon, T. Optimizing Solute–Water van der Waals Interactions To Reproduce Solvation Free Energies. *J. Phys. Chem. B* **116**, 4524–4534 (2012).
144. Mercadante, D., Wagner, J. A., Aramburu, I. V., Lemke, E. A. & Gräter, F. Sampling Long- versus Short-Range Interactions Defines the Ability of Force Fields To Reproduce the Dynamics of Intrinsically Disordered Proteins. *J. Chem. Theory Comput.* **13**, 3964–3974 (2017).
145. Grossfield, A. & Zuckerman, D. M. Quantifying uncertainty and sampling quality in biomolecular simulations. *Annu. Rep. Comput. Chem.* **5**, 23–48 (2009).
146. Grossfield, A. *et al.* Best Practices for Quantification of Uncertainty and Sampling Quality in Molecular Simulations [Article v1.0]. *Living J Comput Mol Sci* **1**, (2018).
147. Stryer, L. Fluorescence energy transfer as a spectroscopic ruler. *Annu. Rev. Biochem.* **47**, 819–846 (1978).
148. Förster, T. Zwischenmolekulare Energiewanderung und Fluoreszenz. *Ann. Phys.* **437**, 55–75 (1948).
149. Gopich, I. & Szabo, A. Theory of photon statistics in single-molecule Förster resonance energy transfer. *J. Chem. Phys.* **122**, 14707 (2005).
150. Gopich, I. V. & Szabo, A. Single-molecule FRET with diffusion and conformational dynamics. *J. Phys. Chem. B* **111**, 12925–12932 (2007).
151. Torella, J. P., Holden, S. J., Santoso, Y., Hohlbein, J. & Kapanidis, A. N. Identifying molecular dynamics in single-molecule FRET experiments with burst variance analysis. *Biophys. J.* **100**, 1568–1577 (2011).
152. Kalinin, S., Valeri, A., Antonik, M., Felekyan, S. & Seidel, C. A. M. Detection of structural dynamics by FRET: a photon distribution and fluorescence lifetime analysis of systems with multiple states. *J. Phys. Chem. B* **114**, 7983–7995 (2010).

153. Ingargiola, A., Weiss, S. & Lerner, E. Monte Carlo Diffusion-Enhanced Photon Inference: Distance Distributions and Conformational Dynamics in Single-Molecule FRET. *J. Phys. Chem. B* **122**, 11598–11615 (2018).
154. Antonik, M., Felekyan, S., Gaiduk, A. & Seidel, C. A. M. Separating structural heterogeneities from stochastic variations in fluorescence resonance energy transfer distributions via photon distribution analysis. *J. Phys. Chem. B* **110**, 6970–6978 (2006).
155. Chung, H. S., McHale, K., Louis, J. M. & Eaton, W. A. Single-molecule fluorescence experiments determine protein folding transition path times. *Science* **335**, 981–984 (2012).
156. Chung, H. S. *et al.* Extracting rate coefficients from single-molecule photon trajectories and FRET efficiency histograms for a fast-folding protein. *J. Phys. Chem. A* **115**, 3642–3656 (2011).
157. Rhoades, E., Cohen, M., Schuler, B. & Haran, G. Two-state folding observed in individual protein molecules. *J. Am. Chem. Soc.* **126**, 14686–14687 (2004).
158. Rhoades, E., Gussakovsky, E. & Haran, G. Watching proteins fold one molecule at a time. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 3197–3202 (2003).
159. Pirchi, M. *et al.* Single-molecule fluorescence spectroscopy maps the folding landscape of a large protein. *Nat. Commun.* **2**, 493 (2011).
160. Chung, H. S., Cellmer, T., Louis, J. M. & Eaton, W. A. Measuring ultrafast protein folding rates from photon-by-photon analysis of single molecule fluorescence trajectories. *Chem. Phys.* **422**, 229–237 (2013).
161. Gopich, I. V. & Szabo, A. FRET efficiency distributions of multistate single molecules. *J. Phys. Chem. B* **114**, 15221–15226 (2010).
162. Schuler, B., Lipman, E. A. & Eaton, W. A. Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. *Nature* **419**, 743 (2002).
163. Santoso, Y., Torella, J. P. & Kapanidis, A. N. Characterizing single-molecule FRET dynamics with probability distribution analysis. *Chemphyschem* **11**, 2209–2219 (2010).
164. Chung, H. S. & Gopich, I. V. Fast single-molecule FRET spectroscopy: theory and experiment. *Phys. Chem. Chem. Phys.* **16**, 18644–18657 (2014).
165. Gopich, I. V. & Szabo, A. Theory of the energy transfer efficiency and fluorescence lifetime distribution in single-molecule FRET. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 7747–7752 (2012).
166. Widengren, J. *et al.* Single-molecule detection and identification of multiple species by multiparameter fluorescence detection. *Anal. Chem.* **78**, 2039–2050 (2006).
167. Kühnemuth, R. & Seidel, C. A. M. Principles of single molecule multiparameter fluorescence spectroscopy. *Single Mol.* **2**, 251–254 (2001).
168. Soranno, A. *et al.* Quantifying internal friction in unfolded and intrinsically disordered proteins with single-molecule spectroscopy. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 17800–17806 (2012).
169. Chung, H. S., Louis, J. M. & Gopich, I. V. Analysis of Fluorescence Lifetime and Energy Transfer Efficiency in Single-Molecule Photon Trajectories of Fast-Folding Proteins. *J. Phys. Chem. B* **120**, 680–699 (2016).
170. Nettels, D., Gopich, I. V., Hoffmann, A. & Schuler, B. Ultrafast dynamics of protein collapse from single-molecule photon statistics. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 2655–2660 (2007).

171. Sindbert, S. *et al.* Accurate distance determination of nucleic acids via Förster resonance energy transfer: implications of dye linker length and rigidity. *J. Am. Chem. Soc.* **133**, 2463–2480 (2011).
172. Gopich, I. V. & Szabo, A. Single-Macromolecule Fluorescence Resonance Energy Transfer and Free-Energy Profiles. *J. Phys. Chem. B* **107**, 5058–5063 (2003).
173. Zheng, W. *et al.* Inferring properties of disordered chains from FRET transfer efficiencies. *J. Chem. Phys.* **148**, 123329 (2018).
174. Sanchez, I. C. Phase Transition Behavior of the Isolated Polymer Chain. *Macromolecules* **12**, 980–988 (1979).
175. Kratky, O. & Porod, G. Röntgenuntersuchung gelöster Fadenmoleküle. *Recl. Trav. Chim. Pays-Bas* **68**, 1106–1122 (1949).
176. Rayleigh, Lord. XXXI. On the problem of random vibrations, and of random flights in one, two, or three dimensions. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **37**, 321–347 (1919).
177. Lin, Y.-H., Brady, J. P., Chan, H. S. & Ghosh, K. A unified analytical theory of heteropolymers for sequence-specific phase behaviors of polyelectrolytes and polyampholytes. *J. Chem. Phys.* **152**, 045102 (2020).
178. Firman, T. & Ghosh, K. Sequence charge decoration dictates coil-globule transition in intrinsically disordered proteins. *J. Chem. Phys.* **148**, 123305 (2018).
179. Huihui, J., Firman, T. & Ghosh, K. Modulating charge patterning and ionic strength as a strategy to induce conformational changes in intrinsically disordered proteins. *J. Chem. Phys.* **149**, 085101 (2018).
180. Sawle, L. & Ghosh, K. A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins. *J. Chem. Phys.* **143**, 085101 (2015).
181. Lin, Y.-H., Forman-Kay, J. D. & Chan, H. S. Sequence-Specific Polyampholyte Phase Separation in Membraneless Organelles. *Phys. Rev. Lett.* **117**, 178101 (2016).
182. Ziv, G., Thirumalai, D. & Haran, G. Collapse transition in proteins. *Phys. Chem. Chem. Phys.* **11**, 83–93 (2009).
183. Sherman, E. & Haran, G. Coil–globule transition in the denatured state of a small protein. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 11539–11543 (2006).
184. Schäfer, L. *Excluded Volume Effects in Polymer Solutions: as Explained by the Renormalization Group*. (Springer Science & Business Media, 2012).
185. Soranno, A. *et al.* Integrated view of internal friction in unfolded proteins from single-molecule FRET, contact quenching, theory, and simulations. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E1833–E1839 (2017).
186. Peran, I. *et al.* Unfolded states under folding conditions accommodate sequence-specific conformational preferences with random coil-like dimensions. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 12301–12310 (2019).
187. Cubuk, J. *et al.* The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA. *Nat. Commun.* (2021).
188. Magde, D., Elson, E. & Webb, W. W. Thermodynamic fluctuations in a reacting system—

measurement by fluorescence correlation spectroscopy. *Phys. Rev. Lett.* (1972).
189. Magde, D., Elson, E. L. & Webb, W. W. Fluorescence correlation spectroscopy. II. An experimental realization. *Biopolymers* **13**, 29–61 (1974).
190. Elson, E. L. & Magde, D. Fluorescence correlation spectroscopy. I. Conceptual basis and theory. *Biopolymers: Original Research on* (1974).
191. Axelrod, D., Koppel, D. E., Schlessinger, J., Elson, E. & Webb, W. W. Mobility measurement by analysis of fluorescence photobleaching recovery kinetics. *Biophys. J.* **16**, 1055–1069 (1976).
192. Elson, E. L. Fluorescence correlation spectroscopy: past, present, future. *Biophys. J.* **101**, 2855–2870 (2011).
193. Crick, S. L., Jayaraman, M., Frieden, C., Wetzel, R. & Pappu, R. V. Fluorescence correlation spectroscopy shows that monomeric polyglutamine molecules form collapsed structures in aqueous solutions. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 16764–16769 (2006).
194. Neuweiler, H., Löllmann, M., Doose, S. & Sauer, M. Dynamics of unfolded polypeptide chains in crowded environment studied by fluorescence correlation spectroscopy. *J. Mol. Biol.* **365**, 856–869 (2007).
195. Zosel, F., Haenni, D., Soranno, A., Nettels, D. & Schuler, B. Combining short- and long-range fluorescence reporters with simulations to explore the intramolecular dynamics of an intrinsically disordered protein. *J. Chem. Phys.* **147**, 152708 (2017).
196. Doose, S., Neuweiler, H. & Sauer, M. Fluorescence quenching by photoinduced electron transfer: a reporter for conformational dynamics of macromolecules. *Chemphyschem* **10**, 1389–1398 (2009).
197. Gopich, I. V., Nettels, D., Schuler, B. & Szabo, A. Protein dynamics from single-molecule fluorescence intensity correlation functions. *J. Chem. Phys.* **131**, 095102 (2009).
198. Meng, F. *et al.* Highly Disordered Amyloid-β Monomer Probed by Single-Molecule FRET and MD Simulation. *Biophys. J.* **114**, 870–884 (2018).
199. Soranno, A., Zosel, F. & Hofmann, H. Internal friction in an intrinsically disordered protein—Comparing Rouse-like models with experiments. *J. Chem. Phys.* **148**, 123326 (2018).
200. Khatri, B. S. & McLeish, T. C. B. Rouse Model with Internal Friction: A Coarse Grained Framework for Single Biopolymer Dynamics. *Macromolecules* **40**, 6770–6777 (2007).
201. Cheng, R. R., Hawk, A. T. & Makarov, D. E. Exploring the role of internal friction in the dynamics of unfolded proteins using simple polymer models. *J. Chem. Phys.* **138**, 074112 (2013).
202. de Gennes, P.-G. & Gennes, P.-G. *Scaling Concepts in Polymer Physics*. (Cornell University Press, 1979).
203. Doi, M., Edwards, S. F. & Edwards, S. F. *The Theory of Polymer Dynamics*. (Clarendon Press, 1988).
204. Sauer, M. & Neuweiler, H. PET-FCS: probing rapid structural fluctuations of proteins and nucleic acids by single-molecule fluorescence quenching. *Methods Mol. Biol.* **1076**, 597–615 (2014).
205. Doose, S., Neuweiler, H. & Sauer, M. A close look at fluorescence quenching of organic dyes by tryptophan. *Chemphyschem* **6**, 2277–2285 (2005).
206. Wang, Z. & Makarov, D. E. Nanosecond Dynamics of Single Polypeptide Molecules Revealed by Photoemission Statistics of Fluorescence Resonance Energy Transfer: A Theoretical Study. *J. Phys. Chem. B* **107**, 5617–5622 (2003).
207. Makarov, D. E. Spatiotemporal correlations in denatured proteins: The dependence of fluorescence

resonance energy transfer (FRET)-derived protein reconfiguration times on the location of the FRET probes. *J. Chem. Phys.* **132**, 035104 (2010).
208. Szabo, A., Schulten, K. & Schulten, Z. First passage time approach to diffusion controlled reactions. *J. Chem. Phys.* **72**, 4350–4357 (1980).
209. Hellenkamp, B. *et al.* Precision and accuracy of single-molecule FRET measurements-a multi-laboratory benchmark study. *Nat. Methods* **15**, 669–676 (2018).
210. Kudryavtsev, V. *et al.* Combining MFD and PIE for accurate single-pair Förster resonance energy transfer measurements. *Chemphyschem* **13**, 1060–1078 (2012).
211. Lee, N. K. *et al.* Accurate FRET measurements within single diffusing biomolecules using alternating-laser excitation. *Biophys. J.* **88**, 2939–2953 (2005).
212. Holmstrom, E. D. *et al.* Accurate Transfer Efficiencies, Distance Distributions, and Ensembles of Unfolded and Intrinsically Disordered Proteins From Single-Molecule FRET. *Methods Enzymol.* **611**, 287–325 (2018).
213. Kapanidis, A. N. *et al.* Alternating-laser excitation of single molecules. *Acc. Chem. Res.* **38**, 523–533 (2005).
214. Hohlbein, J., Craggs, T. D. & Cordes, T. Alternating-laser excitation: single-molecule FRET and beyond. *Chem. Soc. Rev.* **43**, 1156–1171 (2014).
215. Müller, B. K., Zaychikov, E., Bräuchle, C. & Lamb, D. C. Pulsed interleaved excitation. *Biophys. J.* **89**, 3508–3522 (2005).
216. Hendrix, J. & Lamb, D. C. Pulsed interleaved excitation: principles and applications. *Methods Enzymol.* **518**, 205–243 (2013).
217. Niaki, A. G. *et al.* Loss of Dynamic RNA Interaction and Aberrant Phase Separation Induced by Two Distinct Types of ALS/FTD-Linked FUS Mutations. *Mol. Cell* **77**, 82–94.e4 (2020).
218. Holden, S. J. *et al.* Defining the limits of single-molecule FRET resolution in TIRF microscopy. *Biophys. J.* **99**, 3102–3111 (2010).
219. Chung, H. S., Piana-Agostinetti, S., Shaw, D. E. & Eaton, W. A. Structural origin of slow diffusion in protein folding. *Science* **349**, 1504–1510 (2015).
220. Levene, M. J. *et al.* Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299**, 682–686 (2003).
221. Zhu, P. & Craighead, H. G. Zero-mode waveguides for single-molecule analysis. *Annu. Rev. Biophys.* **41**, 269–293 (2012).
222. Polinkovsky, M. E. *et al.* Ultrafast cooling reveals microsecond-scale biomolecular dynamics. *Nat. Commun.* **5**, 5737 (2014).
223. Zhi, Z., Liu, P., Wang, P., Huang, Y. & Zhao, X. S. Domain-specific folding kinetics of staphylococcal nuclease observed through single-molecule FRET in a microfluidic mixer. *Chemphyschem* **12**, 3515–3518 (2011).
224. Lemke, E. A. *et al.* Microfluidic device for single-molecule experiments with enhanced photostability. *J. Am. Chem. Soc.* **131**, 13610–13612 (2009).
225. Horrocks, M. H. *et al.* Single-molecule measurements of transient biomolecular complexes through microfluidic dilution. *Anal. Chem.* **85**, 6855–6859 (2013).

226. Tyagi, S. *et al.* Continuous throughput and long-term observation of single-molecule FRET without immobilization. *Nat. Methods* **11**, 297–300 (2014).
227. Benke, S., Nettels, D., Hofmann, H. & Schuler, B. Quantifying kinetics from time series of single-molecule Förster resonance energy transfer efficiency histograms. *Nanotechnology* **28**, 114002 (2017).
228. Segal, M. *et al.* High-throughput smFRET analysis of freely diffusing nucleic acid molecules and associated proteins. *Methods* **169**, 21–45 (2019).
229. Hoffmann, A. *et al.* Quantifying heterogeneity and conformational dynamics from single molecule FRET of diffusing molecules: recurrence analysis of single particles (RASP). *Phys. Chem. Chem. Phys.* **13**, 1857–1871 (2011).
230. König, I. *et al.* Single-molecule spectroscopy of protein conformational dynamics in live eukaryotic cells. *Nat. Methods* **12**, 773–779 (2015).
231. Müller-Späth, S. *et al.* Charge interactions can dominate the dimensions of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 14609–14614 (2010).
232. Schuler, B., Lipman, E. A., Steinbach, P. J., Kumke, M. & Eaton, W. A. Polyproline and the 'spectroscopic ruler' revisited with single-molecule fluorescence. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 2754–2759 (2005).
233. Aznauryan, M. *et al.* Comprehensive structural and dynamical view of an unfolded protein from the combination of single-molecule FRET, NMR, and SAXS. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E5389–98 (2016).
234. Nettels, D. *et al.* Single-molecule spectroscopy of the temperature-induced collapse of unfolded proteins. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 20740–20745 (2009).
235. McCarney, E. R. *et al.* Site-specific dimensions across a highly denatured protein; a single molecule study. *J. Mol. Biol.* **352**, 672–682 (2005).
236. Best, R. B. *et al.* Effect of flexibility and cis residues in single-molecule FRET studies of polyproline. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 18964–18969 (2007).
237. Dimura, M. *et al.* Quantitative FRET studies and integrative modeling unravel the structure and dynamics of biomolecular systems. *Curr. Opin. Struct. Biol.* **40**, 163–185 (2016).
238. Dimura, M. *et al.* Automated and optimally FRET-assisted structural modeling. *Nat. Commun.* **11**, 5394 (2020).
239. Renault, K., Fredy, J. W., Renard, P.-Y. & Sabot, C. Covalent Modification of Biomolecules through Maleimide-Based Labeling Strategies. *Bioconjug. Chem.* **29**, 2497–2513 (2018).
240. Teng, F.-Y. *et al.* A Toolbox for Site-Specific Labeling of RecQ Helicase With a Single Fluorophore Used in the Single-Molecule Assay. *Front Mol Biosci* **7**, 586450 (2020).
241. Chen, L. *et al.* Improved variants of SrtA for site-specific conjugation on antibodies and proteins with high efficiency. *Sci. Rep.* **6**, 31899 (2016).
242. Levary, D. A., Parthasarathy, R., Boder, E. T. & Ackerman, M. E. Protein-protein fusion catalyzed by sortase A. *PLoS One* **6**, e18342 (2011).
243. Popp, M. W.-L. Site-specific labeling of proteins via sortase: protocols for the molecular biologist. *Methods Mol. Biol.* **1266**, 185–198 (2015).

244. Sarpong, K. & Bose, R. Efficient sortase-mediated N-terminal labeling of TEV protease cleaved recombinant proteins. *Anal. Biochem.* **521**, 55–58 (2017).
245. Lee, T. C., Moran, C. R., Cistrone, P. A., Dawson, P. E. & Deniz, A. A. Site-Specific Three-Color Labeling of α-Synuclein via Conjugation to Uniquely Reactive Cysteines during Assembly by Native Chemical Ligation. *Cell Chem Biol* **25**, 797–801.e4 (2018).
246. Yang, J.-Y. & Yang, W. Y. Site-specific two-color protein labeling for FRET studies using split inteins. *J. Am. Chem. Soc.* **131**, 11644–11645 (2009).
247. Lai, W.-J. C. & Ermolenko, D. N. Ensemble and single-molecule FRET studies of protein synthesis. *Methods* **137**, 37–48 (2018).
248. Chinnaraj, M. *et al.* Bioorthogonal Chemistry Enables Single-Molecule FRET Measurements of Catalytically Active Protein Disulfide Isomerase. *Chembiochem* **22**, 134–138 (2021).
249. Lee, N. K. *et al.* Three-color alternating-laser excitation of single molecules: monitoring multiple interactions and distances. *Biophys. J.* **92**, 303–312 (2007).
250. Barth, A., Seidel, C. A. M. & Lamb, D. C. Studying Complex Biomolecular Dynamics by Single-Molecule Three-Color FRET. *Biophysical Journal* vol. 116 476a–477a Preprint at https://doi.org/10.1016/j.bpj.2018.11.2574 (2019).
251. Yoo, J., Louis, J. M., Gopich, I. V. & Chung, H. S. Three-Color Single-Molecule FRET and Fluorescence Lifetime Analysis of Fast Protein Folding. *J. Phys. Chem. B* **122**, 11702–11720 (2018).
252. Hohng, S., Joo, C. & Ha, T. Single-molecule three-color FRET. *Biophys. J.* **87**, 1328–1337 (2004).
253. Clamme, J.-P. & Deniz, A. A. Three-color single-molecule fluorescence resonance energy transfer. *Chemphyschem* **6**, 74–77 (2005).
254. Lin, C.-W. & Ting, A. Y. Transglutaminase-catalyzed site-specific conjugation of small-molecule probes to proteins in vitro and on the surface of living cells. *J. Am. Chem. Soc.* **128**, 4542–4543 (2006).
255. Lu, M. *et al.* Real-Time Conformational Dynamics of SARS-CoV-2 Spikes on Virus Particles. *Cell Host Microbe* **28**, 880–891.e8 (2020).
256. Yin, J., Lin, A. J., Golan, D. E. & Walsh, C. T. Site-specific protein labeling by Sfp phosphopantetheinyl transferase. *Nat. Protoc.* **1**, 280–285 (2006).
257. Lu, M. *et al.* Shedding-Resistant HIV-1 Envelope Glycoproteins Adopt Downstream Conformations That Remain Responsive to Conformation-Preferring Ligands. *J. Virol.* **94**, (2020).
258. Lu, M. *et al.* Associating HIV-1 envelope glycoprotein structures with states on the virus observed by smFRET. *Nature* **568**, 415–419 (2019).
259. Mofid, M. R., Finking, R. & Marahiel, M. A. Recognition of hybrid peptidyl carrier proteins/acyl carrier proteins in nonribosomal peptide synthetase modules by the 4'-phosphopantetheinyl transferases AcpS and Sfp. *J. Biol. Chem.* **277**, 17023–17031 (2002).
260. Flugel, R. S., Hwangbo, Y., Lambalot, R. H., Cronan, J. E., Jr & Walsh, C. T. Holo-(acyl carrier protein) synthase and phosphopantetheinyl transfer in Escherichia coli. *J. Biol. Chem.* **275**, 959–968 (2000).

261. Gehring, A. M., Lambalot, R. H., Vogel, K. W., Drueckhammer, D. G. & Walsh, C. T. Ability of Streptomyces spp. acyl carrier proteins and coenzyme A analogs to serve as substrates in vitro for E. coli holo-ACP synthase. *Chem. Biol.* **4**, 17–24 (1997).
262. Riback, J. A. *et al.* Commonly used FRET fluorophores promote collapse of an otherwise disordered protein. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 8889–8894 (2019).
263. Watkins, H. M. *et al.* Random coil negative control reproduces the discrepancy between scattering and FRET measurements of denatured protein dimensions. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 6631–6636 (2015).
264. Fuertes, G. *et al.* Comment on 'Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water'. *Science* **361**, (2018).
265. Riback, J. A. *et al.* Response to Comment on 'Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water'. *Science* vol. 361 (2018).
266. Zerze, G. H., Best, R. B. & Mittal, J. Modest influence of FRET chromophores on the properties of unfolded proteins. *Biophys. J.* **107**, 1654–1660 (2014).
267. Gomes, G.-N. W. *et al.* Conformational Ensembles of an Intrinsically Disordered Protein Consistent with NMR, SAXS, and Single-Molecule FRET. *J. Am. Chem. Soc.* **142**, 15697–15710 (2020).
268. Zhang, Z., Yomo, D. & Gradinaru, C. Choosing the right fluorophore for single-molecule fluorescence studies in a lipid environment. *Biochim. Biophys. Acta Biomembr.* **1859**, 1242–1253 (2017).
269. Sisamakis, E., Valeri, A., Kalinin, S., Rothwell, P. J. & Seidel, C. A. M. Accurate single-molecule FRET studies using multiparameter fluorescence detection. *Methods Enzymol.* **475**, 455–514 (2010).
270. Schröder, G. F., Alexiev, U. & Grubmüller, H. Simulation of Fluorescence Anisotropy Experiments: Probing Protein Dynamics. *Biophysical Journal* vol. 89 3757–3770 Preprint at https://doi.org/10.1529/biophysj.105.069500 (2005).
271. Zheng, W. *et al.* Probing the action of chemical denaturant on an intrinsically disordered protein by simulation and experiment. *J. Am. Chem. Soc.* **138**, 11702–11713 (2016).
272. Mao, A. H., Crick, S. L., Vitalis, A., Chicoine, C. L. & Pappu, R. V. Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 8183–8188 (2010).
273. Fleming, P. J. & Fleming, K. G. HullRad: Fast Calculations of Folded and Disordered Protein and Nucleic Acid Hydrodynamic Properties. *Biophys. J.* **114**, 856–869 (2018).
274. Nygaard, M., Kragelund, B. B., Papaleo, E. & Lindorff-Larsen, K. An Efficient Method for Estimating the Hydrodynamic Radius of Disordered Protein Conformations. *Biophys. J.* **113**, 550–557 (2017).
275. Rezaei-Ghaleh, N. *et al.* Local and Global Dynamics in Intrinsically Disordered Synuclein. *Angew. Chem. Int. Ed.* **57**, 15262–15266 (2018).
276. Echeverria, I., Makarov, D. E. & Papoian, G. A. Concerted dihedral rotations give rise to internal friction in unfolded proteins. *J. Am. Chem. Soc.* **136**, 8708–8713 (2014).

277. Best, R. B., Hofmann, H., Nettels, D. & Schuler, B. Quantitative interpretation of FRET experiments via molecular simulation: force field and validation. *Biophys. J.* **108**, 2721–2731 (2015).
278. Walczewska-Szewc, K. & Corry, B. Accounting for dye diffusion and orientation when relating FRET measurements to distances: three simple computational methods. *Phys. Chem. Chem. Phys.* **16**, 12317–12326 (2014).
279. Kalinin, S. *et al.* A toolkit and benchmark study for FRET-restrained high-precision structural modeling. *Nat. Methods* **9**, 1218–1225 (2012).
280. Woźniak, A. K., Schröder, G. F., Grubmüller, H., Seidel, C. A. M. & Oesterhelt, F. Single-molecule FRET measures bends and kinks in DNA. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 18337–18342 (2008).
281. Lehmann, K. *et al.* Dynamics of the nucleosomal histone H3 N-terminal tail revealed by high precision single-molecule FRET. *Nucleic Acids Res.* **48**, 1551–1571 (2020).
282. Craggs, T. D. & Kapanidis, A. N. Six steps closer to FRET-driven structural biology. *Nature methods* vol. 9 1157–1158 (2012).
283. Zimmerman, M. I. *et al.* Citizen Scientists Create an Exascale Computer to Combat COVID-19. *bioRxiv* (2020) doi:10.1101/2020.06.27.175430.
284. Shirts, M. & Pande, V. S. Screen Savers of the World Unite! *Science* **290**, 1903–1904 (2000).
285. Shaw, D. E. *et al.* Anton, a Special-purpose Machine for Molecular Dynamics Simulation. *Commun. ACM* **51**, 91–97 (2008).
286. Jaynes, E. T. Information Theory and Statistical Mechanics. *Phys. Rev.* **106**, 620–630 (1957).
287. Bonomi, M., Heller, G. T., Camilloni, C. & Vendruscolo, M. Principles of protein structural ensemble determination. *Curr. Opin. Struct. Biol.* **42**, 106–116 (2017).
288. Boomsma, W., Ferkinghoff-Borg, J. & Lindorff-Larsen, K. Combining experiments and simulations using the maximum entropy principle. *PLoS Comput. Biol.* **10**, e1003406 (2014).
289. Hummer, G. & Köfinger, J. Bayesian ensemble refinement by replica simulations and reweighting. *J. Chem. Phys.* **143**, 243150 (2015).
290. Köfinger, J., Różycki, B. & Hummer, G. Inferring Structural Ensembles of Flexible and Dynamic Macromolecules Using Bayesian, Maximum Entropy, and Minimal-Ensemble Refinement Methods. *Methods Mol. Biol.* **2022**, 341–352 (2019).
291. Lincoff, J. *et al.* Extended experimental inferential structure determination method in determining the structural ensembles of disordered protein states. *Communications Chemistry* **3**, 74 (2020).
292. Bottaro, S., Bengtsen, T. & Lindorff-Larsen, K. Integrating Molecular Simulation and Experimental Data: A Bayesian/Maximum Entropy Reweighting Approach. *Methods Mol. Biol.* **2112**, 219–240 (2020).
293. Crehuet, R., Buigues, P. J., Salvatella, X. & Lindorff-Larsen, K. Bayesian-Maximum-Entropy Reweighting of IDP Ensembles Based on NMR Chemical Shifts. *Entropy* **21**, 898 (2019).
294. Leung, H. T. A. *et al.* A Rigorous and Efficient Method To Reweight Very Large Conformational Ensembles Using Average Experimental Data and To Determine Their Relative Information Content. *J. Chem. Theory Comput.* **12**, 383–394 (2016).

295. Larsen, A. H. *et al.* Combining molecular dynamics simulations with small-angle X-ray and neutron scattering data to study multi-domain proteins in solution. *PLoS Comput. Biol.* **16**, e1007870 (2020).
296. Camilloni, C., Langkilde, A. E. & Lindorff-Larsen, K. Refinement of α-synuclein ensembles against SAXS data: Comparison of force fields and methods. *bioRxiv* (2021).
297. Gaalswyk, K., Muniyat, M. I. & MacCallum, J. L. The emerging role of physical modeling in the future of structure determination. *Curr. Opin. Struct. Biol.* **49**, 145–153 (2018).
298. Morrone, J. A., Brini, E. & MacCallum, J. L. Blind protein structure prediction using accelerated free-energy simulations. *Science* (2016).
299. Perez, A., MacCallum, J. L. & Dill, K. A. Accelerating molecular simulations of proteins using Bayesian inference on weak information. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 11846–11851 (2015).
300. MacCallum, J. L., Perez, A. & Dill, K. A. Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 6985–6990 (2015).
301. Rangan, R. *et al.* Determination of Structural Ensembles of Proteins: Restraining vs Reweighting. *J. Chem. Theory Comput.* **14**, 6632–6641 (2018).
302. Ghosh, K., Dixit, P. D., Agozzino, L. & Dill, K. A. The Maximum Caliber Variational Principle for Nonequilibria. *Annu. Rev. Phys. Chem.* **71**, 213–238 (2020).
303. Deniz, A. A. *et al.* Single-molecule protein folding: diffusion fluorescence resonance energy transfer studies of the denaturation of chymotrypsin inhibitor 2. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 5179–5184 (2000).
304. Merchant, K. A., Best, R. B., Louis, J. M., Gopich, I. V. & Eaton, W. A. Characterizing the unfolded states of proteins using single-molecule FRET spectroscopy and molecular simulations. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 1528–1533 (2007).
305. Thirumalai, D., Samanta, H. S., Maity, H. & Reddy, G. Universal Nature of Collapsibility in the Context of Protein Folding and Evolution. *Trends Biochem. Sci.* **44**, 675–687 (2019).
306. Holehouse, A. S., Garai, K., Lyle, N., Vitalis, A. & Pappu, R. V. Quantitative assessments of the distinct contributions of polypeptide backbone amides versus side chain groups to chain expansion via chemical denaturation. *J. Am. Chem. Soc.* **137**, 2984–2995 (2015).
307. Best, R. B. Emerging consensus on the collapse of unfolded and intrinsically disordered proteins in water. *Curr. Opin. Struct. Biol.* **60**, 27–38 (2020).
308. Riback, J. A. *et al.* Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water. *Science* **358**, 238–241 (2017).
309. Yoo, T. Y. *et al.* Small-Angle X-ray Scattering and Single-Molecule FRET Spectroscopy Produce Highly Divergent Views of the Low-Denaturant Unfolded State. *J. Mol. Biol.* **418**, 226–236 (2012).
310. Bowman, M. A. *et al.* Properties of protein unfolded states suggest broad selection for expanded conformational ensembles. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 23356–23364 (2020).
311. Moses, D. *et al.* Revealing the Hidden Sensitivity of Intrinsically Disordered Proteins to their Chemical Environment. *J. Phys. Chem. Lett.* 10131–10136 (2020).

312. Clark, P. L., Plaxco, K. W. & Sosnick, T. R. Water as a Good Solvent for Unfolded Proteins: Folding and Collapse are Fundamentally Different. *J. Mol. Biol.* **432**, 2882–2889 (2020).
313. Diehl, R. C., Guinn, E. J., Capp, M. W., Tsodikov, O. V. & Record, M. T., Jr. Quantifying additive interactions of the osmolyte proline with individual functional groups of proteins: comparisons with urea and glycine betaine, interpretation of m-values. *Biochemistry* **52**, 5997–6010 (2013).
314. Record, M. T., Jr, Anderson, C. F. & Lohman, T. M. Thermodynamic analysis of ion effects on the binding and conformational equilibria of proteins and nucleic acids: the roles of ion association or release, screening, and ion effects on water activity. *Q. Rev. Biophys.* **11**, 103–178 (1978).
315. Wang, Y., Sukenik, S., Davis, C. M. & Gruebele, M. Cell Volume Controls Protein Stability and Compactness of the Unfolded State. *J. Phys. Chem. B* **122**, 11762–11770 (2018).
316. Sukenik, S., Salam, M., Wang, Y. & Gruebele, M. In-Cell Titration of Small Solutes Controls Protein Stability and Aggregation. *J. Am. Chem. Soc.* **140**, 10497–10503 (2018).
317. Ruff, K. M. & Holehouse, A. S. SAXS versus FRET: A Matter of Heterogeneity? *Biophys. J.* **113**, 971–973 (2017).
318. Stenzoski, N. E. *et al.* The Cold-Unfolded State Is Expanded but Contains Long- and Medium-Range Contacts and Is Poorly Described by Homopolymer Models. *Biochemistry* **59**, 3290–3299 (2020).
319. Das, R. K. & Pappu, R. V. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 13392–13397 (2013).
320. Metskas, L. A. & Rhoades, E. Conformation and Dynamics of the Troponin I C-Terminal Domain: Combining Single-Molecule and Computational Approaches for a Disordered Protein Region. *J. Am. Chem. Soc.* **137**, 11962–11969 (2015).
321. Medina, E. *et al.* Intrinsically Disordered Regions of the DNA-Binding Domain of Human FoxP1 Facilitate Domain Swapping. *J. Mol. Biol.* **432**, 5411–5429 (2020).
322. Gopich, I. V. & Szabo, A. Decoding the pattern of photon colors in single-molecule FRET. *J. Phys. Chem. B* **113**, 10965–10973 (2009).
323. Heidarsson, P. O. *et al.* Disordered Proteins Enable Histone Chaperoning on the Nucleosome. *bioRxiv* 2020.04.17.046243 (2020) doi:10.1101/2020.04.17.046243.
324. Holmstrom, E. D., Liu, Z., Nettels, D., Best, R. B. & Schuler, B. Disordered RNA chaperones can enhance nucleic acid folding via local charge screening. *Nat. Commun.* **10**, 2453 (2019).
325. Alshareedah, I., Moosa, M. M., Raju, M., Potoyan, D. A. & Banerjee, P. R. Phase transition of RNA−protein complexes into ordered hollow condensates. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 15650–15658 (2020).
326. Kaur, T. *et al.* Sequence-encoded and composition-dependent protein-RNA interactions control multiphasic condensate morphologies. *Nat. Commun.* **12**, 872 (2021).
327. Wente, S. R. & Rout, M. P. The nuclear pore complex and nuclear transport. *Cold Spring Harb. Perspect. Biol.* **2**, a000562 (2010).
328. Görlich, D. & Kutay, U. Transport between the cell nucleus and the cytoplasm. *Annu. Rev. Cell Dev. Biol.* **15**, 607–660 (1999).

329. Görlich, D. & Mattaj, I. W. Nucleocytoplasmic transport. *Science* (1996).
330. Schmidt, H. B. & Görlich, D. Transport Selectivity of Nuclear Pores, Phase Separation, and Membraneless Organelles. *Trends Biochem. Sci.* **41**, 46–61 (2016).
331. Milles, S. *et al.* Plasticity of an ultrafast interaction between nucleoporins and nuclear transport receptors. *Cell* **163**, 734–745 (2015).
332. Wuttke, R. *et al.* Temperature-dependent solvation modulates the dimensions of disordered proteins. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 5213–5218 (2014).
333. Drubin, D. G. & Kirschner, M. W. Tau protein function in living cells. *J. Cell Biol.* **103**, 2739–2746 (1986).
334. Weingarten, M. D., Lockwood, A. H., Hwo, S. Y. & Kirschner, M. W. A protein factor essential for microtubule assembly. *Proc. Natl. Acad. Sci. U. S. A.* **72**, 1858–1862 (1975).
335. Melo, A. M. *et al.* A functional role for intrinsic disorder in the tau-tubulin complex. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 14336–14341 (2016).
336. Nerenberg, P. S. & Head-Gordon, T. New developments in force fields for biomolecular simulations. *Curr. Opin. Struct. Biol.* **49**, 129–138 (2018).
337. Shi, Y. *et al.* The Polarizable Atomic Multipole-based AMOEBA Force Field for Proteins. *J. Chem. Theory Comput.* **9**, 4046–4063 (2013).
338. Rackers, J. A. *et al.* An optimized charge penetration model for use with the AMOEBA force field. *Phys. Chem. Chem. Phys.* **19**, 276–291 (2016).
339. Lemkul, J. A., Huang, J., Roux, B. & MacKerell, A. D., Jr. An Empirical Polarizable Force Field Based on the Classical Drude Oscillator Model: Development History and Recent Applications. *Chem. Rev.* **116**, 4983–5013 (2016).
340. Lagardère, L. *et al.* Tinker-HP: a massively parallel molecular dynamics package for multiscale simulations of large complex systems with advanced point dipole polarizable force fields. *Chem. Sci.* **9**, 956–972 (2018).
341. Inizan, T. J. *et al.* High-resolution mining of the SARS-CoV-2 main protease conformational space: supercomputer-driven unsupervised adaptive sampling. *Chem. Sci.* (2021) doi:10.1039/D1SC00145K.
342. Brown, S. & Head-Gordon, T. Cool walking: a new Markov chain Monte Carlo sampling method. *J. Comput. Chem.* **24**, 68–76 (2003).
343. Lincoff, J., Sasmal, S. & Head-Gordon, T. The combined force field-sampling problem in simulations of disordered amyloid-β peptides. *J. Chem. Phys.* **150**, 104108 (2019).
344. Henzler-Wildman, K. & Kern, D. Dynamic personalities of proteins. *Nature* **450**, 964–972 (2007).
345. Netherton, C. L. & Wileman, T. Virus factories, double membrane vesicles and viroplasm generated in animal cells. *Curr. Opin. Virol.* **1**, 381–387 (2011).
346. Wolff, G., Melia, C. E., Snijder, E. J. & Bárcena, M. Double-Membrane Vesicles as Platforms for Viral Replication. *Trends Microbiol.* **28**, 1022–1033 (2020).
347. García-Sastre, A. Ten Strategies of Interferon Evasion by Viruses. *Cell Host Microbe* **22**, 176–184 (2017).
348. Vale-Costa, S. & Amorim, M. J. Recycling Endosomes and Viral Infection. *Viruses* **8**, 64 (2016).

349. de Castro, I. F., Volonté, L. & Risco, C. Virus factories: biogenesis and structural design. *Cell. Microbiol.* **15**, 24–34 (2013).
350. Fernández de Castro, I., Tenorio, R. & Risco, C. Virus assembly factories in a lipid world. *Curr. Opin. Virol.* **18**, 20–26 (2016).
351. Wang, L. *et al.* Development of Small-Molecule Inhibitors Against Zika Virus Infection. *Front. Microbiol.* **10**, 2725 (2019).
352. Shin, Y. & Brangwynne, C. P. Liquid phase condensation in cell physiology and disease. *Science* **357**, (2017).
353. Vandelli, A., Vocino, G. & Tartaglia, G. G. Phase Separation Drives SARS-CoV-2 Replication: A Hypothesis. *Front Mol Biosci* **9**, 893067 (2022).
354. Wang, S. *et al.* Targeting liquid-liquid phase separation of SARS-CoV-2 nucleocapsid protein promotes innate antiviral immunity by elevating MAVS activity. *Nat. Cell Biol.* **23**, 718–732 (2021).
355. Yu, X. *et al.* The STING phase-separator suppresses innate immune signalling. *Nat. Cell Biol.* **23**, 330–340 (2021).
356. Said, E. A., Tremblay, N., Al-Balushi, M. S., Al-Jabri, A. A. & Lamarre, D. Viruses Seen by Our Cells: The Role of Viral RNA Sensors. *J Immunol Res* **2018**, 9480497 (2018).
357. Banani, S. F., Lee, H. O., Hyman, A. A. & Rosen, M. K. Biomolecular condensates: organizers of cellular biochemistry. *Nat. Rev. Mol. Cell Biol.* **18**, 285–298 (2017).
358. Mittag, T. & Pappu, R. V. A conceptual framework for understanding phase separation and addressing open questions and challenges. *Mol. Cell* **82**, 2201–2214 (2022).
359. Choi, J.-M., Holehouse, A. S. & Pappu, R. V. Physical Principles Underlying the Complex Biology of Intracellular Phase Transitions. *Annu. Rev. Biophys.* **49**, 107–133 (2020).
360. Soranno, A. The Trap in the FRAP: A Cautionary Tale about Transport Measurements in Biomolecular Condensates. *Biophysical journal* vol. 117 2041–2042 (2019).
361. Taylor, N. O., Wei, M.-T., Stone, H. A. & Brangwynne, C. P. Quantifying Dynamics in Phase-Separated Condensates Using Fluorescence Recovery after Photobleaching. *Biophys. J.* **117**, 1285–1300 (2019).
362. Flint, S. J., Racaniello, V. R., Rall, G. F., Hatziioannou, T. & Skalka, A. M. *Principles of Virology: Molecular Biology*. (Wiley, 2020).
363. Varsani, A., Lefeuvre, P., Roumagnac, P. & Martin, D. Notes on recombination and reassortment in multipartite/segmented viruses. *Curr. Opin. Virol.* **33**, 156–166 (2018).
364. Nevers, Q., Albertini, A. A., Lagaudrière-Gesbert, C. & Gaudin, Y. Negri bodies and other virus membrane-less replication compartments. *Biochim. Biophys. Acta Mol. Cell Res.* **1867**, 118831 (2020).
365. Negri, A. *Contributo allo studio dell'eziologia della rabbia*. (Tipografia e Legatoria Cooperativa, 1905).
366. Lahaye, X. *et al.* Functional characterization of Negri bodies (NBs) in rabies virus-infected cells: Evidence that NBs are sites of viral transcription and replication. *J. Virol.* **83**, 7948–7958 (2009).
367. Lahaye, X., Vidy, A., Fouquet, B. & Blondel, D. Hsp70 protein positively regulates rabies virus infection. *J. Virol.* **86**, 4743–4751 (2012).

368. Sagan, S. M. & Weber, S. C. Let's phase it: viruses are master architects of biomolecular condensates. *Trends Biochem. Sci.* (2022) doi:10.1016/j.tibs.2022.09.008.
369. Zhou, Y., Su, J. M., Samuel, C. E. & Ma, D. Measles Virus Forms Inclusion Bodies with Properties of Liquid Organelles. *J. Virol.* **93**, (2019).
370. Ma, D. *et al.* Upon Infection, Cellular WD Repeat-Containing Protein 5 (WDR5) Localizes to Cytoplasmic Inclusion Bodies and Enhances Measles Virus Replication. *J. Virol.* **92**, (2018).
371. Heinrich, B. S., Maliga, Z., Stein, D. A., Hyman, A. A. & Whelan, S. P. J. Phase Transitions Drive the Formation of Vesicular Stomatitis Virus Replication Compartments. *MBio* **9**, (2018).
372. Geiger, F. *et al.* Liquid-liquid phase separation underpins the formation of replication factories in rotaviruses. *EMBO J.* **40**, e107711 (2021).
373. Caragliano, E. *et al.* Human cytomegalovirus forms phase-separated compartments at viral genomes to facilitate viral replication. *Cell Rep.* **38**, 110469 (2022).
374. Charlier, C. M. *et al.* Analysis of borna disease virus trafficking in live infected cells by using a virus encoding a tetracysteine-tagged p protein. *J. Virol.* **87**, 12339–12348 (2013).
375. Nikolic, J. *et al.* Negri bodies are viral factories with properties of liquid organelles. *Nat. Commun.* **8**, 58 (2017).
376. Ménager, P. *et al.* Toll-like receptor 3 (TLR3) plays a major role in the formation of rabies virus Negri Bodies. *PLoS Pathog.* **5**, e1000315 (2009).
377. Wolozin, B. & Ivanov, P. Stress granules and neurodegeneration. *Nat. Rev. Neurosci.* **20**, 649–666 (2019).
378. Marcelo, A., Koppenol, R., de Almeida, L. P., Matos, C. A. & Nóbrega, C. Stress granules, RNA-binding proteins and polyglutamine diseases: too much aggregation? *Cell Death Dis.* **12**, 592 (2021).
379. Qin, Q., Hastings, C. & Miller, C. L. Mammalian orthoreovirus particles induce and are recruited into stress granules at early times postinfection. *J. Virol.* **83**, 11090–11101 (2009).
380. Qin, Q., Carroll, K., Hastings, C. & Miller, C. L. Mammalian orthoreovirus escape from host translational shutoff correlates with stress granule disruption and is independent of eIF2alpha phosphorylation and PKR. *J. Virol.* **85**, 8798–8810 (2011).
381. McInerney, G. M., Kedersha, N. L., Kaufman, R. J., Anderson, P. & Liljeström, P. Importance of eIF2α Phosphorylation and Stress Granule Assembly in Alphavirus Translation Regulation. *MBoC* **16**, 3753–3763 (2005).
382. Mazroui, R. *et al.* Inhibition of Ribosome Recruitment Induces Stress Granule Formation Independently of Eukaryotic Initiation Factor 2α Phosphorylation. *MBoC* **17**, 4212–4219 (2006).
383. White, J. P., Cardenas, A. M., Marissen, W. E. & Lloyd, R. E. Inhibition of cytoplasmic mRNA stress granule formation by a viral proteinase. *Cell Host Microbe* **2**, 295–305 (2007).
384. White, J. P. & Lloyd, R. E. Poliovirus unlinks TIA1 aggregation and mRNA stress granule formation. *J. Virol.* **85**, 12442–12454 (2011).
385. Ariumi, Y. *et al.* Hepatitis C virus hijacks P-body and stress granule components around lipid droplets. *J. Virol.* **85**, 6882–6892 (2011).
386. Eisfeld, A. J., Neumann, G. & Kawaoka, Y. At the centre: influenza A virus ribonucleoproteins.

Nat. Rev. Microbiol. **13**, 28–41 (2015).
387. Lakdawala, S. S. *et al.* Influenza a virus assembly intermediates fuse in the cytoplasm. *PLoS Pathog.* **10**, e1003971 (2014).
388. Alenquer, M. *et al.* Influenza A virus ribonucleoproteins form liquid organelles at endoplasmic reticulum exit sites. *Nat. Commun.* **10**, 1629 (2019).
389. Chou, Y.-Y. *et al.* Colocalization of different influenza viral RNA segments in the cytoplasm before viral budding as shown by single-molecule sensitivity FISH analysis. *PLoS Pathog.* **9**, e1003358 (2013).
390. Guseva, S. *et al.* Measles virus nucleo- and phosphoproteins form liquid-like phase-separated compartments that promote nucleocapsid assembly. *Sci Adv* **6**, eaaz7095 (2020).
391. Lu, S. *et al.* The SARS-CoV-2 Nucleocapsid phosphoprotein forms mutually exclusive condensates with RNA and the membrane-associated M protein. 2020.07.30.228023 (2020) doi:10.1101/2020.07.30.228023.
392. Savastano, A., Ibáñez de Opakua, A., Rankovic, M. & Zweckstetter, M. Nucleocapsid protein of SARS-CoV-2 phase separates into RNA-rich polymerase-containing condensates. *Nat. Commun.* **11**, 6041 (2020).
393. Soranno, A. *et al.* Shelterin Components Modulate Nucleic Acids Condensation and Phase Separation in the Context of Telomeric DNA. *J. Mol. Biol.* **434**, 167685 (2022).
394. Etibor, T. A., Yamauchi, Y. & Amorim, M. J. Liquid Biomolecular Condensates and Viral Lifecycles: Review and Perspectives. *Viruses* **13**, (2021).
395. Scoca, V. & Di Nunzio, F. Membraneless organelles restructured and built by pandemic viruses: HIV-1 and SARS-CoV-2. *J. Mol. Cell Biol.* **13**, 259–268 (2021).
396. Brocca, S., Grandori, R., Longhi, S. & Uversky, V. Liquid-Liquid Phase Separation by Intrinsically Disordered Protein Regions of Viruses: Roles in Viral Life Cycle and Control of Virus-Host Interactions. *Int. J. Mol. Sci.* **21**, (2020).
397. Dolnik, O., Gerresheim, G. K. & Biedenkopf, N. New Perspectives on the Biogenesis of Viral Inclusion Bodies in Negative-Sense RNA Virus Infections. *Cells* **10**, (2021).
398. Rubinstein, M. & Colby, R. H. *Polymer Physics*. (Oxford University Press, 2003).
399. Murphy, M. C., Rasnik, I., Cheng, W., Lohman, T. M. & Ha, T. Probing single-stranded DNA conformational flexibility using fluorescence spectroscopy. *Biophys. J.* **86**, 2530–2537 (2004).
400. Kohn, J. E. *et al.* Random-coil behavior and the dimensions of chemically unfolded proteins. *Proceedings of the National Academy of Sciences* **101**, 12491–12496 (2004).
401. Soranno, A. Physical basis of the disorder-order transition. *Arch. Biochem. Biophys.* **685**, 108305 (2020).
402. Quiroz, F. G. & Chilkoti, A. Sequence heuristics to encode phase behaviour in intrinsically disordered protein polymers. *Nat. Mater.* **14**, 1164–1171 (2015).
403. Varanko, A. K., Su, J. C. & Chilkoti, A. Elastin-Like Polypeptides for Biomedical Applications. *Annu. Rev. Biomed. Eng.* **22**, 343–369 (2020).
404. Zeng, X., Holehouse, A. S., Chilkoti, A., Mittag, T. & Pappu, R. V. Connecting Coil-to-Globule Transitions to Full Phase Diagrams for Intrinsically Disordered Proteins. *Biophys. J.* **119**, 402–418

(2020).
405. Joseph, J. A. *et al.* Physics-driven coarse-grained model for biomolecular phase separation with near-quantitative accuracy. *Nat Comput Sci* **1**, 732–743 (2021).
406. Joseph, J. A. *et al.* Thermodynamics and kinetics of phase separation of protein-RNA mixtures by a minimal model. *Biophys. J.* **120**, 1219–1230 (2021).
407. Holehouse, A. S. & Pappu, R. V. *PIMMS (0.24 pre-beta).* (2019). doi:10.5281/zenodo.3588456.
408. Choi, J.-M., Dar, F. & Pappu, R. V. LASSI: A lattice model for simulating phase transitions of multivalent proteins. *PLoS Comput. Biol.* **15**, e1007028 (2019).
409. Semenov, A. N. & Rubinstein, M. Thermoreversible Gelation in Solutions of Associative Polymers. 1. Statics. *Macromolecules* **31**, 1373–1385 (1998).
410. Ginell, G. M. & Holehouse, A. S. An Introduction to the Stickers-and-Spacers Framework as Applied to Biomolecular Condensates. in *Phase-Separated Biomolecular Condensates: Methods and Protocols* (eds. Zhou, H.-X., Spille, J.-H. & Banerjee, P. R.) 95–116 (Springer US, 2023).
411. Nott, T. J. *et al.* Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. *Mol. Cell* **57**, 936–947 (2015).
412. Molliex, A. *et al.* Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. *Cell* **163**, 123–133 (2015).
413. Wang, J. *et al.* A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins. *Cell* vol. 174 688–699.e16 Preprint at https://doi.org/10.1016/j.cell.2018.06.006 (2018).
414. Bremer, A. *et al.* Deciphering how naturally occurring sequence features impact the phase behaviors of disordered prion-like domains. bioRxiv [Preprint](2021). Preprint at https://doi.org/10.1101/2021.01.01.425046.
415. Han, T. W. *et al.* Cell-free Formation of RNA Granules: Bound RNAs Identify Features and Components of Cellular Assemblies. *Cell* vol. 149 768–779 Preprint at https://doi.org/10.1016/j.cell.2012.04.016 (2012).
416. Kato, M. *et al.* Cell-free Formation of RNA Granules: Low Complexity Sequence Domains Form Dynamic Fibers within Hydrogels. *Cell* vol. 149 753–767 Preprint at https://doi.org/10.1016/j.cell.2012.04.017 (2012).
417. Frey, S., Richter, R. P. & Görlich, D. FG-rich repeats of nuclear pore proteins form a three-dimensional meshwork with hydrogel-like properties. *Science* **314**, 815–817 (2006).
418. Yoshizawa, T. *et al.* Nuclear Import Receptor Inhibits Phase Separation of FUS through Binding to Multiple Sites. *Cell* **173**, 693–705.e22 (2018).
419. Qamar, S. *et al.* FUS Phase Separation Is Modulated by a Molecular Chaperone and Methylation of Arginine Cation-π Interactions. *Cell* **173**, 720–734.e15 (2018).
420. Li, H.-R., Chiang, W.-C., Chou, P.-C., Wang, W.-J. & Huang, J.-R. TAR DNA-binding protein 43 (TDP-43) liquid–liquid phase separation is mediated by just a few aromatic residues. *J. Biol. Chem.* **293**, 6090–6098 (2018).
421. Banerjee, P. R., Milin, A. N., Moosa, M. M., Onuchic, P. L. & Deniz, A. A. Reentrant Phase Transition Drives Dynamic Substructure Formation in Ribonucleoprotein Droplets. *Angew.*

Chem. Int. Ed Engl. **56**, 11354–11359 (2017).
422. Chang, L.-W. *et al.* Sequence and entropy-based control of complex coacervates. *Nat. Commun.* **8**, 1273 (2017).
423. Pak, C. W. *et al.* Sequence determinants of intracellular phase separation by complex coacervation of a disordered protein. *Mol. Cell* **63**, 72–85 (2016).
424. Lin, Y.-H. & Chan, H. S. Phase Separation and Single-Chain Compactness of Charged Disordered Proteins Are Strongly Correlated. *Biophys. J.* **112**, 2043–2046 (2017).
425. Cummings, C. S. & Obermeyer, A. C. Phase Separation Behavior of Supercharged Proteins and Polyelectrolytes. *Biochemistry* **57**, 314–323 (2018).
426. Mitrea, D. M. *et al.* Nucleophosmin integrates within the nucleolus via multi-modal interactions with proteins displaying R-rich linear motifs and rRNA. *Elife* **5**, (2016).
427. Galvanetto, N. *et al.* Ultrafast molecular dynamics observed within a dense protein condensate. *bioRxiv* 2022.12.12.520135 (2022) doi:10.1101/2022.12.12.520135.
428. Schmidt, H. B., Barreau, A. & Rohatgi, R. Phase separation-deficient TDP43 remains functional in splicing. *Nat. Commun.* **10**, 4890 (2019).
429. Holehouse, A. S., Ginell, G. M., Griffith, D. & Böke, E. Clustering of aromatic residues in prion-like domains can tune the formation, state, and organization of biomolecular condensates. *Biochemistry* **60**, 3566–3581 (2021).
430. Weiner, B. G., Pyo, A. G. T., Meir, Y. & Wingreen, N. S. Motif-pattern dependence of biomolecular phase separation driven by specific interactions. *PLoS Comput. Biol.* **17**, e1009748 (2021).
431. Harmon, T. S., Holehouse, A. S. & Pappu, R. V. Differential solvation of intrinsically disordered linkers drives the formation of spatially organized droplets in ternary systems of linear multivalent proteins. *New J. Phys.* **20**, 045002 (2018).
432. Ranganathan, S. & Shakhnovich, E. I. Dynamic metastable long-living droplets formed by sticker-spacer proteins. *Elife* **9**, (2020).
433. Choi, J.-M., Hyman, A. A. & Pappu, R. V. Generalized models for bond percolation transitions of associative polymers. *Phys Rev E* **102**, 042403 (2020).
434. Cohan, M. C. & Pappu, R. V. Making the Case for Disordered Proteins and Biomolecular Condensates in Bacteria. *Trends Biochem. Sci.* **45**, 668–680 (2020).
435. Harmon, T. S., Holehouse, A. S., Rosen, M. K. & Pappu, R. V. Intrinsically disordered linkers determine the interplay between phase separation and gelation in multivalent proteins. *Elife* **6**, (2017).
436. Bhandari, K., Cotten, M. A., Kim, J., Rosen, M. K. & Schmit, J. D. Structure–Function Properties in Disordered Condensates. *J. Phys. Chem. B* **125**, 467–476 (2021).
437. Schmit, J. D., Bouchard, J. J., Martin, E. W. & Mittag, T. Protein Network Structure Enables Switching between Liquid and Gel States. *J. Am. Chem. Soc.* **142**, 874–883 (2020).
438. Li, P. *et al.* Phase transitions in the assembly of multivalent signalling proteins. *Nature* **483**, 336–340 (2012).
439. Teif, V. B. & Bohinc, K. Condensed DNA: condensing the concepts. *Prog. Biophys. Mol. Biol.* **105**,

208–222 (2011).
440. Post, C. B. & Zimm, B. H. Theory of DNA condensation: collapse versus aggregation. *Biopolymers* **21**, 2123–2137 (1982).
441. Cubuk, J. *et al.* The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA. *bioRxiv* 2020.06.17.158121 (2020) doi:10.1101/2020.06.17.158121.
442. Iserman, C. *et al.* Genomic RNA Elements Drive Phase Separation of the SARS-CoV-2 Nucleocapsid. *Mol. Cell* **80**, 1078–1091.e6 (2020).
443. Lu, S. *et al.* The SARS-CoV-2 nucleocapsid phosphoprotein forms mutually exclusive condensates with RNA and the membrane-associated M protein. *Nat. Commun.* **12**, 502 (2021).
444. Adam, M. A. & Miller, A. D. Identification of a signal in a murine retrovirus that is sufficient for packaging of nonretroviral RNA into virions. *J. Virol.* **62**, 3802–3806 (1988).
445. Dornburg, R. & Temin, H. M. Presence of a retroviral encapsidation sequence in nonretroviral RNA increases the efficiency of formation of cDNA genes. *J. Virol.* **64**, 886–889 (1990).
446. Mansky, L. M., Krueger, A. E. & Temin, H. M. The bovine leukemia virus encapsidation signal is discontinuous and extends into the 5' end of the gag gene. *J. Virol.* **69**, 3282–3289 (1995).
447. Vile, R. G., Ali, M., Hunter, E. & McClure, M. O. Identification of a generalised packaging sequence for D-type retroviruses and generation of a D-type retroviral vector. *Virology* **189**, 786–791 (1992).
448. Ding, P. *et al.* Identification of the initial nucleocapsid recognition element in the HIV-1 RNA packaging signal. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 17737–17746 (2020).
449. Bessa, L. M. *et al.* The intrinsically disordered SARS-CoV-2 nucleoprotein in dynamic complex with its viral partner nsp3a. *Sci Adv* **8**, eabm4034 (2022).
450. Klein, S. *et al.* SARS-CoV-2 structure and replication characterized by in situ cryo-electron tomography. doi:10.1101/2020.06.23.167064.
451. Hsu, C. C. & Prausnitz, J. M. Thermodynamics of Polymer Compatibility in Ternary Systems. *Macromolecules* **7**, 320–324 (1974).
452. Koningsveld, R. & Staverman, A. J. Liquid-liquid phase separation in multicomponent polymer solutions. *Colloid Polym. Sci.* **220**, 31–40 (1967).
453. Deviri, D. & Safran, S. A. Physical theory of biological noise buffering by multicomponent phase separation. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
454. Brown, C. J., Johnson, A. K., Dunker, A. K. & Daughdrill, G. W. Evolution and disorder. *Curr. Opin. Struct. Biol.* **21**, 441–446 (2011).
455. Holehouse, A. S., Das, R. K., Ahad, J. N., Richardson, M. O. G. & Pappu, R. V. CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. *Biophys. J.* **112**, 16–21 (2017).
456. Zhang, Y., Xu, B., Weiner, B. G., Meir, Y. & Wingreen, N. S. Decoding the physical principles of two-component biomolecular phase separation. *Elife* **10**, (2021).
457. Prusty, D., Pryamitsyn, V. & Olvera de la Cruz, M. Thermodynamics of Associative Polymer Blends. *Macromolecules* **51**, 5918–5932 (2018).
458. Nandi, S. K., Österle, D., Heidenreich, M., Levy, E. D. & Safran, S. A. Affinity and Valence Impact

the Extent and Symmetry of Phase Separation of Multivalent Proteins. *Phys. Rev. Lett.* **129**, 128102 (2022).
459. Risso-Ballester, J. *et al.* A condensate-hardening drug blocks RSV replication in vivo. *Nature* **595**, 596–599 (2021).
460. Van Lindt, J. *et al.* A generic approach to study the kinetics of liquid-liquid phase separation under near-native conditions. *Commun Biol* **4**, 77 (2021).
461. Ceballos, A. V., McDonald, C. J. & Elbaum-Garfinkle, S. Methods and Strategies to Quantify Phase Separation of Disordered Proteins. *Methods Enzymol.* **611**, 31–50 (2018).
462. Li, P. *et al.* Rapid Determination of Phase Diagrams for Biomolecular Liquid-Liquid Phase Separation with Microfluidics. *Anal. Chem.* **94**, 687–694 (2022).
463. Milkovic, N. M. & Mittag, T. Determination of Protein Phase Diagrams by Centrifugation. in *Intrinsically Disordered Proteins: Methods and Protocols* (eds. Kragelund, B. B. & Skriver, K.) 685–702 (Springer US, 2020).
464. Carlson, C. R. *et al.* Phosphoregulation of Phase Separation by the SARS-CoV-2 N Protein Suggests a Biophysical Basis for its Dual Functions. *Mol. Cell* **80**, 1092–1103.e4 (2020).
465. Ganser, L. R. & Myong, S. Methods to Study Phase-Separated Condensates and the Underlying Molecular Interactions. *Trends Biochem. Sci.* **45**, 1004–1005 (2020).
466. Henninger, J. E. *et al.* RNA-mediated feedback control of transcriptional condensates. *Cell* **184**, 207–225.e24 (2021).
467. Kaur, T. *et al.* Molecular Crowding Tunes Material States of Ribonucleoprotein Condensates. *Biomolecules* **9**, (2019).
468. Berry, J., Weber, S. C., Vaidya, N., Haataja, M. & Brangwynne, C. P. RNA transcription modulates phase transition-driven nuclear body assembly. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E5237–45 (2015).
469. Wang, J., Shi, C., Xu, Q. & Yin, H. SARS-CoV-2 nucleocapsid protein undergoes liquid-liquid phase separation into stress granules through its N-terminal intrinsically disordered region. *Cell Discov* **7**, 5 (2021).
470. Wang, Z., Zhang, G. & Zhang, H. Protocol for analyzing protein liquid–liquid phase separation. *Biophysics Reports* **5**, 1–9 (2019).
471. Alberti, S., Gladfelter, A. & Mittag, T. Considerations and Challenges in Studying Liquid-Liquid Phase Separation and Biomolecular Condensates. *Cell* **176**, 419–434 (2019).
472. Peng, S. *et al.* Phase separation at the nanoscale quantified by dcFCCS. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 27124–27131 (2020).
473. Guillén-Boixet, J. *et al.* RNA-Induced Conformational Switching and Clustering of G3BP Drive Stress Granule Assembly by Condensation. *Cell* **181**, 346–361.e17 (2020).
474. Magde, D., Elson, E. & Webb, W. W. Thermodynamic Fluctuations in a Reacting System---Measurement by Fluorescence Correlation Spectroscopy. *Phys. Rev. Lett.* **29**, 705–708 (1972).
475. Kral, T., Hof, M., Jurkiewicz, P. & Langner, M. Fluorescence correlation spectroscopy (FCS) as a tool to study DNA condensation with hexadecyltrimethylammonium bromide (HTAB). *Cell. Mol. Biol. Lett.* **7**, 203–211 (2002).

476. Kral, T., Hof, M. & Langner, M. The effect of spermine on plasmid condensation and dye release observed by fluorescence correlation spectroscopy. *Biol. Chem.* **383**, 331–335 (2002).
477. Ramisetty, S. K., Langlete, P., Lale, R. & Dias, R. S. In vitro studies of DNA condensation by bridging protein in a crowding environment. *Int. J. Biol. Macromol.* **103**, 845–853 (2017).
478. Hodges, C. & Meiners, J.-C. Fluorescence Correlation Spectroscopy on Genomic DNA in Living Cells. in *Nanoscale Imaging: Methods and Protocols* (ed. Lyubchenko, Y. L.) 415–424 (Springer New York, 2018).
479. Sabanayagam, C. R., Oram, M., Lakowicz, J. R. & Black, L. W. Viral DNA packaging studied by fluorescence correlation spectroscopy. *Biophys. J.* **93**, L17–9 (2007).
480. Gopal, A. *et al.* Viral RNAs are unusually compact. *PLoS One* **9**, e105875 (2014).
481. Borodavka, A. *et al.* Sizes of Long RNA Molecules Are Determined by the Branching Patterns of Their Secondary Structures. *Biophys. J.* **111**, 2077–2085 (2016).
482. Stryer, L. & Haugland, R. P. Energy transfer: a spectroscopic ruler. *Proc. Natl. Acad. Sci. U. S. A.* **58**, 719–726 (1967).
483. Pelicci, S., Diaspro, A. & Lanzanò, L. Chromatin nanoscale compaction in live cells visualized by acceptor-to-donor ratio corrected Förster resonance energy transfer between DNA dyes. *J. Biophotonics* **12**, e201900164 (2019).
484. Levchenko, S. M., Pliss, A., Peng, X., Prasad, P. N. & Qu, J. Fluorescence lifetime imaging for studying DNA compaction and gene activities. *Light Sci Appl* **10**, 224 (2021).
485. König, I. *et al.* Single-molecule spectroscopy of protein conformational dynamics in live eukaryotic cells. *Nat. Methods* **12**, 773–779 (2015).
486. Schuler, B., König, I., Soranno, A. & Nettels, D. Impact of in-cell and in-vitro crowding on the conformations and dynamics of an intrinsically disordered protein. *Angew. Chem. Weinheim Bergstr. Ger.* (2021) doi:10.1002/ange.202016804.
487. Wen, J. *et al.* Conformational Expansion of Tau in Condensates Promotes Irreversible Aggregation. *J. Am. Chem. Soc.* **143**, 13056–13064 (2021).
488. Ray, S., Singh, N., Patel, K., Krishnamoorthy, G. & Maji, S. K. FRAP and FRET Investigation of α-Synuclein Fibrillization via Liquid-Liquid Phase Separation In Vitro and in HeLa Cells. *Methods Mol. Biol.* **2551**, 395–423 (2023).
489. Wang, H., Musier-Forsyth, K., Falk, C. & Barbara, P. F. Single-molecule spectroscopic study of dynamic nanoscale DNA bending behavior of HIV-1 nucleocapsid protein. *J. Phys. Chem. B* **117**, 4183–4196 (2013).
490. Bustamante, C. J., Chemla, Y. R., Liu, S. & Wang, M. D. Optical tweezers in single-molecule biophysics. *Nat Rev Methods Primers* **1**, (2021).
491. Fabian, R., Jr *et al.* A Horizontal Magnetic Tweezers for Studying Single DNA Molecules and DNA-Binding Proteins. *Molecules* **26**, (2021).
492. Butt, H.-J., Cappella, B. & Kappl, M. Force measurements with the atomic force microscope: Technique, interpretation and applications. *Surf. Sci. Rep.* **59**, 1–152 (2005).
493. Wang, Y. *et al.* Direct Demonstration of DNA Compaction Mediated by Divalent Counterions. *J. Phys. Chem. B* **123**, 79–85 (2019).

494. Gao, T., Zhang, W., Wang, Y. & Yang, G. DNA Compaction and Charge Neutralization Regulated by Divalent Ions in very Low pH Solution. *Polymers* **11**, (2019).
495. van den Broek, B. *et al.* Visualizing the formation and collapse of DNA toroids. *Biophys. J.* **98**, 1902–1910 (2010).
496. Baumann, C. G. *et al.* Stretching of single collapsed DNA molecules. *Biophys. J.* **78**, 1965–1978 (2000).
497. Baumann, C. G., Smith, S. B., Bloomfield, V. A. & Bustamante, C. Ionic effects on the elasticity of single DNA molecules. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 6185–6190 (1997).
498. Nguyen, T. *et al.* Chromatin sequesters pioneer transcription factor Sox2 from exerting force on DNA. *Nat. Commun.* **13**, 3988 (2022).
499. Renger, R. *et al.* Co-condensation of proteins with single- and double-stranded DNA. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2107871119 (2022).
500. Gien, H. *et al.* HIV-1 Nucleocapsid Protein Binds Double-Stranded DNA in Multiple Modes to Regulate Compaction and Capsid Uncoating. *Viruses* **14**, (2022).
501. Xi, B. & Ran, S.-Y. Formation of DNA pearl-necklace structures on mica surface governed by kinetics and thermodynamics. *J. Polym. Sci. B Polym. Phys.* **55**, 971–979 (2017).
502. Quail, T. *et al.* Force generation by protein–DNA co-condensation. *Nat. Phys.* **17**, 1007–1012 (2021).
503. Alshareedah, I., Kaur, T. & Banerjee, P. R. Methods for characterizing the material properties of biomolecular condensates. *Methods Enzymol.* **646**, 143–183 (2021).
504. Gui, X. *et al.* Structural basis for reversible amyloids of hnRNPA1 elucidates their role in stress granule assembly. *Nat. Commun.* **10**, 2006 (2019).
505. Jawerth, L. M. *et al.* Salt-Dependent Rheology and Surface Tension of Protein Condensates Using Optical Traps. *Phys. Rev. Lett.* **121**, 258101 (2018).
506. Jawerth, L. *et al.* Protein condensates as aging Maxwell fluids. *Science* **370**, 1317–1323 (2020).
507. Greene, E. C., Wind, S., Fazio, T., Gorman, J. & Visnapuu, M.-L. Chapter 14 - DNA Curtains for High-Throughput Single-Molecule Optical Imaging. in *Methods in Enzymology* (ed. Walter, N. G.) vol. 472 293–315 (Academic Press, 2010).
508. Schaub, J. M., Zhang, H., Soniat, M. M. & Finkelstein, I. J. Assessing Protein Dynamics on Low-Complexity Single-Stranded DNA Curtains. *Langmuir* **34**, 14882–14890 (2018).
509. Zuo, L. *et al.* Loci-specific phase separation of FET fusion oncoproteins promotes gene transcription. *Nat. Commun.* **12**, 1491 (2021).
510. Keenen, M. M. *et al.* HP1 proteins compact DNA into mechanically and positionally stable phase separated domains. *Elife* **10**, (2021).
511. Calcines-Cruz, C., Finkelstein, I. J. & Hernandez-Garcia, A. CRISPR-Guided Programmable Self-Assembly of Artificial Virus-Like Nucleocapsids. *Nano Lett.* **21**, 2752–2757 (2021).
512. Murthy, A. C. & Fawzi, N. L. The (un)structural biology of biomolecular liquid-liquid phase separation using NMR spectroscopy. *Journal of Biological Chemistry* vol. 295 2375–2384 Preprint at https://doi.org/10.1074/jbc.rev119.009847 (2020).
513. Summers, M. F., South, T. L., Kim, B. & Hare, D. R. High-resolution structure of an HIV zinc fingerlike domain via a new NMR-based distance geometry approach. *Biochemistry* **29**, 329–340

(1990).
514. South, T. L., Blake, P. R., Hare, D. R. & Summers, M. F. C-terminal retroviral-type zinc finger domain from the HIV-1 nucleocapsid protein is structurally similar to the N-terminal zinc finger domain. *Biochemistry* **30**, 6342–6349 (1991).
515. Cheong, H. K., Cheong, C. & Choi, B. S. Secondary structure of the panhandle RNA of influenza virus A studied by NMR spectroscopy. *Nucleic Acids Res.* **24**, 4197–4201 (1996).
516. Lee, M.-K. *et al.* A single-nucleotide natural variation (U4 to C4) in an influenza A virus promoter exhibits a large structural change: implications for differential viral RNA synthesis by RNA-dependent RNA polymerase. *Nucleic Acids Res.* **31**, 1216–1223 (2003).
517. Lu, K. *et al.* NMR detection of structures in the HIV-1 5'-leader RNA that regulate genome packaging. *Science* **334**, 242–245 (2011).
518. Brown, J. D. *et al.* Structural basis for transcriptional start site control of HIV-1 RNA fate. *Science* **368**, 413–417 (2020).
519. Wacker, A. *et al.* Secondary structure determination of conserved SARS-CoV-2 RNA elements by NMR spectroscopy. *Nucleic Acids Res.* **48**, 12415–12435 (2020).
520. Sun, Y.-T. & Varani, G. Structure of the dengue virus RNA promoter. *RNA* **28**, 1210–1223 (2022).
521. Tang, C., Ndassa, Y. & Summers, M. F. Structure of the N-terminal 283-residue fragment of the immature HIV-1 Gag polyprotein. *Nat. Struct. Biol.* **9**, 537–543 (2002).
522. Shen, Q. & Cho, J.-H. The structure and conformational plasticity of the nonstructural protein 1 of the 1918 influenza A virus. *Biochem. Biophys. Res. Commun.* **518**, 178–182 (2019).
523. Whitehead, R. D., 3rd, Teschke, C. M. & Alexandrescu, A. T. NMR Mapping of Disordered Segments from a Viral Scaffolding Protein Enclosed in a 23 MDa Procapsid. *Biophys. J.* **117**, 1387–1392 (2019).
524. Schiavina, M., Pontoriero, L., Uversky, V. N., Felli, I. C. & Pierattelli, R. The highly flexible disordered regions of the SARS-CoV-2 nucleocapsid N protein within the 1–248 residue construct: sequence-specific resonance assignments through NMR. *Biomolecular NMR Assignments* vol. 15 219–227 Preprint at https://doi.org/10.1007/s12104-021-10009-8 (2021).
525. Dinesh, D. C. *et al.* Structural basis of RNA recognition by the SARS-CoV-2 nucleocapsid phosphoprotein. *PLoS Pathog.* **16**, e1009100 (2020).
526. Redzic, J. S. *et al.* The Inherent Dynamics and Interaction Sites of the SARS-CoV-2 Nucleocapsid N-Terminal Region. *J. Mol. Biol.* **433**, 167108 (2021).
527. Pontoriero, L. *et al.* NMR Reveals Specific Tracts within the Intrinsically Disordered Regions of the SARS-CoV-2 Nucleocapsid Protein Involved in RNA Encountering. *Biomolecules* **12**, (2022).
528. Burke, K. A., Janke, A. M., Rhine, C. L. & Fawzi, N. L. Residue-by-Residue View of In Vitro FUS Granules that Bind the C-Terminal Domain of RNA Polymerase II. *Mol. Cell* **60**, 231–241 (2015).
529. Emmanouilidis, L. *et al.* NMR and EPR reveal a compaction of the RNA-binding protein FUS upon droplet formation. *Nat. Chem. Biol.* **17**, 608–614 (2021).
530. Emmanouilidis, L., Esteban-Hofer, L., Jeschke, G. & Allain, F. H.-T. Structural biology of RNA-binding proteins in the context of phase separation: What NMR and EPR can bring? *Curr. Opin.*

Struct. Biol. **70**, 132–138 (2021).
531. Pollack, L. SAXS Studies of Ion–Nucleic Acid Interactions. *Annu. Rev. Biophys.* **40**, 225–242 (2011).
532. Kikhney, A. G. & Svergun, D. I. A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Lett.* **589**, 2570–2577 (2015).
533. Lake, J. A. Yeast transfer RNA: a small-angle x-ray study. *Science* **156**, 1371–1373 (1967).
534. Russell, R. *et al.* Rapid compaction during RNA folding. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 4266–4271 (2002).
535. Caliskan, G. *et al.* Persistence length changes dramatically as RNA folds. *Phys. Rev. Lett.* **95**, 268303 (2005).
536. Welty, R. *et al.* Divalent ions tune the kinetics of a bacterial GTPase center rRNA folding transition from secondary to tertiary structure. *RNA* **24**, 1828–1838 (2018).
537. Koltover, I., Wagner, K. & Safinya, C. R. DNA condensation in two dimensions. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 14046–14051 (2000).
538. Qiu, X., Andresen, K., Lamb, J. S., Kwok, L. W. & Pollack, L. Abrupt transition from a free, repulsive to a condensed, attractive DNA phase, induced by multivalent polyamine cations. *Phys. Rev. Lett.* **101**, 228101 (2008).
539. Baker, M. A. B. *et al.* Dimensions and Global Twist of Single-Layer DNA Origami Measured by Small-Angle X-ray Scattering. *ACS Nano* **12**, 5791–5799 (2018).
540. Ober, M. F., Baptist, A., Wassermann, L., Heuer-Jungemann, A. & Nickel, B. In situ small-angle X-ray scattering reveals strong condensation of DNA origami during silicification. *Nat. Commun.* **13**, 5668 (2022).
541. Souza, B. B. S. *et al.* A biophysical study of DNA condensation mediated by histones and protamines. *J. Mol. Liq.* **368**, 120745 (2022).
542. Li, L., Pabit, S. A., Meisburger, S. P. & Pollack, L. Double-stranded RNA resists condensation. *Phys. Rev. Lett.* **106**, 108101 (2011).
543. Martin, E. W. *et al.* A multi-step nucleation process determines the kinetics of prion-like domain phase separation. *Nat. Commun.* **12**, 4513 (2021).
544. Alston, J. J., Soranno, A. & Holehouse, A. S. Integrating single-molecule spectroscopy and simulations for the study of intrinsically disordered proteins. *Methods* **193**, 116–135 (2021).
545. Young, M. A., Jayaram, B. & Beveridge, D. L. Intrusion of Counterions into the Spine of Hydration in the Minor Groove of B-DNA: Fractional Occupancy of Electronegative Pockets. *J. Am. Chem. Soc.* **119**, 59–69 (1997).
546. Giambaşu, G. M., Luchko, T., Herschlag, D., York, D. M. & Case, D. A. Ion Counting from Explicit-Solvent Simulations and 3D-RISM. *Biophysical Journal* vol. 106 883–894 Preprint at https://doi.org/10.1016/j.bpj.2014.01.021 (2014).
547. Savelyev, A. & MacKerell, A. D., Jr. Competition among Li(+), Na(+), K(+), and Rb(+) monovalent ions for DNA in molecular dynamics simulations using the additive CHARMM36 and Drude polarizable force fields. *J. Phys. Chem. B* **119**, 4428–4440 (2015).
548. Yoo, J. & Aksimentiev, A. Competitive binding of cations to duplex DNA revealed through molecular dynamics simulations. *J. Phys. Chem. B* **116**, 12946–12954 (2012).

549. Pasi, M., Maddocks, J. H. & Lavery, R. Analyzing ion distributions around DNA: sequence-dependence of potassium ion distributions from microsecond molecular dynamics. *Nucleic Acids Res.* **43**, 2412–2423 (2015).
550. Xi, K., Wang, F.-H., Xiong, G., Zhang, Z.-L. & Tan, Z.-J. Competitive Binding of Mg2+ and Na+ Ions to Nucleic Acids: From Helices to Tertiary Structures. *Biophys. J.* **114**, 1776–1790 (2018).
551. Kar, M. *et al.* Phase-separating RNA-binding proteins form heterogeneous distributions of clusters in subsaturated solutions. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2202222119 (2022).
552. Cubuk, J. *et al.* The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA. *Nat. Commun.* **12**, 1936 (2021).
553. Ashbaugh, H. S. & Hatch, H. W. Natively unfolded protein stability as a coil-to-globule transition in charge/hydropathy space. *J. Am. Chem. Soc.* **130**, 9536–9542 (2008).
554. Ashbaugh, H. S. Tuning the globular assembly of hydrophobic/hydrophilic heteropolymer sequences. *J. Phys. Chem. B* **113**, 14043–14046 (2009).
555. Dignon, G. L., Zheng, W. & Mittal, J. Simulation methods for liquid–liquid phase separation of disordered proteins. *Curr. Opin. Chem. Eng.* **23**, 92–98 (2019).
556. Regy, R. M., Dignon, G. L., Zheng, W., Kim, Y. C. & Mittal, J. Sequence dependent phase separation of protein-polynucleotide mixtures elucidated using molecular simulations. *Nucleic Acids Res.* **48**, 12593–12603 (2020).
557. Tejedor, A. R., Garaizar, A., Ramírez, J. & Espinosa, J. R. 'RNA modulation of transport properties and stability in phase-separated condensates. *Biophys. J.* **120**, 5169–5186 (2021).
558. Tesei, G., Schulze, T. K., Crehuet, R. & Lindorff-Larsen, K. Accurate model of liquid-liquid phase behaviour of intrinsically-disordered proteins from optimization of single-chain properties. Preprint at https://doi.org/10.1101/2021.06.23.449550.
559. Kapcha, L. H. & Rossky, P. J. A simple atomic-level hydrophobicity scale reveals protein interfacial structure. *J. Mol. Biol.* **426**, 484–498 (2014).
560. Kim, Y. C. & Hummer, G. Coarse-grained models for simulations of multiprotein complexes: application to ubiquitin binding. *J. Mol. Biol.* **375**, 1416–1433 (2008).
561. Vernon, R. M. *et al.* Pi-Pi contacts are an overlooked protein feature relevant to phase separation. *Elife* **7**, (2018).
562. Fisher, R. S. & Elbaum-Garfinkle, S. Tunable multiphase dynamics of arginine and lysine liquid condensates. *Nat. Commun.* **11**, 4628 (2020).
563. Overbeek, J. T. G. & Voorn, M. J. Phase separation in polyelectrolyte solutions. Theory of complex coacervation. *Journal of Cellular and Comparative Physiology* vol. 49 7–26 Preprint at https://doi.org/10.1002/jcp.1030490404 (1957).
564. Sing, C. E. Development of the modern theory of polymeric complex coacervation. *Adv. Colloid Interface Sci.* **239**, 2–16 (2017).
565. Sing, C. E. & Perry, S. L. Recent progress in the science of complex coacervation. *Soft Matter* **16**, 2885–2914 (2020).
566. Boeynaems, S. *et al.* Spontaneous driving forces give rise to protein−RNA condensates with coexisting phases and complex material properties. *Proceedings of the National Academy of Sciences*

vol. 116 7889–7898 Preprint at https://doi.org/10.1073/pnas.1821038116 (2019).
567. Alshareedah, I. *et al.* Interplay between Short-Range Attraction and Long-Range Repulsion Controls Reentrant Liquid Condensation of Ribonucleoprotein-RNA Complexes. *J. Am. Chem. Soc.* **141**, 14593–14602 (2019).
568. Laghmach, R. *et al.* RNA chain length and stoichiometry govern surface tension and stability of protein-RNA condensates. *iScience* **25**, 104105 (2022).
569. Alshareedah, I., Moosa, M. M., Pham, M., Potoyan, D. A. & Banerjee, P. R. Programmable viscoelasticity in protein-RNA condensates with disordered sticker-spacer polypeptides. *Nat. Commun.* **12**, 6620 (2021).
570. Cubuk, J. & Soranno, A. Macromolecular crowding and intrinsically disordered proteins: A polymer physics perspective. *ChemSystemsChem* (2022) doi:10.1002/syst.202100051.
571. Soranno, A. *et al.* Single-molecule spectroscopy reveals polymer effects of disordered proteins in crowded environments. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 4874–4879 (2014).
572. Dill, K. A. & Shortle, D. Denatured states of proteins. *Annu. Rev. Biochem.* **60**, 795–825 (1991).
573. Mao, A. H., Lyle, N. & Pappu, R. V. Describing sequence–ensemble relationships for intrinsically disordered proteins. *Biochem. J* **449**, 307–318 (2013).
574. Chan, H. S. & Dill, K. A. Polymer principles in protein structure and stability. *Annu. Rev. Biophys. Biophys. Chem.* **20**, 447–490 (1991).
575. Pappu, R. V., Wang, X., Vitalis, A. & Crick, S. L. A polymer physics perspective on driving forces and mechanisms for protein aggregation - Highlight Issue: Protein Folding. *Arch. Biochem. Biophys.* **469**, 132–141 (2008).
576. Lin, Y.-H., Forman-Kay, J. D. & Chan, H. S. Theories for Sequence-Dependent Phase Behaviors of Biomolecular Condensates. *Biochemistry* **57**, 2499–2508 (2018).
577. Thirumalai, D., O'Brien, E. P., Morrison, G. & Hyeon, C. Theoretical perspectives on protein folding. *Annu. Rev. Biophys.* **39**, 159–183 (2010).
578. Wilkins, D. K. *et al.* Hydrodynamic radii of native and denatured proteins measured by pulse field gradient NMR techniques. *Biochemistry* **38**, 16424–16431 (1999).
579. Damaschun, G., Damaschun, H., Gast, K., Zirwer, D. & Bychkova, V. E. Solvent dependence of dimensions of unfolded protein chains. *Int. J. Biol. Macromol.* **13**, 217–221 (1991).
580. Calmettes, P. *et al.* How random is a highly denatured protein? *Biophys. Chem.* **53**, 105–113 (1994).
581. Mok, Y. K., Kay, C. M., Kay, L. E. & Forman-Kay, J. NOE data demonstrating a compact unfolded state for an SH3 domain under non-denaturing conditions. *J. Mol. Biol.* **289**, 619–638 (1999).
582. Meng, W., Luan, B., Lyle, N., Pappu, R. V. & Raleigh, D. P. The Denatured State Ensemble Contains Significant Local and Long-Range Structure under Native Conditions: Analysis of the N-Terminal Domain of Ribosomal Protein L9. *Biochemistry* **52**, 2662–2671 (2013).
583. Bottaro, S. & Lindorff-Larsen, K. Biophysical experiments and biomolecular simulations: A perfect match? *Science* **361**, 355–360 (2018).
584. Lalmansingh, J. M., Keeley, A. T., Ruff, K. M., Pappu, R. V. & Holehouse, A. S. SOURSOP: A Python package for the analysis of simulations of intrinsically disordered proteins. *bioRxiv (in press*

at JCTC) (2023) doi:10.1101/2023.02.16.528879.
585. de Gennes, P. G. *Scaling concepts in polymer physics*. (Cornell University Press, 1979).
586. Flory, P. J. *Statistical Mechanics of Chain Molecules*. (Oxford University Press, 1969).
587. Song, J., Gomes, G.-N., Shi, T., Gradinaru, C. C. & Chan, H. S. Conformational Heterogeneity and FRET Data Interpretation for Dimensions of Unfolded Proteins. *Biophys. J.* **113**, 1012–1024 (2017).
588. Canchi, D. R. & García, A. E. Cosolvent effects on protein stability. *Annual Reviews of Physical Chemistry* **64**, 273–293 (2013).
589. Borgia, A. *et al.* Consistent View of Polypeptide Chain Expansion in Chemical Denaturants from Multiple Experimental Methods. *J. Am. Chem. Soc.* **138**, 11714–11726 (2016).
590. Tran, H. T. & Pappu, R. V. Toward an accurate theoretical framework for describing ensembles for proteins under strongly denaturing conditions. *Biophys. J.* **91**, 1868–1886 (2006).
591. Bernadó, P. & Svergun, D. I. Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering. *Mol. Biosyst.* **8**, 151–167 (2011).
592. Bremer, A. *et al.* Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. *Nat. Chem.* **14**, 196–207 (2022).
593. Martin, E. W. & Holehouse, A. S. Intrinsically disordered protein regions and phase separation: sequence determinants of assembly or lack thereof. *Emerg Top Life Sci* **4**, 307–329 (2020).
594. Huihui, J. & Ghosh, K. An analytical theory to describe sequence-specific inter-residue distance profiles for polyampholytes and intrinsically disordered proteins. *J. Chem. Phys.* **152**, 161102 (2020).
595. Das, S., Lin, Y.-H., Vernon, R. M., Forman-Kay, J. D. & Chan, H. S. Comparative roles of charge, π, and hydrophobic interactions in sequence-dependent phase separation of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 28795–28805 (2020).
596. Volkenstein, M. V. *Molecular Biophysics*. (Academic Press, New York, 1977).
597. Kentsis, A., Mezei, M., Gindin, T. & Osman, R. Unfolded state of polyalanine is a segmented polyproline II helix. *Proteins* **55**, 493–501 (2004).
598. Ruff, K. M. *et al.* Sequence grammar underlying the unfolding and phase separation of globular proteins. *Mol. Cell* **82**, 3193–3208.e8 (2022).
599. Lhuillier, D. A Simple-Model for Polymeric Fractals in a Good Solvent and an Improved Version of the Flory Approximation. *Journal De Physique* **49**, 705–710 (1988).
600. Kirkwood, J. G. & Riseman, J. The Intrinsic Viscosities and Diffusion Constants of Flexible Macromolecules in Solution. *J. Chem. Phys.* **16**, 565–573 (1948).
601. Pesce, F. *et al.* Assessment of models for calculating the hydrodynamic radius of intrinsically disordered proteins. *Biophys. J.* (2022) doi:10.1016/j.bpj.2022.12.013.
602. Sherry, K. P., Das, R. K., Pappu, R. V. & Barrick, D. Control of transcriptional activity by design of charge patterning in the intrinsically disordered RAM region of the Notch receptor. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E9243–E9252 (2017).
603. Das, R. K., Huang, Y., Phillips, A. H., Kriwacki, R. W. & Pappu, R. V. Cryptic sequence features within the disordered protein p27Kip1 regulate cell cycle signaling. *Proc. Natl. Acad. Sci. U. S. A.*

113, 5616–5621 (2016).
604. Marsh, J. A. *et al.* Improved structural characterizations of the drkN SH3 domain unfolded state suggest a compact ensemble with native-like and non-native structure. *J. Mol. Biol.* **367**, 1494–1510 (2007).
605. Bezsonova, I., Singer, A., Choy, W.-Y., Tollinger, M. & Forman-Kay, J. D. Structural comparison of the unstable drkN SH3 domain and a stable mutant. *Biochemistry* **44**, 15550–15560 (2005).
606. Demarest, S. J. *et al.* Mutual synergistic folding in recruitment of CBP/p300 by p160 nuclear receptor coactivators. *Nature* **415**, 549–553 (2002).
607. Wells, M. *et al.* Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 5762–5767 (2008).
608. Longhi, S. *et al.* The C-terminal domain of the measles virus nucleoprotein is intrinsically disordered and folds upon binding to the C-terminal moiety of the phosphoprotein. *J. Biol. Chem.* **278**, 18638–18648 (2003).
609. Syme, C. D. *et al.* A Raman optical activity study of rheomorphism in caseins, synucleins and tau. New insight into the structure and behaviour of natively unfolded proteins. *Eur. J. Biochem.* **269**, 148–156 (2002).
610. Theillet, F.-X. *et al.* Structural disorder of monomeric α-synuclein persists in mammalian cells. *Nature* **530**, 45–50 (2016).
611. Moses, D. *et al.* Structural biases in disordered proteins are prevalent in the cell. *bioRxiv* 2021.11.24.469609 (2022) doi:10.1101/2021.11.24.469609.
612. Mohanty, P. *et al.* Aliphatic residues contribute significantly to the phase separation of TDP-43 C-terminal domain. *bioRxiv* 2022.11.10.516004 (2022) doi:10.1101/2022.11.10.516004.
613. Murthy, A. C. *et al.* Molecular interactions contributing to FUS SYGQ LC-RGG phase separation and co-partitioning with RNA polymerase II heptads. *Nat. Struct. Mol. Biol.* **28**, 923–935 (2021).
614. Rekhi, S. *et al.* Role of Strong Localized vs. Weak Distributed Interactions in Disordered Protein Phase Separation. *bioRxiv* 2023.01.27.525976 (2023) doi:10.1101/2023.01.27.525976.
615. Griep, S. & Hobohm, U. PDBselect 1992–2009 and PDBfilter-select. *Nucleic Acids Res.* **38**, D318–D319 (2009).
616. Dill, K. & Bromberg, S. *Molecular driving forces: statistical thermodynamics in biology, chemistry, physics, and nanoscience.* (Garland Science, 2010).
617. Martin, E. W., Hopkins, J. B. & Mittag, T. Small-angle X-ray scattering experiments of monodisperse intrinsically disordered protein samples close to the solubility limit. *Methods Enzymol.* **646**, 185–222 (2021).
618. González-Foutel, N. S. *et al.* Conformational buffering underlies functional selection in intrinsically disordered protein regions. *Nat. Struct. Mol. Biol.* **29**, 781–790 (2022).
619. Zerze, G. H., Best, R. B. & Mittal, J. Sequence- and Temperature-Dependent Properties of Unfolded and Disordered Proteins from Atomistic Simulations. *J. Phys. Chem. B* **119**, 14622–14630 (2015).
620. Sørensen, C. S. & Kjaergaard, M. Effective concentrations enforced by intrinsically disordered linkers are governed by polymer physics. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 23124–23131 (2019).

621. Riback, J. A. *et al.* Stress-Triggered Phase Separation Is an Adaptive, Evolutionarily Tuned Response. *Cell* **168**, 1028–1040.e19 (2017).
622. Zhu, N. *et al.* A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N. Engl. J. Med.* **382**, 727–733 (2020).
623. Corman, V. M., Muth, D., Niemeyer, D. & Drosten, C. Chapter Eight - Hosts and Sources of Endemic Human Coronaviruses. in *Advances in Virus Research* (eds. Kielian, M., Mettenleiter, T. C. & Roossinck, M. J.) vol. 100 163–188 (Academic Press, 2018).
624. Roser, M., Ritchie, H., Ortiz-Ospina, E. & Hasell, J. Coronavirus Pandemic (COVID-19). *Our World in Data* (2020).
625. Lurie, N., Saville, M., Hatchett, R. & Halton, J. Developing Covid-19 Vaccines at Pandemic Speed. *N. Engl. J. Med.* **382**, 1969–1973 (2020).
626. Gordon, D. E. *et al.* A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* **583**, 459–468 (2020).
627. Sanders, J. M., Monogue, M. L., Jodlowski, T. Z. & Cutrell, J. B. Pharmacologic Treatments for Coronavirus Disease 2019 (COVID-19): A Review. *JAMA* (2020) doi:10.1001/jama.2020.6019.
628. Walls, A. C. *et al.* Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **181**, 281–292.e6 (2020).
629. Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **181**, 271–280.e8 (2020).
630. Shang, J. *et al.* Structural basis of receptor recognition by SARS-CoV-2. *Nature* **581**, 221–224 (2020).
631. Lan, J. *et al.* Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* **581**, 215–220 (2020).
632. Masters, P. S. Coronavirus genomic RNA packaging. *Virology* **537**, 198–207 (2019).
633. Laude, H. & Masters, P. S. The Coronavirus Nucleocapsid Protein. in *The Coronaviridae* (ed. Siddell, S. G.) 141–163 (Springer US, 1995).
634. Jack, A. *et al.* SARS CoV-2 nucleocapsid protein forms condensates with viral genomic RNA. 2020.09.14.295824 (2020) doi:10.1101/2020.09.14.295824.
635. Perdikari, T. M. *et al.* SARS-CoV-2 nucleocapsid protein phase-separates with RNA and with human hnRNPs. *EMBO J.* **n/a**, e106478 (2020).
636. Wu, C. *et al.* Characterization of SARS-CoV-2 N protein reveals multiple functional consequences of the C-terminal domain. *bioRxiv* 2020.11.30.404905 (2020).
637. Chen, H. *et al.* Liquid–liquid phase separation by SARS-CoV-2 nucleocapsid protein and RNA. *Cell Res.* **30**, 1143–1145 (2020).
638. McBride, R., van Zyl, M. & Fielding, B. C. The coronavirus nucleocapsid is a multifunctional protein. *Viruses* **6**, 2991–3018 (2014).
639. Chang, C.-K. *et al.* Multiple nucleic acid binding sites and intrinsic disorder of severe acute respiratory syndrome coronavirus nucleocapsid protein: implications for ribonucleocapsid protein packaging. *J. Virol.* **83**, 2255–2264 (2009).
640. Grossoehme, N. E. *et al.* Coronavirus N protein N-terminal domain (NTD) specifically binds the transcriptional regulatory sequence (TRS) and melts TRS-cTRS RNA duplexes. *J. Mol. Biol.* **394**,

544–557 (2009).
641. Cui, L. *et al.* The Nucleocapsid Protein of Coronaviruses Acts as a Viral Suppressor of RNA Silencing in Mammalian Cells. *J. Virol.* **89**, 9029–9043 (2015).
642. Takeda, M. *et al.* Solution structure of the c-terminal dimerization domain of SARS coronavirus nucleocapsid protein solved by the SAIL-NMR method. *J. Mol. Biol.* **380**, 608–622 (2008).
643. Jayaram, H. *et al.* X-ray structures of the N- and C-terminal domains of a coronavirus nucleocapsid protein: implications for nucleocapsid formation. *J. Virol.* **80**, 6612–6620 (2006).
644. Yu, I.-M. *et al.* Recombinant severe acute respiratory syndrome (SARS) coronavirus nucleocapsid protein forms a dimer through its C-terminal domain. *J. Biol. Chem.* **280**, 23280–23286 (2005).
645. Luo, H., Chen, J., Chen, K., Shen, X. & Jiang, H. Carboxyl terminus of severe acute respiratory syndrome coronavirus nucleocapsid protein: self-association analysis and nucleic acid binding characterization. *Biochemistry* **45**, 11827–11835 (2006).
646. Chang, C.-K., Chen, C.-M. M., Chiang, M.-H., Hsu, Y.-L. & Huang, T.-H. Transient oligomerization of the SARS-CoV N protein--implication for virus ribonucleoprotein packaging. *PLoS One* **8**, e65045 (2013).
647. Robbins, S. G., Frana, M. F., McGowan, J. J., Boyle, J. F. & Holmes, K. V. RNA-binding proteins of coronavirus MHV: detection of monomeric and multimeric N protein with an RNA overlay-protein blot assay. *Virology* **150**, 402–410 (1986).
648. He, R. *et al.* Analysis of multimerization of the SARS coronavirus nucleocapsid protein. *Biochem. Biophys. Res. Commun.* **316**, 476–483 (2004).
649. Kang, S. *et al.* Crystal structure of SARS-CoV-2 nucleocapsid protein RNA binding domain reveals potential unique drug targeting sites. *Acta Pharm Sin B* (2020) doi:10.1016/j.apsb.2020.04.009.
650. Zinzula, L., Nagy, M. O. & Bracher, A. 1.45 Angstrom Resolution Crystal Structure of C-terminal Dimerization Domain of Nucleocapsid Phosphoprotein from SARS-CoV-2 (PDB: 6YUN). *Protein Data Bank* (2020).
651. Ye, Q., West, A. M. V., Silletti, S. & Corbett, K. D. Architecture and self-assembly of the SARS-CoV -2 nucleocapsid protein. *Protein Sci.* **29**, 1890–1901 (2020).
652. Zeng, W. *et al.* Biochemical characterization of SARS-CoV-2 nucleocapsid protein. *Biochem. Biophys. Res. Commun.* **527**, 618–623 (2020).
653. Borgia, A. *et al.* Localizing internal friction along the reaction coordinate of protein folding by combining ensemble and single-molecule fluorescence spectroscopy. *Nat. Commun.* **3**, 1195 (2012).
654. Soranno, A., Cabassi, F. & Orselli, M. E. Dynamics of Structural Elements of GB1 β-Hairpin Revealed by Tryptophan–Cysteine Contact Formation Experiments. *The Journal of* (2018).
655. Schellman, J. A. Selective binding and solvent denaturation. *Biopolymers* **26**, 549–559 (1987).
656. Tompa, P. & Fuxreiter, M. Fuzzy complexes: polymorphism and structural disorder in protein–protein interactions. *Trends Biochem. Sci.* **33**, 2–8 (2008/1).
657. Haenni, D., Zosel, F., Reymond, L., Nettels, D. & Schuler, B. Intramolecular distances and dynamics from the combined photon statistics of single-molecule FRET and photoinduced electron transfer. *J. Phys. Chem. B* **117**, 13015–13028 (2013).

658. Brangwynne, C. P. *et al.* Germline P granules are liquid droplets that localize by controlled dissolution/condensation. *Science* **324**, 1729–1732 (2009).
659. Lin, Y., Protter, D. S. W., Rosen, M. K. & Parker, R. Formation and maturation of phase-separated liquid droplets by RNA-binding proteins. *Mol. Cell* **60**, 208–219 (2015).
660. Wang, J. *et al.* A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins. *Cell* **174**, 688–699.e16 (2018).
661. Stockmayer, W. H. Light Scattering in Multi-Component Systems. *J. Chem. Phys.* **18**, 58–61 (1950).
662. Posey, A. E., Holehouse, A. S. & Pappu, R. V. Chapter One - Phase Separation of Intrinsically Disordered Proteins. in *Methods in Enzymology* (ed. Rhoades, E.) vol. 611 1–30 (Academic Press, 2018).
663. Sanders, D. W. *et al.* Competing Protein-RNA Interaction Networks Control Multiphase Intracellular Organization. *Cell* **181**, 306–324.e28 (2020).
664. Riback, J. A. *et al.* Composition-dependent thermodynamics of intracellular phase separation. *Nature* **581**, 209–214 (2020).
665. Post, C. B. & Zimm, B. H. Internal condensation of a single DNA molecule. *Biopolymers* **18**, 1487–1501 (1979).
666. Hsieh, P.-K. *et al.* Assembly of severe acute respiratory syndrome coronavirus RNA packaging signal into virus-like particles is nucleocapsid dependent. *J. Virol.* **79**, 13848–13855 (2005).
667. Woo, K., Joo, M., Narayanan, K., Kim, K. H. & Makino, S. Murine coronavirus packaging signal confers packaging to nonviral RNA. *J. Virol.* **71**, 824–827 (1997).
668. Cologna, R. & Hogue, B. G. Identification of a bovine coronavirus packaging signal. *J. Virol.* **74**, 580–583 (2000).
669. Pool, R. & Bolhuis, P. G. Sampling the kinetic pathways of a micelle fusion and fission transition. *J. Chem. Phys.* **126**, 244703 (2007).
670. Denkova, A. G., Mendes, E. & Coppens, M.-O. Non-equilibrium dynamics of block copolymer micelles in solution: recent insights and open questions. *Soft Matter* **6**, 2351–2357 (2010).
671. Leyrat, C. *et al.* The N0-binding region of the vesicular stomatitis virus phosphoprotein is globally disordered but contains transient α-helices. *Protein Sci.* **20**, 542–556 (2011).
672. Feuerstein, S. *et al.* Transient structure and SH3 interaction sites in an intrinsically disordered fragment of the hepatitis C virus protein NS5A. *J. Mol. Biol.* **420**, 310–323 (2012).
673. Jensen, M. R. *et al.* Quantitative conformational analysis of partially folded proteins from residual dipolar couplings: application to the molecular recognition element of Sendai virus nucleoprotein. *J. Am. Chem. Soc.* **130**, 8055–8061 (2008).
674. Das, R. K., Crick, S. L. & Pappu, R. V. N-terminal segments modulate the α-helical propensities of the intrinsically disordered basic regions of bZIP proteins. *J. Mol. Biol.* **416**, 287–299 (2012).
675. Harmon, T. S. *et al.* GADIS: Algorithm for designing sequences to achieve target secondary structure profiles of intrinsically disordered proteins. *Protein Eng. Des. Sel.* **29**, 339–346 (2016).
676. Bayer, T. S., Booth, L. N., Knudsen, S. M. & Ellington, A. D. Arginine-rich motifs present multiple interfaces for specific binding by RNA. *RNA* **11**, 1848–1857 (2005).
677. Battiste, J. L. *et al.* Alpha helix-RNA major groove recognition in an HIV-1 rev peptide-RRE RNA

complex. *Science* **273**, 1547–1551 (1996).
678. Hurst, K. R., Koetzner, C. A. & Masters, P. S. Characterization of a critical interaction between the coronavirus nucleocapsid protein and nonstructural protein 3 of the viral replicase-transcriptase complex. *J. Virol.* **87**, 9159–9172 (2013).
679. Hurst, K. R., Ye, R., Goebel, S. J., Jayaraman, P. & Masters, P. S. An Interaction between the Nucleocapsid Protein and a Component of the Replicase-Transcriptase Complex Is Crucial for the Infectivity of Coronavirus Genomic RNA. *J. Virol.* **84**, 10276–10288 (2010).
680. Verheije, M. H. *et al.* The coronavirus nucleocapsid protein is dynamically associated with the replication-transcription complexes. *J. Virol.* **84**, 11575–11579 (2010).
681. Surjit, M. *et al.* The severe acute respiratory syndrome coronavirus nucleocapsid protein is phosphorylated and localizes in the cytoplasm by 14-3-3-mediated translocation. *J. Virol.* **79**, 11476–11486 (2005).
682. Timani, K. A. *et al.* Nuclear/nucleolar localization properties of C-terminal nucleocapsid protein of SARS coronavirus. *Virus Res.* **114**, 23–34 (2005).
683. Kuo, L. & Masters, P. S. Genetic evidence for a structural interaction between the carboxy termini of the membrane and nucleocapsid proteins of mouse hepatitis virus. *J. Virol.* **76**, 4987–4999 (2002).
684. Hurst, K. R. *et al.* A major determinant for membrane protein interaction localizes to the carboxy-terminal domain of the mouse coronavirus nucleocapsid protein. *J. Virol.* **79**, 13285–13297 (2005).
685. Verma, S., Bednar, V., Blount, A. & Hogue, B. G. Identification of functionally important negatively charged residues in the carboxy end of mouse hepatitis coronavirus A59 nucleocapsid protein. *J. Virol.* **80**, 4344–4355 (2006).
686. Brass, V. *et al.* An Amino-terminal Amphipathic α-Helix Mediates Membrane Association of the Hepatitis C Virus Nonstructural Protein 5A. *J. Biol. Chem.* **277**, 8130–8139 (2002).
687. Braun, A. R., Lacy, M. M., Ducas, V. C., Rhoades, E. & Sachs, J. N. α-Synuclein's Uniquely Long Amphipathic Helix Enhances its Membrane Binding and Remodeling Capacity. *J. Membr. Biol.* **250**, 183–193 (2017).
688. Wyman, J. & Gill, S. J. *Binding and Linkage: Functional Chemistry of Biological Macromolecules.* (University Science Books, 1990).
689. Metrick, C. M., Koenigsberg, A. L. & Heldwein, E. E. Conserved Outer Tegument Component UL11 from Herpes Simplex Virus 1 Is an Intrinsically Disordered, RNA-Binding Protein. *MBio* **11**, (2020).
690. Monette, A. *et al.* Pan-retroviral Nucleocapsid-Mediated Phase Separation Regulates Genomic RNA Positioning and Trafficking. *Cell Rep.* **31**, 107520 (2020).
691. Monette, A. & Mouland, A. J. Zinc and Copper Ions Differentially Regulate Prion-Like Phase Separation Dynamics of Pan-Virus Nucleocapsid Biomolecular Condensates. *Viruses* **12**, (2020).
692. Klein, S. *et al.* SARS-CoV-2 structure and replication characterized by in situ cryo-electron tomography. *Nat. Commun.* **11**, 5885 (2020).
693. Cong, Y., Kriegenburg, F., de Haan, C. A. M. & Reggiori, F. Coronavirus nucleocapsid proteins

assemble constitutively in high molecular oligomers. *Sci. Rep.* **7**, 5740 (2017).
694. Chang, C.-K., Hou, M.-H., Chang, C.-F., Hsiao, C.-D. & Huang, T.-H. The SARS coronavirus nucleocapsid protein--forms and functions. *Antiviral Res.* **103**, 39–50 (2014).
695. Borodavka, A., Tuma, R. & Stockley, P. G. A two-stage mechanism of viral RNA compaction revealed by single molecule fluorescence. *RNA Biol.* **10**, 481–489 (2013).
696. He, R. *et al.* Characterization of protein–protein interactions between the nucleocapsid protein and membrane protein of the SARS coronavirus. *Virus Res.* **105**, 121–125 (2004).
697. Bergeron-Sandoval, L.-P. *et al.* Endocytosis caused by liquid-liquid phase separation of proteins. *bioRxiv* 145664 (2018) doi:10.1101/145664.
698. Bergeron-Sandoval, L.-P. & Michnick, S. W. Mechanics, Structure and Function of Biopolymer Condensates. *J. Mol. Biol.* **430**, 4754–4761 (2018).
699. Holmstrom, E. D., Nettels, D. & Schuler, B. Conformational Plasticity of Hepatitis C Virus Core Protein Enables RNA-Induced Formation of Nucleocapsid-like Particles. *J. Mol. Biol.* **430**, 2453–2467 (2018).
700. Rodríguez, L., Cuesta, I., Asenjo, A. & Villanueva, N. Human respiratory syncytial virus matrix protein is an RNA-binding protein: binding properties, location and identity of the RNA contact residues. *J. Gen. Virol.* **85**, 709–719 (2004).
701. Linger, B. R., Kunovska, L., Kuhn, R. J. & Golden, B. L. Sindbis virus nucleocapsid assembly: RNA folding promotes capsid protein dimerization. *RNA* **10**, 128–138 (2004).
702. Zúñiga, S. *et al.* Coronavirus nucleocapsid protein is an RNA chaperone. *Virology* **357**, 215–227 (2007).
703. Luo, H. *et al.* The nucleocapsid protein of SARS coronavirus has a high binding affinity to the human cellular heterogeneous nuclear ribonucleoprotein A1. *FEBS Lett.* **579**, 2623–2628 (2005).
704. Yang, P. *et al.* G3BP1 Is a Tunable Switch that Triggers Phase Separation to Assemble Stress Granules. *Cell* **181**, 325–345.e28 (2020).
705. Nabeel-Shah, S. *et al.* SARS-CoV-2 Nucleocapsid protein attenuates stress granule formation and alters gene expression via direct interaction with host mRNAs. *bioRxiv* (2020) doi:10.1101/2020.10.23.342113.
706. van Rosmalen, M. G. M. *et al.* Revealing in real-time a multistep assembly mechanism for SV40 virus-like particles. *Science Advances* **6**, eaaz1639 (2020).
707. Patel, A. *et al.* A liquid-to-solid phase transition of the ALS protein FUS accelerated by disease mutation. *Cell* **162**, 1066–1077 (2015).
708. Alberti, S. & Dormann, D. Liquid-Liquid Phase Separation in Disease. *Annu. Rev. Genet.* **53**, 171–194 (2019).
709. Weber, S. C. & Brangwynne, C. P. Getting RNA and protein in phase. *Cell* **149**, 1188–1191 (2012).
710. Dignon, G. L., Zheng, W., Best, R. B., Kim, Y. C. & Mittal, J. Relation between single-molecule properties and phase behavior of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 9929–9934 (2018).
711. Mészáros, B., Erdos, G. & Dosztányi, Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* **46**, W329–W337

(2018).
712. Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 10037–10041 (2001).
713. Holehouse, A. S. *protfasta*. (2020). doi:10.5281/zenodo.3862728.
714. Mao, A. H. & Pappu, R. V. Crystal lattice properties fully determine short-range interaction parameters for alkali and halide ions. *J. Chem. Phys.* **137**, 064104 (2012).
715. McGibbon, R. T. *et al.* MDTraj: a modern, open library for the analysis of molecular dynamics trajectories. *Biophys. J.* **109**, 1528–1532 (2015).
716. Ortega, E. *et al.* Transcription factor dimerization activates the p300 acetyltransferase. *Nature* **562**, 538–544 (2018).
717. Guex, N. & Peitsch, M. C. SWISS-MODEL and the Swiss-Pdb Viewer: an environment for comparative protein modeling. *Electrophoresis* **18**, 2714–2723 (1997).
718. Duan, Y. *et al.* A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **24**, 1999–2012 (2003).
719. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015/9).
720. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).
721. Feenstra, K. A., Hess, B. & Berendsen, H. J. C. Improving efficiency of large time-scale molecular dynamics simulations of hydrogen-rich systems. *J. Comput. Chem.* **20**, 786–798 (1999).
722. Hess, B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *J. Chem. Theory Comput.* **4**, 116–122 (2008).
723. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126**, 014101 (2007).
724. Zimmerman, M. I., Porter, J. R., Sun, X., Silva, R. R. & Bowman, G. R. Choice of Adaptive Sampling Strategy Impacts State Discovery, Transition Probabilities, and the Apparent Mechanism of Conformational Changes. *J. Chem. Theory Comput.* **14**, 5459–5475 (2018).
725. Zimmerman, M. I. & Bowman, G. R. FAST Conformational Searches by Balancing Exploration/Exploitation Trade-Offs. *J. Chem. Theory Comput.* **11**, 5747–5757 (2015).
726. Zimmerman, M. I. *et al.* Prediction of New Stabilizing Mutations Based on Mechanistic Insights from Markov State Models. *ACS Cent Sci* **3**, 1311–1321 (2017).
727. Porter, J. R., Zimmerman, M. I. & Bowman, G. R. Enspara: Modeling molecular ensembles with scalable data structures and parallel computing. *J. Chem. Phys.* **150**, 044108 (2019).
728. Boeynaems, S. *et al.* Spontaneous driving forces give rise to protein-RNA condensates with coexisting phases and complex material properties. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 7889–7898 (2019).
729. Fichthorn, K. A. & Weinberg, W. H. Theoretical foundations of dynamical Monte Carlo simulations. *J. Chem. Phys.* **95**, 1090–1096 (1991).
730. Bieler, N. S., Knowles, T. P. J., Frenkel, D. & Vácha, R. Connecting macroscopic observables and

microscopic assembly events in amyloid formation using coarse grained simulations. *PLoS Comput. Biol.* **8**, e1002692 (2012).
731. Šarić, A. *et al.* Physical determinants of the self-replication of protein fibrils. *Nat. Phys.* **12**, 874–880 (2016).
732. Schuler, B., Müller-Späth, S., Soranno, A. & Nettels, D. Application of confocal single-molecule FRET to intrinsically disordered proteins. *Methods Mol. Biol.* **896**, 21–45 (2012).
733. Hoffmann, A. *et al.* Mapping protein collapse with single-molecule fluorescence and kinetic synchrotron radiation circular dichroism spectroscopy. *Proceedings of the National Academy of Sciences* **104**, 105–110 (2007).
734. Rigler, R., Mets, Ü., Widengren, J. & Kask, P. Fluorescence correlation spectroscopy with high count rate and low background: analysis of translational diffusion. *Eur. Biophys. J.* **22**, 169–175 (1993).
735. Krichevsky, O. & Bonnet, G. Fluorescence correlation spectroscopy: the technique and its applications. *Rep. Prog. Phys.* **65**, 251 (2002).
736. Nettels, D., Hoffmann, A. & Schuler, B. Unfolded protein and peptide dynamics investigated with single-molecule FRET and correlation spectroscopy from picoseconds to seconds. *J. Phys. Chem. B* **112**, 6137–6146 (2008).
737. Higgs, P. G. & Joanny, J. Theory of polyampholyte solutions. *J. Chem. Phys.* **94**, 1543–1554 (1991).
738. Michelson, A. M. 270. Polynucleotides. Part I. Synthesis and properties of some polyribonucleotides. *J. Chem. Soc.* 1371–1394 (1959).
739. Pace, C. N., Vajdos, F., Fee, L., Grimsley, G. & Gray, T. How to measure and predict the molar absorption coefficient of a protein. *Protein Sci.* **4**, 2411–2423 (1995).
740. Meng, W., Lyle, N., Luan, B., Raleigh, D. P. & Pappu, R. V. Experiments and simulations show how long-range contacts can form in expanded unfolded proteins with negligible secondary structure. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 2123–2128 (2013).
741. McGuffin, L. J., Bryson, K. & Jones, D. T. The PSIPRED protein structure prediction server. *Bioinformatics* **16**, 404–405 (2000).
742. Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* **395**, 565–574 (2020).
743. Laue, M. *et al.* Morphometry of SARS-CoV and SARS-CoV-2 particles in ultrathin plastic sections of infected Vero cell cultures. *Sci. Rep.* **11**, 3515 (2021).
744. Yao, H. *et al.* Molecular Architecture of the SARS-CoV-2 Virus. *Cell* **183**, 730–738.e13 (2020).
745. Bárcena, M. *et al.* Cryo-electron tomography of mouse hepatitis virus: Insights into the structure of the coronavirion. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 582–587 (2009).
746. Macneughton, M. R. & Davies, H. A. Ribonucleoprotein-like structures from coronavirus particles. *J. Gen. Virol.* **39**, 545–549 (1978).
747. Perdikari, T. M. *et al.* SARS-CoV-2 nucleocapsid protein undergoes liquid-liquid phase separation stimulated by RNA and partitions into phases of human ribonucleoproteins. Preprint at https://doi.org/10.1101/2020.06.09.141101.
748. Wu, Y. *et al.* RNA-induced liquid phase separation of SARS-CoV-2 nucleocapsid protein facilitates

NF-κB hyper-activation and inflammation. *Signal Transduct Target Ther* **6**, 167 (2021).
749. Roden, C. A. *et al.* Double-stranded RNA drives SARS-CoV-2 nucleocapsid protein to undergo phase separation at specific temperatures. *bioRxiv* (2021) doi:10.1101/2021.06.14.448452.
750. Seim, I., Roden, C. A. & Gladfelter, A. S. Role of spatial patterning of N-protein interactions in SARS-CoV-2 genome packaging. *Biophys. J.* **120**, 2771–2784 (2021).
751. Jack, A. *et al.* SARS-CoV-2 nucleocapsid protein forms condensates with viral genomic RNA. *PLoS Biol.* **19**, e3001425 (2021).
752. Levy, Y., Onuchic, J. N. & Wolynes, P. G. Fly-casting in protein-DNA binding: frustration between protein folding and electrostatics facilitates target recognition. *J. Am. Chem. Soc.* **129**, 738–739 (2007).
753. Kruse, T. *et al.* Large scale discovery of coronavirus-host factor protein interaction motifs reveals SARS-CoV-2 specific mechanisms and vulnerabilities. *Nat. Commun.* **12**, 6761 (2021).
754. Dupuis, N. F., Holmstrom, E. D. & Nesbitt, D. J. Molecular-crowding effects on single-molecule RNA folding/unfolding thermodynamics and kinetics. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 8464–8469 (2014).
755. Sen, S. & Holmstrom, E. D. A single-molecule FRET approach for investigating the binding mechanisms of anti-viral aptamers. in *RNA Nanotechnology and Therapeutics* 193–209 (CRC Press, 2022).
756. Feng, X. A., Poyton, M. F. & Ha, T. Multicolor single-molecule FRET for DNA and RNA processes. *Curr. Opin. Struct. Biol.* **70**, 26–33 (2021).
757. Nabeel-Shah, S. *et al.* SARS-CoV-2 nucleocapsid protein binds host mRNAs and attenuates stress granules to impair host stress response. *iScience* **25**, 103562 (2022).
758. Xiang, J. S. *et al.* Discovery and functional interrogation of SARS-CoV-2 protein-RNA interactions. Preprint at https://doi.org/10.1101/2022.02.21.481223.
759. Joseph, J. A. *et al.* Physics-driven coarse-grained model for biomolecular phase separation with near-quantitative accuracy. *Biophysical Journal* vol. 121 307a Preprint at https://doi.org/10.1016/j.bpj.2021.11.1214 (2022).
760. Sanchez-Burgos, I., Espinosa, J. R., Joseph, J. A. & Collepardo-Guevara, R. RNA length has a non-trivial effect in the stability of biomolecular condensates formed by RNA-binding proteins. *PLoS Comput. Biol.* **18**, e1009810 (2022).
761. Zimmerman, M. I. *et al.* SARS-CoV-2 simulations go exascale to predict dramatic spike opening and cryptic pockets across the proteome. *Nature Chemistry* vol. 13 651–659 Preprint at https://doi.org/10.1038/s41557-021-00707-0 (2021).
762. Zhu, H. *et al.* The Chromatin Regulator HMGA1a Undergoes Phase Separation in the Nucleus. *Chembiochem* **24**, e202200450 (2023).
763. Tejedor, A. R. *et al.* Protein structural transitions critically transform the network connectivity and viscoelasticity of RNA-binding protein condensates but RNA can prevent it. *Nat. Commun.* **13**, 5717 (2022).
764. Manning, G. S. Limiting Laws and Counterion Condensation in Polyelectrolyte Solutions I. Colligative Properties. *J. Chem. Phys.* **51**, 924–933 (1969).

765. Vander Meulen & Saecker. Formation of a wrapped DNA–protein interface: experimental characterization and analysis of the large contributions of ions and water to the thermodynamics …. *J. Mol. Appl. Genet.*

766. Korn, S., Dhamotharan, K. & Schlundt, A. The preference signature of the SARS-CoV-2 Nucleocapsid NTD for its 5'-genomic RNA elements. *Research Square* (2022) doi:10.21203/rs.3.rs-1445747/v1.

767. Chen, S.-C. & Olsthoorn, R. C. L. Group-specific structural features of the 5'-proximal sequences of coronavirus genomic RNAs. *Virology* **401**, 29–41 (2010).

768. Zhao, H. *et al.* Plasticity in structure and assembly of SARS-CoV-2 nucleocapsid protein. *PNAS Nexus* **1**, gac049 (2022).

769. Gangavarapu, K. *et al.* Outbreak.info genomic reports: scalable and dynamic surveillance of SARS-CoV-2 variants and mutations. *bioRxiv* (2022) doi:10.1101/2022.01.27.22269965.

770. Taneja, I. & Holehouse, A. S. Folded domain charge properties influence the conformational behavior of disordered tails. *Curr Res Struct Biol* **3**, 216–228 (2021).

771. Holmstrom, Nettels & Sottini. Binding without folding–the biomolecular function of disordered polyelectrolyte complexes. *Curr. Opin. Allergy Clin. Immunol.*

772. Sottini, A. *et al.* Polyelectrolyte interactions enable rapid association and dissociation in high-affinity disordered protein complexes. *Nat. Commun.* **11**, 5736 (2020).

773. Wu, C. *et al.* Characterization of SARS-CoV-2 nucleocapsid protein reveals multiple functional consequences of the C-terminal domain. *iScience* vol. 24 102681 Preprint at https://doi.org/10.1016/j.isci.2021.102681 (2021).

774. Shkel, I. A., Ballin, J. D. & Record, M. T., Jr. Interactions of cationic ligands and proteins with small nucleic acids: analytic treatment of the large coulombic end effect on binding free energy as a function of salt concentration. *Biochemistry* **45**, 8411–8426 (2006).

775. Luo, H. & Sharp, K. On the calculation of absolute macromolecular binding free energies. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 10399–10404 (2002).

776. Wang, X., Ramírez-Hinestrosa, S., Dobnikar, J. & Frenkel, D. The Lennard-Jones potential: when (not) to use it. *Phys. Chem. Chem. Phys.* **22**, 10624–10633 (2020).

777. Tesei, G., Schulze, T. K., Crehuet, R. & Lindorff-Larsen, K. Accurate model of liquid–liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).

778. Siebenmorgen, T. & Zacharias, M. Computational prediction of protein–protein binding affinities. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **10**, (2020).

779. C. Branden, J. Tooze, Introduction to protein structure, Garland Pub. *Inc., New York* (1991).

780. I. A. Adzhubei, *et al.*, A method and server for predicting damaging missense mutations. *Nat. Methods* 7, 248–249 (2010).

781. K. E. Samocha, *et al.*, A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* 46, 944–950 (2014).

782. B. W. Matthews, Structural and genetic analysis of protein stability. *Annu. Rev. Biochem.* 62, 139–160

(1993).

783. P. E. Wright, H. J. Dyson, Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **293**, 321–331 (1999).

784. R. van der Lee, *et al.*, Classification of intrinsically disordered regions and proteins. *Chem. Rev.* **114**, 6589–6631 (2014).

785. A. K. Dunker, *et al.*, Intrinsically disordered protein. *J. Mol. Graph. Model.* **19**, 26–59 (2001).

786. V. N. Uversky, Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.* **11**, 739–756 (2002).

787. F. Wiggers, *et al.*, Diffusion of a disordered protein on its folded ligand. *Proc. Natl. Acad. Sci. U. S. A.* 118 (2021).

788. M. D. Stuchell-Brereton, *et al.*, Apolipoprotein E4 has extensive conformational heterogeneity in lipid-free and lipid-bound forms. *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2215371120 (2023).

789. R. D. Finn, J. Clements, S. R. Eddy, HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–37 (2011).

790. R. D. Finn, *et al.*, Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–30 (2014).

791. A. G. Murzin, S. E. Brenner, T. Hubbard, C. Chothia, SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540 (1995).

792. C. A. Orengo, *et al.*, The CATH Database provides insights into protein structure/function relationships. *Nucleic Acids Res.* **27**, 275–279 (1999).

793. S. McGinnis, T. L. Madden, BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* **32**, W20–5 (2004).

794. O. Olmea, B. Rost, A. Valencia, Effective use of sequence correlation and conservation in fold recognition. *J. Mol. Biol.* **293**, 1221–1239 (1999).

795. K. Tunyasuvunakool, *et al.*, Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).

796. I. R. Humphreys, *et al.*, Computed structures of core eukaryotic protein complexes. *Science* **374**, eabm4805 (2021).

797. I. Langstein-Skora, *et al.*, Sequence- and chemical specificity define the functional landscape of intrinsically disordered regions. *bioRxiv*, 2022.02.10.480018 (2022).

798. T. Zarin, *et al.*, Proteome-wide signatures of function in highly diverged intrinsically disordered regions. *Elife* **8** (2019).

799. M. C. Cohan, M. K. Shinn, J. M. Lalmansingh, R. V. Pappu, Uncovering Non-random Binary Patterns Within Sequences of Intrinsically Disordered Proteins. *J. Mol. Biol.* **434**, 167373 (2022).

800. M. Kumar, *et al.*, The Eukaryotic Linear Motif resource: 2022 release. *Nucleic Acids Res.* **50**, D497–D508 (2022).

801. A. G. Sangster, T. Zarin, A. M. Moses, Evolution of short linear motifs and disordered proteins Topic: yeast as model system to study evolution. *Curr. Opin. Genet. Dev.* **76**, 101964 (2022).

802. N. E. Davey, *et al.*, Attributes of short linear motifs. *Mol. Biosyst.* **8**, 268–281 (2012).

803. T. Zarin, C. N. Tsai, A. N. Nguyen Ba, A. M. Moses, Selection maintains signaling function of a highly diverged intrinsically disordered region. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E1450–E1459 (2017).

804. A. X. Lu, *et al.*, Discovering molecular features of intrinsically disordered regions by using evolution for contrastive learning. *PLoS Comput. Biol.* **18**, e1010238 (2022).

805. T. Zarin, *et al.*, Identifying molecular features that are associated with biological function of intrinsically disordered protein regions. *Elife* **10**, e60220 (2021).

806. A. Toth-Petroczy, *et al.*, Structured States of Disordered Proteins from Genomic Sequences. *Cell* **167**, 158–170.e12 (2016).

807. B. Xue, A. K. Dunker, V. N. Uversky, Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. *J. Biomol. Struct. Dyn.* **30**, 137–149 (2012).

808. F. Mihalič, *et al.*, Large-scale phage-based screening reveals extensive pan-viral mimicry of host short linear motifs. *Nat. Commun.* **14**, 2409 (2023).

809. H. J. Dyson, Vital for Viruses: Intrinsically Disordered Proteins. *J. Mol. Biol.*, 167860 (2022).

810. E. Domingo, *et al.*, Basic concepts in RNA virus evolution. *FASEB J.* **10**, 859–864 (1996).

811. R. W. Hendrix, J. G. Lawrence, G. F. Hatfull, S. Casjens, The origins and ongoing evolution of viruses. *Trends Microbiol.* **8**, 504–508 (2000).

812. E. V. Koonin, V. V. Dolja, M. Krupovic, Origins and evolution of viruses of eukaryotes: The ultimate modularity. *Virology* 479-480, 2–25 (2015).

813. P. S. Masters, The molecular biology of coronaviruses. *Adv. Virus Res.* 66, 193–292 (2006).

814. J. Cui, F. Li, Z.-L. Shi, Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol.* **17**, 181–192 (2019).

815. A. R. Fehr, S. Perlman, Coronaviruses: an overview of their replication and pathogenesis. *Methods Mol. Biol.* **1282**, 1–23 (2015).

816. J. J. Alston, A. Soranno, Condensation goes viral: a polymer physics perspective. *J. Mol. Biol.*, 167988 (2023).

817. Thompson, A. P. *et al.* LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comput. Phys. Commun.* **271**, 108171 (2022).

818. Chen, H. *et al.* Ionic strength-dependent persistence lengths of single-stranded RNA and DNA. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 799–804 (2012).

819. Cubuk, J. *et al.* The disordered N-terminal tail of SARS CoV-2 Nucleocapsid protein forms a dynamic complex with RNA. *bioRxiv* 2023.02.10.527914 (2023) doi:10.1101/2023.02.10.527914.

820. Tomasello, G., Armenia, I. & Molla, G. The Protein Imager: a full-featured online molecular viewer interface with server-side HQ-rendering capabilities. *Bioinformatics* **36**, 2909–2911 (2020).

821. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph. Model.* **14**, 33–8, 27–8 (1996).

822. Schneidman-Duhovny, D., Hammel, M., Tainer, J. A. & Sali, A. Accurate SAXS profile computation and its assessment by contrast variation experiments. *Biophys. J.* **105**, 962–974 (2013).

823. Gussow, A. B. *et al.* Genomic determinants of pathogenicity in SARS-CoV-2 and other human coronaviruses. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 15193–15199 (2020).

824. Peng, Y. *et al.* Structures of the SARS-CoV-2 nucleocapsid and their perspectives for drug design. *EMBO J.* **39**, e105938 (2020).

825. Terry, J. S. *et al.* Development of a SARS-CoV-2 nucleocapsid specific monoclonal antibody. *Virology* **558**, 28–37 (2021).

826. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).

827. Boeynaems, S. *et al.* Aberrant phase separation is a common killing strategy of positively charged peptides in biology and human disease. *bioRxiv* (2023) doi:10.1101/2023.03.09.531820.

828. Yang, Z., Deng, X., Liu, Y., Gong, W. & Li, C. Analyses on clustering of the conserved residues at protein-RNA interfaces and its application in binding site identification. *BMC Bioinformatics* **21**, 1–14 (2020).

829. Jankowski, M. S. *et al.* The formation of a fuzzy complex in the negative arm regulates the robustness of the circadian clock. *bioRxiv* 2022.01.04.474980 (2022) doi:10.1101/2022.01.04.474980.

830. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods*

17, 261–272 (2020).
831. Shazman, S. & Mandel-Gutfreund, Y. Classifying RNA-binding proteins based on electrostatic properties. *PLoS Comput. Biol.* **4**, e1000146 (2008).
832. Hsu, I. S. *et al.* A functionally divergent intrinsically disordered region underlying the conservation of stochastic signaling. *PLoS Genet.* **17**, e1009629 (2021).
833. Kuo, L., Koetzner, C. A. & Masters, P. S. A key role for the carboxy-terminal tail of the murine coronavirus nucleocapsid protein in coordination of genome packaging. *Virology* **494**, 100–107 (2016).
834. Schütz, S. *et al.* The Disordered MAX N-terminus Modulates DNA Binding of the Transcription Factor MYC:MAX. *J. Mol. Biol.* **434**, 167833 (2022).
835. Guo, X., Bulyk, M. L. & Hartemink, A. J. Intrinsic disorder within and flanking the DNA-binding domains of human transcription factors. *Pac. Symp. Biocomput.* 104–115 (2012).
836. Periole, X., Cavalli, M., Marrink, S.-J. & Ceruso, M. A. Combining an Elastic Network With a Coarse-Grained Molecular Force Field: Structure, Dynamics, and Intermolecular Recognition. *J. Chem. Theory Comput.* **5**, 2531–2543 (2009).
837. Zhang, Z., Pfaendtner, J., Grafmüller, A. & Voth, G. A. Defining coarse-grained representations of large biomolecules and biomolecular complexes from elastic network models. *Biophys. J.* **97**, 2327–2337 (2009).
838. Etibor, T. A. *et al.* Defining basic rules for hardening influenza A virus liquid condensates. *Elife* **12**, (2023).