Single-molecule Spectroscopy of the SARS-CoV-2 Nucleocapsid Protein

Jasmine Cubuk

Follow this and additional works at: https://openscholarship.wustl.edu/art_sci_etds

Recommended Citation

This Dissertation is brought to you for free and open access by the Arts & Sciences at Washington University Open Scholarship. It has been accepted for inclusion in Arts & Sciences Electronic Theses and Dissertations by an authorized administrator of Washington University Open Scholarship. For more information, please contact digital@wumail.wustl.edu.
WASHINGTON UNIVERSITY IN ST. LOUIS
Division of Biology and Biomedical Sciences
Biochemistry, Biophysics, and Structural Biology

Dissertation Examination Committee:
Andrea Soranno, Chair
Yuna Ayala
Kathleen Hall
Tim Lohman
Janice Robertson

Single-molecule Spectroscopy of the SARS-CoV-2 Nucleocapsid Protein
by
Jasmine Cubuk

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

May 2023
St. Louis, Missouri
# Table of Contents

List of Figures x
List of Tables xvi
Acknowledgments xviii
Abstract xxiii

## Chapter 1: The SARS-CoV-2 Nucleocapsid protein

1.1 The SARS-CoV-2 pandemic 2
1.2 Viruses 2
1.3 The family Coronaviridae 3
1.4 The SARS-CoV-2 Virus 4
1.5 SarS-CoV Nucleocapsid protein 9
1.6 The SARS-CoV-2 Nucleocapsid protein 11
1.7 Scope of the thesis 13
1.8 References 17

## Chapter 2: The biophysics of disordered proteins from the point of view of single-molecule fluorescence spectroscopy

2.1 Abstract 25
2.2 Introduction 25
2.3 Single-molecule Fluorescence 27
2.4 Accessing the disordered state 34
2.5 Effect of sequence on chain conformations 38
2.6 Effect of Temperature 40
2.7 Accessing chain dynamics 41
2.8 Accessing ligand interactions 43
2.9 IDPs in complex environment 44
Chapter 3: The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA

3.1 Abstract
3.2 Introduction
3.3 Results
  3.3.1 The NTD is disordered, flexible, and transiently interacts with the RBD
  3.3.2 The linker is highly dynamic and there is minimal interaction between the RBD and the dimerization domain
  3.3.3 The CTD engages in transient but non-negligible interactions with the dimerization domain
  3.3.4 N protein undergoes phase separation with RNA
  3.3.5 A simple polymer model shows symmetry-breaking can facilitate multiple metastable single-polymer condensates instead of a single multi-polymer condensate
3.4 Discussion
  3.4.1 All three IDRs are highly dynamic
  3.4.2 Simulations identify multiple transient helices
  3.4.3 The physiological relevance of nucleocapsid protein phase separation in SARS-CoV-2 physiology
  3.4.4 The physics of single polymer condensates
3.5 Acknowledgements
3.6 Methods
  3.6.1 All-atom Simulation
3.6.2 Coarse-grained polymer simulations 96
3.6.3 Protein Expression, Purification, and Labeleing 97
3.6.4 Single-molecule fluorescence spectroscopy 98

3.7 Supplementary Material 99
3.7.1 Sequence Analysis 99
3.7.2 All-atom Monte Carlo Simulations 99
3.7.3 Molecular Dynamics Simulations 102
3.7.4 Sequential Molecular Dynamics + Monte Carlo Sampling Approach 104
3.7.5 Coarse-Grained Polymer Simulations 105
3.7.6 Extended Discussion on Coarse-Grained Simulations 106
3.7.7 Plasmid Construct Design 107
3.7.8 Protein Expression and Purification 108
3.7.9 Choice of labeling positions 110
3.7.10 Fluorescent Dye Labeling 111
3.7.11 Experimental Setup and Procedure 112
3.7.12 PEGylation of Glass Surfaces 114
3.7.13 FRET Efficiency Histograms 115
3.7.14 Fluorescence Lifetimes and Anisotropies Analysis 118
3.7.15 Fluorescence Correlation Spectroscopy (FCS) Analysis 119
3.7.16 Nanosecond Fluorescence Correlation Spectroscopy 121
3.7.17 Polymer Models of Distance Distributions 122
3.7.18 Binding of Denaturant and Folding 123
3.7.19 Folding of RBD Domain 124
3.7.20 Salt Dependence of NTD, LINK, and CTD Conformations 125
Chapter 4: The dimerization domain of SARS CoV 2 Nucleocapsid protein is partially disordered as a monomer and forms a high affinity dynamic complex.

4.1 Abstract 172
4.2 Introduction 173
4.3 Methods 174
4.3.1 Single-molecule FRET 174
4.4 Results 175
4.4.1 The Dimerization Domain is partially disordered in the monomer form 176
4.4.2 Dimer formation induces folding of the DD domain 178
4.4.3 Denaturant effect on dimer stability 180
4.4.4 Temperature effect on dimer stability 183
4.4.5 Does dimerization affect other domains? 186
4.5 Discussion 187
4.6 Conclusions 190
4.7 Supporting Information 191
4.7.1 Protein expression, purification, and labeling 191
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.7.2</td>
<td>Experimental setup and procedure for single-molecule fluorescence experiments</td>
<td>192</td>
</tr>
<tr>
<td>4.7.3</td>
<td>Construction of transfer efficiency histograms</td>
<td>193</td>
</tr>
<tr>
<td>4.7.4</td>
<td>Fit of transfer efficiency distributions</td>
<td>194</td>
</tr>
<tr>
<td>4.7.5</td>
<td>Root-mean-square interdye distance</td>
<td>195</td>
</tr>
<tr>
<td>4.7.6</td>
<td>Binding Denturant</td>
<td>196</td>
</tr>
<tr>
<td>4.7.7</td>
<td>Fluorescence lifetime</td>
<td>196</td>
</tr>
<tr>
<td>4.7.8</td>
<td>Modeling Dimerization</td>
<td>197</td>
</tr>
<tr>
<td>4.8</td>
<td>Supporting Figures</td>
<td>199</td>
</tr>
<tr>
<td>4.9</td>
<td>Supporting Tables</td>
<td>204</td>
</tr>
<tr>
<td>4.10</td>
<td>References</td>
<td>209</td>
</tr>
</tbody>
</table>

Chapter 5: The disordered N-terminal tail of SARS CoV-2 Nucleocapsid protein forms a dynamic complex with RNA

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Abstract</td>
<td>214</td>
</tr>
<tr>
<td>5.2</td>
<td>Introduction</td>
<td>215</td>
</tr>
<tr>
<td>5.3</td>
<td>Materials and Methods</td>
<td>217</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Protein expression and purification</td>
<td>217</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Protein Labeling</td>
<td>219</td>
</tr>
<tr>
<td>5.3.3</td>
<td>RNA Preparation</td>
<td>220</td>
</tr>
<tr>
<td>5.3.4</td>
<td>Instrumentation</td>
<td>220</td>
</tr>
<tr>
<td>5.3.5</td>
<td>Analysis of Binding Experiments</td>
<td>221</td>
</tr>
<tr>
<td>5.3.6</td>
<td>Statistical Analysis</td>
<td>222</td>
</tr>
<tr>
<td>5.3.7</td>
<td>Data Availability, Software, Algorithms</td>
<td>222</td>
</tr>
<tr>
<td>5.3.8</td>
<td>Simulations</td>
<td>223</td>
</tr>
</tbody>
</table>
5.3.9 Database Refencing

5.4 Results

5.4.1 Folding Stability of RBD

5.4.2 Binding of non-specific RNA to RBD

5.4.3 Binding of non-specific RNA to NTD-RBD

5.4.4 Simulations of RNA binding to NTD-RBD

5.4.5 Effect of Salt

5.4.6 Interaction with specific single-stranded RNA

5.4.7 Interaction with specific RNA hairpins

5.4.8 Omicron Variant

5.5 Discussion

5.5.1 The NTD is essential for RBD function

5.5.2 The NTD-RBD forms a dynamic complex with RNA

5.5.3 The NTD-RBD region prefers single-stranded RNA

5.5.4 NTD mutations alter RNA binding

5.6 Conclusions

5.7 Supporting Information

5.7.1 Experimental setup and procedure for single-molecule fluorescence experiments

5.7.2 Construction of transfer efficiency histograms

5.7.3 Fit of transfer efficiency distributions

5.7.4 Determination of root mean square interdye distances from mean FRET transfer efficiencies

5.7.5 Folding equilibrium of the RBD

5.7.6 Equilibrium binding models
5.7.7 Nanosecond FCS analysis 262
5.7.8 Coarse-grained simulations 263
5.7.9 Measuring the stability of double-stranded RNA 265
5.8 Supporting Figures and Tables 267

Chapter 6: Macromolecular Crowding and Intrinsically Disordered Proteins: A Polymer Physics Perspective

6.1 Abstract 294
6.2 Introduction 295
6.3 Conformations and dynamics of IDRs in dilute solutions 297
6.4 Crowding Effects on Disordered proteins 303
   6.4.1 Investigating crowding effects using synthetic polymers 303
   6.4.2 Crowding effects on IDR conformations 305
   6.4.3 Crowding effects on binding equilibrium 310
   6.4.4 Crowding effects on diffusion 315
   6.4.5 Crowding effects on binding kinetics 318
6.5 Summary and Outlook 320
6.7 References 323

Chapter 7: Excluded volume and weak interactions in crowded solutions modulate conformations and RNA binding of an intrinsically disordered tail.

7.1 Abstract 338
7.2 Introduction 340
7.3 Results 342
   7.3.1 Large crowders compact protein conformations 343
   7.3.2 Small crowding agents expand protein conformations 345
   7.3.3 The disordered tail remains dynamic under crowded conditions 347
   7.3.4 Balancing excluded volume screening and local interactions 348
7.3.5 Crowder effects on binding affinity equilibrium 352
7.3.6 Depletion effects can describe binding affinities 353
7.3.7 Temperature effects on protein conformations in absence of crowders 354
7.3.8 Enthalpic and entropic contributions to RNA binding in the absence of crowders 355
7.3.9 Temperature dependence of protein conformations in crowded conditions 358
7.3.10 Enthalpic and entropic components to RNA binding in presence of crowders 359
7.3.11 Sequence specificity 360
7.4 Discussion 362
7.4.1 Crowding and IDR conformations 362
7.4.2 Crowding and IDR interactions 363
7.4.3 PEG interactions 364
7.4.4 Implications for SARS-CoV-2 365
7.5 Conclusions 366
7.6 Methods 367
7.7 Supporting Information 368
7.7.1 Reagents 368
7.7.2 Single-molecule fluorescence setup 368
7.7.3 Measurement conditions 369
7.7.4 Construction of transfer efficiency histograms 370
7.7.5 Fit of transfer efficiency distributions 371
7.7.6 Estimate of fraction bound 371
7.7.7 Nanosecond FCS analysis

7.7.8 Determination of root mean square interdye distances from mean transfer efficiencies

7.7.9 Protein conformations

7.7.10 A note regarding the renormalized group theory model

7.7.11 Flory’s criterion

7.7.12 Protein and nucleic acid binding

7.7.13 Depletion theory

7.8 Supporting Figures

7.9 Supporting Tables

7.10 References

Chapter 8: Conclusions and Future Directions

8.1 Main Findings

8.2 SARS-CoV-2 Nucleocapsid interactions with nucleic acids

8.3 SARS-CoV-2 Nucleocapsid with other proteins

8.4 SARS-CoV-2 Nucleocapsid phase separation

8.5 Future directions

8.6 References

Appendix A

The LINK region significantly enhances the binding affinity of N protein to RNA

Appendix B

Nuclie acid binding induces oligomerization of the SARS-CoV-2 Nucleocapsid Protein

Appendix C

Apolipoprotein E4 has extensive conformational heterogeneity in lipid-free and lipid-bound form
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>The SARS-CoV-2 genome</td>
<td>5</td>
</tr>
<tr>
<td>1.2</td>
<td>Transmission EM of a single virus particle of SARS-CoV-2</td>
<td>6</td>
</tr>
<tr>
<td>1.3</td>
<td>The SARS-CoV-2 Viral Life Cycle</td>
<td>7</td>
</tr>
<tr>
<td>1.4</td>
<td>The SARS-CoV-1 Viral Nucleocapsid</td>
<td>10</td>
</tr>
<tr>
<td>2.1</td>
<td>Spectrum of conformational disorder</td>
<td>26</td>
</tr>
<tr>
<td>2.2</td>
<td>Single-molecule fluorescence spectroscopy</td>
<td>28</td>
</tr>
<tr>
<td>2.3</td>
<td>Experimental methods for the study of disordered proteins</td>
<td>33</td>
</tr>
<tr>
<td>2.4</td>
<td>Example of single-molecule fluorescence measurements</td>
<td>37</td>
</tr>
<tr>
<td>3.1</td>
<td>Sequence and structural summary of N protein</td>
<td>62</td>
</tr>
<tr>
<td>3.2</td>
<td>The N-terminal domain (NTD FL) is disordered with residual helical motifs</td>
<td>67</td>
</tr>
<tr>
<td>3.3</td>
<td>The RNA binding domain (RBD) and dimerization domains are interconnected by a flexible disordered linker</td>
<td>71</td>
</tr>
<tr>
<td>3.4</td>
<td>The C-terminal domain (CTD) is disordered, engages in transient interaction with the dimerization domain, and contains a putative helical binding motif</td>
<td>75</td>
</tr>
<tr>
<td>3.5</td>
<td>Nucleocapsid protein undergoes phase separation with RNA</td>
<td>78</td>
</tr>
<tr>
<td>3.6</td>
<td>A simple polymer suggests symmetry breaking can promote single-polymer condensates over multi-polymer assemblie</td>
<td>82</td>
</tr>
<tr>
<td>3.7</td>
<td>Summary and proposed model</td>
<td>93</td>
</tr>
<tr>
<td>S3.1</td>
<td>Sequence alignment of the coronavirus N-terminal domain (NTD)</td>
<td>140</td>
</tr>
<tr>
<td>S3.2</td>
<td>Sequence alignment of the coronavirus RNA binding domain (RBD)</td>
<td>141</td>
</tr>
<tr>
<td>S3.3</td>
<td>Sequence alignment of the coronavirus linker (LINK)</td>
<td>141</td>
</tr>
<tr>
<td>S3.4</td>
<td>Sequence alignment of the coronavirus dimerization domain</td>
<td>142</td>
</tr>
<tr>
<td>S3.5</td>
<td>Sequence alignment of the coronavirus C-terminal domain (CTD)</td>
<td>142</td>
</tr>
</tbody>
</table>
Figure S3.6  Histograms of transfer efficiency distributions across GdmCl concentrations  143
Figure S3.7  Dependence of fluorescence lifetime on transfer efficiency  144
Figure S3.8  Mean transfer efficiency and width of NTD FL vs NTD-RBD, LINK FL vs LINK-ΔDimer, CTD-FL vs CTD fragment across GdmCl concentration  145
Figure S3.9  Fit of NTD construct with two populations compared to folding of RBD domain  146
Figure S3.10  Effects of Urea denaturation on NTD FL, LINK-ΔDimer, and CTD FL  147
Figure S3.11  Interdye distances of NTD, LINK, CTD in presence of salt  148
Figure S3.12  Chain dynamics measured via ns-FCS  149
Figure S3.13  Turbidity experiments plotted against RNA/protein ratio  150
Figure S3.14  Testing SARS-CoV-2 N protein oligomerization  151
Figure S3.15  Distributions of inter-residue distance from ABSINTH simulations (black) vs. excluded volume simulations  152
Figure S3.16  Scaling maps for IDR-only simulations  153
Figure S3.17  Distributions for the radius of gyration ($R_g$) of for IDR-only simulations  154
Figure S3.18  Monte Carlo simulations reveal slow pseudo-kinetics of condensate fusion  155
Figure S3.19  Comparison of secondary structure in IDRs from bioinformatics predictions  156
Figure 4.1  Domain architecture of SARS-CoV-2 N protein and structural features of the dimerization domain  176
Figure 4.2  Dimer domain conformations  178
Figure 4.3  Stabilization of the dimer conformations upon binding: conformational changes and dimerization affinity  180
Figure 4.4  Denaturant dependace of the dimerization domain  183
Figure 4.5  Temperature dependance of the dimerization domain  186
<table>
<thead>
<tr>
<th>Figure ID</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>S4.1</td>
<td>Estimated transfer efficiency based on the dimer structure</td>
</tr>
<tr>
<td>S4.2</td>
<td>Labeling Stoichiometry Ratio Histograms</td>
</tr>
<tr>
<td>S4.3</td>
<td>Lifetime vs Transfer efficiency of the DD_{FL}</td>
</tr>
<tr>
<td>S4.4</td>
<td>Effects of solution conditions on dimerization</td>
</tr>
<tr>
<td>S4.5</td>
<td>Dimerization alters the conformations of the Linker domain and CTD, but not NTD or RBD</td>
</tr>
<tr>
<td>5.1</td>
<td>Nucleocapsid protein constructs in this study</td>
</tr>
<tr>
<td>5.2</td>
<td>RNA Binding Domain (RBD) folding</td>
</tr>
<tr>
<td>5.3</td>
<td>poly(rU) binding to RBD and NTD-RBD</td>
</tr>
<tr>
<td>5.4</td>
<td>Length dependence of poly(rU) binding to NTD-RBD and RBD</td>
</tr>
<tr>
<td>5.5</td>
<td>Coarse-grained simulations of the Nucleocapsid protein with ssRNA</td>
</tr>
<tr>
<td>5.6</td>
<td>Salt dependence of binding association constant</td>
</tr>
<tr>
<td>5.7</td>
<td>Specific ssRNA binding to NTD_{L}-RBD</td>
</tr>
<tr>
<td>5.8</td>
<td>Specific hairpin RNA (hpRNA) binding to NTD-RBD</td>
</tr>
<tr>
<td>5.9</td>
<td>Omicron variant</td>
</tr>
<tr>
<td>S5.1</td>
<td>Transfer efficiency variation upon binding</td>
</tr>
<tr>
<td>S5.2</td>
<td>Dynamics of the disordered NTD when complex with RNA</td>
</tr>
<tr>
<td>S5.3</td>
<td>NTD-RBD:(rU)_{10} dependence of interacting residues on distance threshold used for contact fraction</td>
</tr>
<tr>
<td>S5.4</td>
<td>The NTD and RNA remain disordered in the NTD-RBD:(rU)_{10} complex</td>
</tr>
<tr>
<td>S5.5</td>
<td>Simulations of N protein construct and RNA binding</td>
</tr>
<tr>
<td>S5.6</td>
<td>Representative transfer efficiency distributions of (rU)_{20} as a function of salt concentration</td>
</tr>
<tr>
<td>S5.7</td>
<td>Representative transfer efficiency distributions of (rU)_{40} as a function of salt concentration</td>
</tr>
</tbody>
</table>
Figure S5.8  Association constant as a function of the total concentration of positive ions 273
Figure S5.9  Transfer efficiency distributions for NTD$_L$-RBD and RNA hairpins 274
Figure S5.10 Thermal melting curves of RNA hairpins 275
Figure S5.11 The NTD does not alter the overall pattern of RBD:RNA interactions 276
Figure S5.12 RNA length tunes the magnitude of protein:RNA interactions but does not alter the overall pattern of RBD:RNA interaction 277
Figure 6.1  Volume Fraction 295
Figure 6.2  Conformations of disordered proteins in dilute solutions 302
Figure 6.3  Dilute and semidilute regime of polymers 305
Figure 6.4  Crowding models and conformational changes of disordered proteins 310
Figure 6.5  Crowding effects on binding 315
Figure 6.6  Quantification of crowding effects on disordered proteins in vitro and in living cells 320
Figure 7.1  N-terminal domain of SARS-CoV-2 and RNA-Binding-Domain 342
Figure 7.2  Impact of crowders on IDR conformations 344
Figure 7.3  Degree of compaction as a function of crowding concentration 347
Figure 7.4  Crowding impact on the NTD-RBD and (rU)$_{40}$ association constant 351
Figure 7.5  NTD-RBD conformations and binding of (rU)$_{40}$ as a function of temperature 357
Figure 7.6  Effects of crowding on the Omicron variant 361
Figure 7.7  Weak interactions and Excluded Volume (EV) screening effects on disordered tails 366
Figure S7.1  Mean transfer efficiency at 15% w/v as a function of PEG Molecular Weight 384
Figure S7.2  NTD$_L$-RBD dynamics under crowding conditions 385
Figure S7.3  Dilute and semi-dilute conditions 386
Figure S7.4  Temperature dependence of RBD₁  387
Figure S7.5  NTD-RBD conformations and binding of (rU)₄₀ as a function of temperature  388
Figure S7.6  NTD-RBD conformations and binding of (rU)₄₀ as a function of temperature in the presence of 15% w/v EG  389
Figure S7.7  NTD-RBD conformations and binding of (rU)₄₀ as a function of temperature in the presence of 15% w/v PEG 600  390
Figure S7.8  NTD-RBD conformations and binding of (rU)₄₀ as a function of temperature in the presence of 15% w/v PEG 8000  391
Figure 7.9  NTD₁-RBD conformations as a function of temperature in absence and presence of 15% w/v PEG solutions  392
Figure 7.10  Figure 10. ΔΔH and -ΔΔS at 15% w/v for EG (red), PEG 600 (yellow), PEG 8000 (gray)  393
Figure 7.11  Temperature dependence of ln(Kₐ) as estimated from different fitting models  394
Figure 7.12  Binding curves from single-molecule FRET experiments for wild type Wuhan-Hu-1 and Omicron variants in absence (empty and cyan dots) and presence (orange and blue dots) of 15% w/v PEG 600  395
Figure 8.1  N protein- nucleic acid interactions  414
Figure 8.2  N protein interactions  416
Figure 8.3  N protein phase separation disruption by small molecules  420
Figure A1  Binding of poly(rU) to the NTD-RBD-LINK  431
Figure A2  Binding of (rU)₂₀ to the NTD-RBD-LINK  432
Figure B1  N protein is monomeric at 100 pM concentrations in the absence of RNA  437
Figure B2  N protein oligomerizes at 100 pM concentrations in the presence of RNA  438
List of Tables

Table S3.1  N protein Constructs  134
Table S3.2  Fit parameters to denaturant binding models  137
Table S3.3  Fit Parameters Higgs and Joany  138
Table S3.4  Table S3.4. Scaling exponents  138
Table S3.5  All-atom simulation summary  138
Table S3.6  Interdye distances  139
Table S3.7  Sequences used in simulations  139
Table S4.1  Sequence of Wildtype (WT) Nucleocapsid Protein  204
Table S4.2  N protein constructs used in this study  204
Table S4.3  FRETTraj parameters for Alexa 488 and Alexa 594  207
Table S4.4  RNAs used in this study  207
Table S4.5  Thermodynamic parameters fit from temperature dependence with 23 °C as reference temperature  208
Table S5.1  Sequence of wild type NTD-RBD  278
Table S5.2  Constructs used in this study  278
Table S5.3  RNA sequences used in this study  280
Table S5.4  Summary of simulation details  281
Table S5.5  RBD Folding parameters  283
Table S5.6  Intrinsic association constants  283
Table S5.7  RBDₜ and NTDₜ-RBD association constants for (rU)ₙ as measured by single-molecule FRET  283
Table S5.8  Simulation-derived association constants (Kₐ) in μM⁻¹  284
Table S5.9  Simulation-derived dissociation constants (K₅₀) in μM  284
Table S5.10  Simulation-derived ratio of association constants Kₐ* defined as (Kₐ of Construct + (rU)ₙ)/(Kₐ of NTD-RBD + (rU)₂₅) 28
Table S5.11 Ion released upon binding of (rU)$_{20}$ and (rU)$_{40}$  
Table S5.12 Association and dissociation constants of NTD$_L$-RBD as a function of salt concentration for (rU)$_{20}$ and (rU)$_{40}$  
Table S5.13 NTD$_L$-RBD association constants for V21 binding  
Table S5.14 NTD$_L$-RBD association constants for hairpin RNA sequences  
Table S5.15 Simulation-derived ratio of association constants $K_A^*$ defined as $(K_A$ of Construct + (rU)$_n$)/(K$_A$ of NTD-RBD + (rU)$_{25}$)  
Table S7.1 Sequence of WT NTD-RBD  
Table S7.2 Constructs used in this study  
Table S7.3 PEGs used in this study  
Table S7.4 Dynamics of the NTD$_L$-RBD in the presence of 15% Crowder  
Table S7.5 Schellman weak binding model individual fit results  
Table S7.6 Schellman weak binding model global fit results  
Table S7.7 Association constants for the binding of NTD$_L$-RBD to (rU)$_{40}$ in presence and absence of crowding agents at 41°C in 50 mM HEPES buffer  
Table S7.8 Association constants for the binding of NTD$_L$-RBD to (rU)$_{40}$ in presence and absence of 15% w/v crowding agents at temperatures from 10 °C to 37 °C in 50 mM HEPES buffer  
Table S7.9 Enthalpic and entropic contribution to the binding free energy  
Table S7.10 Theoretical expectations from depletion theory ($\Delta S_{\text{depl}}$)  
Table S7.11 Association constants for the binding of NTD$_L$-RBD to (rU)$_{40}$ calculated using the change in mean transfer efficiencies in absence of crowding agents from 10 °C to 56 °C in 50 mM HEPES buffer  
Table A1 Sequence of WT NTD-RBD-LINK  
Table A2 Association constants for the NTD-RBD-LINK in comparison to the NTD-RBD (chapter four)
I want to thank every single person, mentioned here or not, who has been there for me and supported me in one way or another these past six years. I am grateful for every experience that’s brought me here today. I never would have accomplished anything had I been alone.

To my parents, thank you for the love and support you have given me my entire life, and for all the sacrifices you’ve made to get me here today. From my childhood to present day, you have always pushed me to the best person I could be and always encouraged all of my curiosities. Thank you for showing me that above all, kindness is the most important quality and that I am always loved, successful or not. Daddy, you’re the strongest man I know, and I am so grateful that you continue to fight to stay here every day. I am grateful that even with stage 4 cancer, you’re living your life and showing up for me, day after day.

To my family, thank you for the love and support as I grew into who I am today.

To my thesis advisor, Dr. Andrea Soranno, thank you for taking me into your lab and allowing me to be your first graduate student. Your guidance and mentorship have been truly invaluable, and you’ve pushed me to be the best scientist I could be. Thank you for being patient and passionate about teaching, and for fostering a safe lab environment. Most importantly, thank you for not giving up on me when I felt incapable of succeeding and for pushing and reassuring me that I was capable.

To Dr. Melissa Brereton, thank you for the CONSTANT support, friendship, and mentorship throughout the entirety of my PhD. I want to thank you for being a sounding board, a shoulder to
cry on, and a genuine friend for the past five years in the lab. I am, and always will be grateful to have you in my life.

To the Soranno lab, current and former members: Thank you for all of the help, guidance, support and friendship. Jere, Debjit, Klaudio, Nirmay, Upasana, Owen, Jhullian, Eun Sil, Marco and Pasquale, I am so lucky to have worked with you all in one way or another. You are all the best group of people, and I can’t express how lucky I am to have fallen into such a supportive, loving, and friendly environment. You guys are amazing.

To my thesis committee, Dr. Tim Lohman, Dr. Kathleen Hall, Dr. Janice Robertson, and Dr. Yuna Ayala, thank you for the guidance, support, and advice over the years. Your comments, suggestions, and encouragement have been essential to the completion of my PhD.

To the BMB admin, thank you for helping my graduate career run smoothly. Your kindness and patience was always appreciated. I am also grateful for the funding from the BMB department and Millipore Sigma.

To my previous mentors, Dr. Cohn and Dr. Rajaraman, your guidance and mentorship throughout my time at Stockton have been life-changing. Working in your labs gave me the opportunity to perform independent research and realize that I wanted to go to grad school. I would not be here had you not encouraged me to apply. To Mrs. Lotterman, Mrs. D’Angelo, and Mr. Maiorino, thank you for showing high school me that it didn’t matter where I grew up, that I was capable, and that I could be anything I wanted. You were all incredible mentors and GHS is lucky to have you. To the Boys and Girls club of Garfield, and everyone I’ve worked with there, thank you for taking me in at fifteen and giving me the opportunity to volunteer and work with you all. The
experience I’ve gained and the friendships I’ve made there will stay with me forever. Nicole, Shane and Sugar, thank you for being wonderful mentors and friends over the years.

To my therapist, Susan, you’ve helped me grow so much over the past five years. Thank you for never giving up on me or minimizing my feelings. Thank you for making sure I stayed accountable on progressing to recovery. Thank you for assuring me that it was okay not to be and to not let my illness define who I am. Thank you for significantly improving my quality of life and helping me care just a little bit less of what others think of me. I appreciate you more than words can describe.

To George and Todd, thank you for being the world’s best uncles. I am so lucky to have lived with you when I first moved to STL and to have your love all of these years. I am forever grateful for everything you have done for my friends and I.

To the best friends I’ve ever known, thank you for everything you’ve done and the friendship you’ve given me. James, thank you for keeping this friendship alive for the past 17 years, you are one of my dearest friends, and I am so grateful for your love and support. You’re one of the most genuine people I know, and I’m so lucky to have you in my life. Brianna, Kevin, Amina, Amel and Brian, thank you for the genuine love and kindness over the years since high school and thank you for being here for me no matter what, no matter when. My best memories of high school are because of you guys, and I love that you’re still a huge part of my life and who I am today. Dhvani and Danielle, thank you for being the crazy loving, kind, amazing people you are and for being incredible humans. You’ve made my time in college unforgettable. To my cohort here at WashU, Victoria, Paige, Matt, Jhullian, Rose, Mickey, Huiming, Naomi, Kellen, I am so lucky to have gone through grad school with you guys. You’re the best group of friends to do this
with and the constant love does not go unnoticed. From the QE practices to the potlucks and ridiculous group chat, I would have never completed my PhD without you all. I love you all so much.

To Victoria, thank you for love and friendship and endless rant sessions. Thank you for the game nights, and heart-to-hearts, and for helping push me to write this thesis.

To my sister, Gabby, I would not be who I am or where I am without you. You keep me sane and love me at my worst. You have stuck by me even when I have been impossible to deal with and even throughout all of the times I’ve pushed you away. You have literally kept me alive, I’m so lucky to have you in my life. I can’t wait to grow old and have families and annoy you forever. I love you, sunshine!

To my husband, Brandon, thank you for loving me and supporting me all of the times when I couldn’t love myself. Thank you for the countless nights you’ve stayed up helping me study, write, or facetime me while I was in the lab late or on weekends. Your endless support and patience mean the world to me. You are my best friend and marrying you is one of the best things I’ve ever done.

To Murdy, you’re the bestest, most beautiful boy on the earth, I love you.

Jasmine Cubuk

Washington University in St. Louis

May 2023
Dedicated to my parents.
ABSTRACT OF THE DISSERTATION

Single-molecule spectroscopy of SARS-CoV-2 Nucleocapsid Protein

by

Jasmine Cubuk

Doctor of Philosophy in Biology and Biomedical Sciences

Biochemistry, Biophysics and Structural Biology

Washington University in St. Louis, 2023

Professor Andrea Soranno, Chair

The COVID pandemic has affected over 760,000,000 individuals worldwide since late 2019. Understanding how SARS-CoV-2, the virus responsible for the disease, functions at a mechanistic level is essential to develop therapeutics and vaccines. SARS-CoV-2 utilizes four structural proteins that work together during the viral life cycle to ensure the spread of infections: spike (S), envelope (E), membrane (M) and nucleocapsid (N). Though much work has focused on the S protein for the purpose of vaccines, the N protein plays a key function in the viral life cycle as well. Nucleocapsid is responsible for packaging the viral genome and incorporating the viral ribonucleocapsid into the virion. In SARS-CoV-2, the packaged viral genome adopts a more “beads on a string” organization than the helical configuration previously observed in other coronaviruses. Little is known about how N protein controls the packaging of the viral genome. N protein is composed of five domains: a folded RNA binding domain, a folded dimerization domain, and three flanking intrinsically disordered regions that were proposed to modulate interaction with RNA. Despite their potential role in modulating genome compaction, properties of corresponding disordered regions in nucleocapsid proteins from other coronaviruses remains largely understudied. At the beginning of the COVID pandemic, there was no insight on whether predicted disordered regions in SARS-CoV-2 remain disordered in the context of the full-length protein and how they modulated protein-RNA interactions.

In my thesis work, I made use of single-molecule confocal fluorescence spectroscopy, and in particular, single-molecule Förster Resonance Energy Transfer (FRET) to close this knowledge
gap and investigate conformations, dynamics, and interactions of the disordered regions within the SARS-CoV-2 nucleocapsid protein.

I first determined that the three predicted disordered regions of N protein are disordered in the context of full-length protein. The combination of single-molecule FRET experiments and all-atom Monte Carlo simulations revealed that the monomeric full-length protein is flexible and dynamic. In addition, we observed that the protein undergoes phase separation when mixed with RNA.

Having characterized the monomeric form of the protein, I next investigated the protein-protein interactions that lead to dimerization. I found that the dimerization domain is partially disordered and flexible when N protein is monomeric. I further determined the concentration under which dimerization occurs (K_D = 11 ± 3 nM) to be in good agreement with previous AUC experiments and found that even in the dimeric state, N protein retains some of the dynamic nature of the monomer. I also quantified that dimer formation does not alter the conformations of the disordered NTD and folded RBD, but causes an expansion of the disordered linker and CTD. These observations were consistent with my previous determination of interactions of the linker and CTD with the dimerization domain.

As a next step, I started to investigate the interactions of the N protein with RNA. To understand the role of a disordered region in aiding the recruitment of RNA, I started to investigate whether the NTD enhances the affinity of RNA to the RBD. For this, I focused on using truncations of the NTD-RBD and RBD in isolation. My experiments showed that the presence of the NTD enhances the affinity by over 50-fold compared to the RBD in isolation. Furthermore, when in complex with RNA, the NTD forms a dynamic fuzzy complex, as seen also in coarse-grained simulations. Comparison of single- and double-stranded RNA provided evidence that the NTD-RBD preferentially binds to single-strandedRNA.

Finally, I examined how a crowded environment (mimicked by polyethylene glycol molecules) can modify binding properties of the NTD-RBD to RNA and found that the NTD binding is sensitive to the solution environment. Comparison of Omicron and wildtype (Wuhan-Hu-1) variants revealed that significant differences in binding affinity observed in absence of crowding are equalized in presence of the crowders.

In conclusion, single-molecule fluorescence spectroscopy has offered a powerful toolbox for investigating protein conformations and interactions of disordered regions. The work has
provided new insights on the molecular interactions encoded in the SARS-CoV-2 N protein and paves the way to quantitative studies of interactions with other binding partners, viral genome RNA, and small molecules.
Chapter 1

The SARS-CoV-2 Nucleocapsid protein
1.1 The SARS-CoV-2 pandemic.

The first case of severe acute respiratory syndrome (SARS) was reported in February 2003, where the illness spread to more than two dozen countries across North America, South America, Europe, and Asia before being contained by 2004\textsuperscript{1,2}. During the outbreak, over 8000 individuals worldwide became infected and 774 died as a result of the infection according to the World Health organization (WHO)\textsuperscript{3}. In 2019, a novel SARS coronavirus, SARS-CoV-2, emerged. Unlike SARS-CoV-1, there was detrimental spread, leading to a worldwide pandemic with over 760,000,000 reported cases of SARS-CoV-2, resulting in over 6,800,000 deaths. Even three and a half years later in 2023, we still see the devastating effect it continues to have. SARS-CoV-2 causes the disease COVID-19, and the severity of the disease remains variable. Though the main effects of COVID-19 target the respiratory system, some people remain asymptomatic, and others have a fatal response, as the virus also attacks other organs\textsuperscript{4-6}.

In the following, I will contextualize the problem of SARS-CoV-2 from a biochemical point of view and highlight the questions my thesis work will address.

1.2 Viruses.

Viruses are infectious microbes consisting of one or multiple segments of nucleic acids (either deoxyribonucleic acid (DNA) or ribonucleic acid (RNA)) encapsulated by a coat of proteins\textsuperscript{7}. Viruses need to infect a host cell in order to replicate and, in the process, often kill the host cell causing host organism damage\textsuperscript{8-10}. Viruses are classified into a taxonomic system, similarly to other organisms. One way of classifying viruses is to use the Baltimore classification, which places viruses in one of seven groups depending on the properties of their genome (e.g. DNA vs
RNA, single vs double strand, positive vs negative sense) and their method of replication. The seven classes are as follows: i) double-stranded DNA viruses, ii) single-stranded DNA viruses, iii) double-stranded RNA viruses, iv) positive sense single-stranded RNA viruses, v) negative-sense single-stranded RNA viruses, vi) single-stranded RNA replication/transcription viruses (i.e. retroviruses), and vii) double-stranded DNA replication/transcription viruses\textsuperscript{11,12}.

### 1.3 The family Coronaviridae

Coronaviruses (CoVs), belonging to the family \textit{Coronaviridae}, are enveloped, single-stranded positive-sense RNA (+ssRNA) viruses that package some of the largest, infectious, viral genomes known to date (26-32 kilobases)\textsuperscript{13}.

Coronaviruses are further characterized into two subfamilies, \textit{Letovirinae} and \textit{Orthocoronavirinae}. The \textit{Orthocoronavirinae} subfamily contains four genera: Alphacoronavirus (α-CoVs), Betacoronavirus (β-CoVs), Gammacoronavirus (γ-CoVs), and Deltacoronavirus (δ-CoVs)\textsuperscript{14}. The main difference in the genera is the organism where they originated and the ones they mostly infect. The ancestral host for most alpha- and betacoronavirus is bats, whereas for gamma- and deltacoronaviruses is birds. Alphacoronaviruses typically infect bats, though some human infections have been reported, and cause both gastrointestinal and respiratory issues\textsuperscript{15}. Betacoronaviruses typically infect humans and cause respiratory issues\textsuperscript{16}. Gammacoronaviruses mainly target birds and cause respiratory issues\textsuperscript{17,18}. Deltacoronaviruses have been shown to infect both birds and pigs, causing gastrointestinal issues\textsuperscript{18}.

SARS-CoV-2 was identified as a new member of the β-CoVss: other members of this family
include Middle East Respiratory Syndrome (MERS), mouse hepatitis virus (MHV) and SARS-CoV-1\textsuperscript{19,20}.

Like other members of coronaviruses, betacoronaviruses have large genomes (typically around 30 kb in size)\textsuperscript{21} Generally, coronavirus genomes are capped at the 5’-end and polyadenylated at the 3’ terminus. The 5’-end of the genome contains untranslated regions (UTR), known as the leader RNA. Though sequence rearrangement may occur due to heterologous recombination, the 5’ UTR is present at the 5’-ends of all the subgenomic RNA as well\textsuperscript{22}. The 3’-end also contains an UTR, usually larger in length than the 5’-end, which is followed by a poly(A) tail. Both UTRs are necessary for regulating RNA replication and transcription\textsuperscript{22,23}. The viral particles containing the genome are roughly spherical, yet relatively pleomorphic, ranging from \textasciitilde80 to \textasciitilde120 nm in size\textsuperscript{22,24}.

Though previous β-CoVs have a similar structural organization and have been extensively investigated\textsuperscript{25}, little is understood of the mechanisms controlling viral assembly. Moreover, in the case of the SARS-CoV-2, a molecular description of its components and their interactions is still incomplete up to today.

1.4 The SARS-CoV-2 Virus

The SARS-CoV-2 viral genome is one of the largest single-stranded genomes, spanning about 30 kilobases\textsuperscript{13} (Fig. 1.1). The first \textasciitilde20 kb encode for the largest gene, ORF1ab. ORF1ab contains overlapping open reading frames, which encode two polyproteins that are cleaved to yield sixteen non-structural proteins (NSP1-16)\textsuperscript{26}, including the papain-like proteinase protein (NSP3)\textsuperscript{27}, 3C-like proteinase (NSP5)\textsuperscript{28}, RNA-dependent RNA polymerase (NSP12/RdRp)\textsuperscript{29},
The non-structural proteins are responsible for suppression of host immune system and host gene expression, proteolytic processing, and viral transcription and replication. The remaining ~10 kilobases encode for the structural and accessory proteins. The main four structural proteins include the spike glycoprotein (S), envelope protein (E), membrane protein (M) and the nucleocapsid protein (N). Much of the initial research responding to the pandemic focused on the S protein and its interaction with human angiotensin-converting enzyme 2 (ACE2), which is necessary to mediate entry of the virus in the host cell. The structure of the S protein has been solved via cryo-electron microscopy, and in the context of the
**Fig. 1.2.** Transmission EM of a single virus particle of SARS-CoV-2. In pink is the spike protein, yellow virus-enveloping membrane and blue the nucleocapsid protein bound to RNA. Coloring done manually.

virus, it is a trimer. However, other structural proteins such as the N protein also play a critical role in viral assembly and replication. The viral ribonucleocapsid (RNP) is formed when the N protein compacts and packages the viral genome, and its interaction with the M protein are what allows the virus RNP to be incorporated into the virion. The overall structure of the virus can be seen in **Fig. 1.2** where the S protein (pink) surrounds and protrudes through the viral membrane (yellow). Inside, the N protein bound to genomic RNA (blue) is arranged in a “beads on a string” manner, in contrast to other CoVs proposed helical arrangement. It is useful to consider these protein components in the context of their role during viral infection and replication.

The life cycle of the SARS-CoV-2 can be broken down into five steps: i) attachment to the host cell, ii) viral entry and uncoating, iii) formation of the replication-transcription complex (RTC), iv) synthesis of viral RNA, v) and molecular assembly/release of SARS-CoV-2 (**Fig 1.3**). **Attachment.** Attachment to the cell surface is mediated by the spike glycoprotein. The spike protein system consists of three monomeric S polypeptides, with each polypeptide containing an
S1 and S2 subunit. The S1 subunit, which contains the receptor binding domain, binds to the host cell receptor, ACE2\(^{40}\), where it is then cleaved at the S1/S2 site. The S2 subunit is responsible for membrane fusion, which takes the virus into the next stage of its cycle\(^{41}\).

**Fig. 1.3: SARS-CoV-2 Viral Life Cycle.** SARS-CoV-2 S protein (blue) interacts with the ACE2 receptor (green), and S protein is cleaved by TMPRSS2 (purple). Uncoating allows the +ssRNA genome to be released into the cytoplasm, where it is translated into the two polyproteins, pp1a and pp1ab. Following translation, the polyproteins are auto proteolytically processed into the NSPs, which assemble the replicase-transcriptase complex (RTC). During formation of the RTC, remodeling of membranes occurs, allowing for the formation of double-membrane vesicles (DMVs) derived from the ER, used for viral RNA synthesis. Following translation, newly synthesized S,M and E proteins anchor to the ER membrane before interaction with N protein shuttles the proteins to the virion assembly site (ERGIC). The new viral nucleocapsid (vRNP) complexes migrate to the ERGIC and bud into the lumen. The newly enveloped virion is finally released from the cells via lysosomes, where the process can begin once again.
Viral entry and uncoating. Entry into the host cell is achieved via endosomal, non-endosomal entry, or both. Endosomal entry is mediated via the cysteine protease cathepsin, which aids in membrane fusion in a low pH environment (early pathway). Non-endosomal entry is facilitated via conformational changes that occur after fusion of S with ACE2, aided by several host factors including type II transmembrane protease serine 2 (TMPRSS2) protease (late pathway). Following entry into the host cell, the genomic RNA is uncoated via lysosomal digestion due to the cells immune response, and then released.

RTC formation. Ribosomal translation of the two open reading frames (ORF1a and ORF1b) occurs immediately after the release and uncoating of the genomic RNA. The resulting polyproteins are co- and post-translationally processed via proteolytic cleavage into the NSPs, thus providing the RTC machinery required for viral replication and transcription.

Viral replication. Along with the expression of the NSPs, the first three steps of the viral life cycle subsequently lead to the biogenesis of viral replication organelles such as double-membrane vesicles (DMVs), convoluted membranes (CMs) and small open double-membrane spherules (DMSs), which create a protective environment for replication and synthesis of multiple copies of genomic RNA. The negative single-stranded genomic RNA (-ssRNA) copies also serve as a template to form a set of nested positive sense subgenomic mRNAs that are required for the production of the viral proteins essential for assembly and release.

Assembly/release. Once the production of subgenomic mRNAs is initiated, the viral RNAs can associate with host cell ribosomes to synthesize the structural and accessory proteins required for the assembly of multiple copies of the virus. Most accessory proteins and three of the four structural proteins (spike, envelope, and membrane) are membrane associated and thereby are
synthesized by endoplasmic reticulum-bound ribosomes. The nucleocapsid protein, however, is translated by free cytosolic ribosomes of the host cells. The translated structural proteins are then transported to the endoplasmic reticulum Golgi intermediate compartment (ERGIC) where virion assembly occurs. During virion assembly, M protein interacts with the spike and envelope proteins to provide a scaffold, while M-N interactions mediate packing of the viral genome. Newly produced genomic RNA results in budding into the lumen of secretory vesicular compartments. Virions are excreted from infected cells via exocytosis after being transported in smooth-wall vesicles using the secretory pathway, where they can now spread to various parts of the body.\(^\text{45}\)

Understanding how each component of the viral life cycle functions is key to identifying new means for disarming the viral replication.

In my work, I focused on characterizing the SARS-CoV-2 nucleocapsid protein and its role in binding and recruiting RNA. When I began working on this problem, the 2020 pandemic had just developed and the only information available was based on previous characterization of the corresponding components from the 2003 SARS-CoV. In the following, I will summarize what was known at the time and how that informed my subsequent steps.

### 1.5 SARS-CoV Nucleocapsid protein

The N protein of SARS-CoV-1 is highly immunogenic and abundantly expressed after human infection\(^\text{46-48}\), making it a strong candidate for serological assays or diagnostic markers.\(^\text{49}\)

The SARS-CoV N protein plays a fundamental role during viral assembly, where it is required for packaging the viral genome into a helical ribonucleocapsid (RNP) through interactions with
the membrane protein, seen in Fig. 1.4. The N protein is composed of five domains: an N-terminal tail, a folded RNA binding domain (RBD), a linker, a dimerization domain, and a C-terminal tail (Fig 1.4). The N-terminal half is identified as the RNA binding domain due to its ability to bind RNA with micromolar affinity. The N protein dimer serves as the basic building block for viral replication, and it has been shown to form a dimer through its C-terminal half. Not only has the C-terminal half been shown to facilitate dimerization, but studies have also revealed the capability of the C-terminal half to bind to nucleic acids. Furthermore, few studies have also shown that the remaining 40% of the protein, though predicted to be disordered, also contributed to nucleic acid binding. Overall, the N protein is consistently depicted and described as an essential RNA binding component of the SARS-CoV life cycle. Of the disordered regions, the central serine-arginine-rich linker has been described as both a hotspot for phosphorylation and an essential component required for interaction with the membrane protein and other binding partners.

The N protein dimer (500 nM) undergoes both chemical and thermal denaturation, with a $C_{m}$ of guanidinium chloride (GdmCl) of about 1.5 M and melting temperature between 35 °C and 55 °C. Fluorescence spectra measuring the stability as a function of pH suggest that the protein is most stable between pH 7-9 at 25 °C.
The N protein also undergoes many post-translational modifications (PTMs) including phosphorylation, acetylation, and sumoylation\(^63\). Phosphorylation of the N protein is demonstrated to occur mainly at the serine residues in the linker region of the protein\(^64,65\) and to be required for cytoplasmic shuttling of the protein. Mass Spectrometry of human sera from infected patients in 2003 showed that the SARS-CoV-1 N protein is acetylated at the N terminus following removal of the N-terminal methionine and oxidation of some remaining methionines\(^66\). Finally, sumoylation occurs at lysine position 62 in the RBD\(^67,68\).

When discussing the function of the nucleocapsid protein, its role in packaging the genome is of utmost importance in terms of viral replication and survival. After assembly is localized at the membrane of the ERGIC, facilitated by interactions of the N, M, and E proteins, N mediates packaging of the viral genome into the assembling virus. However, the interactions that drive genome packaging are not fully understood. Electron cryomicroscopy of the SARS-CoV virus displays the interactions of the N protein with the M protein with poor resolution of the core structure, likely suggesting a dynamic complex\(^48\). Due to the sequence and structural conservation between SARS-CoV and Mouse Hepatitis Virus (MHV), the helical packaged structure observed in MHV is thought to also occur in SARS\(^69\). Other studies have discussed the role of the disordered regions of the N protein as interaction hubs for viral genome binding and packaging\(^58\), helping the genome to wrap around the N protein and form the conventional helical structure.

1.6 The SARS-CoV-2 Nucleocapsid protein.

The SARS-CoV-2 nucleocapsid (N) protein is a 46 kDa multidomain protein with high sequence conservation amongst other CoVs (Chapter 3.10, S3.1-5). Its structural architecture mirrors the
one of the SARS-CoV-1, with two folded domains, the RNA binding domain (RBD) and the dimerization domain, flanked by three predicted disordered regions, N-terminal and C-terminal disordered tails as well as a disordered linker between the two folded domains. After the start of the 2020 pandemic, many studies focused on characterizing the structured domains of the N protein, resolving the structures of both the RBD (6M3M, 6YVO, 6WKP, ) and the Dimerization Domain (6WZO, 6WZQ, 7C22, 7CE0, 6YUN). As expected based on the sequence homology, they exhibit strong structural similarity to previous coronavirus N protein folded domains.

While much of the attention focused on folded domains, the disordered regions were largely neglected, regardless of evidence that the disordered regions are often essential for protein function. Though the sequence of the N proteins from β-CoVs are fairly conserved (Chapter 3.10, S3.1-5), including some of the features of IDRs, they are not identical. In this respect, the sequence composition is particularly important for IDRs, with amino acid content and patterning dictating N protein conformation and, possibly, its interaction with ligands. Indeed, differently from folded regions, conservation of conformational and interacting properties can be achieved by different sequence compositions, while conservation of specific residues in different sequence contexts may result in significantly different protein conformations and interactions.

The disordered nature of the N protein has been postulated as a key mechanism for genome packaging. This is especially of interest, as the typical beta-coronavirus helical RNP has been challenged by current work demonstrating that the SARS-CoV-2 N protein interacts with genomic RNA inside virus particles with a “beads on a string” modality. In this respect, it is interesting to reconsider the conclusions regarding the structure adopted by the viral genome in MHV. In this case, based on EM data, it was reported that “only short coiled
fragments are detectable in the reconstructions, which strongly suggests that the helical nucleocapsid is a very flexible structure that extensively twists and folds upon itself, rapidly adopting orientations that would not be recognizable as coils in tomographic sections.88 Therefore, while locally helical packaging may be occurring, different properties encoded in the viral genome and in the nucleocapsid protein may encode for different modes of packaging, with the MHV having a higher helical content and SARS-CoV-2 having a lower helical content. A better understanding of the conformational properties and interactions of the SARS-CoV-2 N protein and its comparison with previous coronavirus variants may help shedding light on the different structures adopted by the viral genomes.

For the remainder of this work, I will describe my contributions to studying the biochemical and biophysical properties of the SARS-CoV-2 nucleocapsid protein, with a strong focus on the intrinsically disordered regions using single-molecule fluorescence spectroscopy.

1.7. Scope of the thesis.

The Nucleocapsid protein (N) is responsible for compacting and packaging the viral genome. The exact mechanism by which the protein can package a large 30 kb genome into a small ~100 nm virion remains unknown. A fundamental step in decoding the mechanisms of condensation is quantifying the structural properties and interactions of the N protein. In this thesis, I have used single-molecule spectroscopy to study the conformations of the N protein and probe inter- and intra-molecular interactions, within the protein and with ligands. While this first chapter has served as a brief introduction to Coronaviruses and the SARS-CoV-2 components, with a strong emphasis on the nucleocapsid protein, the next chapters will discuss my original contributions to this problem.
In the second chapter, I provide an overview of the properties of disordered regions and discuss the advantages of using single-molecule fluorescence spectroscopy to investigate dynamic conformational ensembles (published in: Cubuk, Stuchell-Brereton, and Soranno, Biochemical Essay 2022).

In the third chapter, I present the conformational analysis of the three predicted disordered regions (NTD, LINK, and CTD) of the Nucleocapsid protein in the context of the full-length protein, which I determined to be highly disordered and dynamic. Furthermore, single-molecule experiments reveal interactions of the CTD with the dimerization domain and dimerization domain-RBD interactions that are modulated by ion concentration. The combination of experiments and all-atom simulations provided insights on atomistic details of the protein conformations, revealing transient helicity regions that may act as local binding interfaces for protein-protein or protein-RNA interactions (published in Cubuk et al., Nat Comm. 2021).

In the fourth chapter, I discuss the conformations of the dimerization domain and its dimerization affinity. My experiments revealed that: i) the dimerization domain adopts more expanded configurations in its monomeric form, folds upon dimerization, and has a strong dimerization affinity ($K_{\text{dim}} = 11 \pm 3$ nM); ii) the dimerization of the protein further modulates the conformations of the disordered regions, with particular respect to the CTD. Furthermore, my experiments investigate the thermal and chemical stability of the dimer. (Cubuk et al., in submission).

In the fifth chapter, I address whether the disordered regions play a role in the protein function, in particular with respect to RNA binding. In particular, I focused on understanding the role of the disordered NTD in aiding the RBD to bind to nucleic acids, primarily single-stranded
RNA (ssRNA). I observed that the NTD-RBD region has a preferential interaction with ssRNA, forms a dynamic complex with the nucleic acid, and strengthens the binding affinity of the RBD for ssRNA by about 50-fold. Overall, our data support a model where the NTD-RBD region can dynamically sample the large SARS-CoV-2 viral genome to identify highly specific binding regions. (Cubuk et al., BioRxiv 2023)

In the sixth chapter, I present an overview of the role of macromolecular crowding on disordered proteins (Cubuk & Soranno, ChemSysChem 2022).

In the seventh chapter, I discussed results of the crowding effects on the NTD-RBD configurations and binding affinities for RNA.

In the eighth chapter, I summarized the conclusion of my work, compared my results with current literature, and discussed potential future directions of the project. (Stringer & Cubuk et al., in submission).

In Appendix A, I have reported preliminary single-molecule FRET measurements highlighting the role of the serine arginine-rich linker in further aiding binding of RNA to the N protein.

In Appendix B, I have presented preliminary measurements of the role of nucleic acid binding in lowering the dimerization association constant of the N protein.

In Appendix C, I have presented results of single-molecule FRET on Apolipoprotein E, where I contributed to express and label some of the constructs as well as measured their reconfiguration times.

Overall, my thesis provides a systematic characterization of the biophysical properties of monomeric and dimeric forms of the SARS-CoV-2 N protein and reveals the important role of intrinsically disordered regions in aiding protein function, such as binding of RNA. These
findings identify key molecular interactions controlling viral genome packaging and pose the basis for further investigating condensation of nucleic acids, interactions with other proteins, and the impact of cellular crowding on protein conformations and association with nucleic acids.
1.8 References


7. Gelderblom, H. R. Structure and Classification of Viruses. in *Medical Microbiology* (ed. Baron, S.) (University of Texas Medical Branch at Galveston).


57. 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,


60. Fang, X. et al. Nucleocapsid amino acids 211 to 254, in particular, tetrad glutamines, are essential for the interaction between the nucleocapsid and membrane proteins of …. Journal of (2006).


70. Wright, P. E. & Dyson, H. J. Intrinsically unstructured proteins: re-assessing the protein


Chapter 2

The biophysics of disordered proteins from the point of view of single molecule fluorescence spectroscopy.

This chapter is adapted from:


Author contributions. J.C., M.D.S.B., and A.S. wrote the manuscript.
2.1 Abstract

Intrinsically disordered proteins and regions have emerged as key-players across many biological functions and diseases. Differently from structured proteins, disordered proteins lack stable structure and are particularly sensitive to changes in the surrounding environment. Investigation of disordered ensembles requires new approaches and concepts for quantifying conformations, dynamics, and interactions. Here, we provide a short description of the fundamental biophysical properties of disordered proteins as understood through the lens of single-molecule fluorescence observations. Single-molecule Foerster Resonance Energy Transfer (FRET) and Fluorescence Correlation Spectroscopy (FCS) provides an extensive and versatile toolbox for quantifying the characteristics of conformational distributions and the dynamics of disordered proteins across many different solution conditions, both in vitro and in living cells.

2.2 Introduction

The last twenty years have seen the emergence of a class of proteins that challenge the classic structure-function paradigm 1. Indeed, a large fraction of proteins in the eukaryotic proteome are completely or partially unstructured 2, but nevertheless play essential roles in biological function 3-6, ranging from transcription 7-10 and translation 11-13 to transport processes 14 and membrane organization 15-17. Their “shape shifting” nature upon binding 18 and the occurrence of short linear interacting motifs 19,20 make them essential components of cellular signaling, allowing for specificity with a multiplicity of signaling targets 19-23. The multivalence of interactions encoded in their sequence can also play an essential role in regulating self-assembly processes 3,24. Because of these key roles, disordered proteins are also central to several diseases, such as neurodegeneration and cancer 25,26.
There is a whole spectrum of conformational heterogeneity in proteins, where disorder can span either the whole protein, some regions of the protein (e.g., tails and linkers that flank and connect folded domains), or just short loops that are involved in the organization of structured proteins. The environment “complexity” can further modulate the properties of disordered proteins, since conformations and dynamics will depend on whether proteins are studied in isolation, in presence of ligands, or within the crowded milieu of a cell (Fig. 2.1).

**Figure 2.1. Spectrum of conformational disorder.** A whole spectrum of disorder can occur in proteins, from completely unstructured sequences to disordered linkers and tails flanking folded domains or just short loops that introduce local disorder in mostly folded proteins. In addition, given the lack of a stable 3D-structure, the complexity of the solution (in vitro and in the cell) can modulate conformations and dynamics of the disordered protein, whether the protein is studied in isolation, bound to a ligand, within a biomolecular condensate or other crowded environment.
These intrinsically disordered proteins (IDPs) and regions (IDRs) do not adopt a unique stable structure, but instead sample many different conformations. As a result, their flexible nature cannot be interpreted in the terms of classical structural biology and requires a different language to describe the properties of these conformational ensembles. At the same time different experimental approaches are required to access and quantify protein conformations, dynamics, and interactions.

The physics of polymers has emerged as a powerful framework to identify key observables and explain the thermodynamic driving forces controlling IDPs. In parallel, single-molecule fluorescence spectroscopy has provided means to directly test and quantify predictions of polymer models. In this review, we have focused on recounting the major findings of the biophysics of disordered proteins in a simple and accessible form, discussing how single-molecule fluorescence spectroscopy can be harnessed to access such properties. We hope this brief summary can provide an entry point in the investigation of IDPs biophysics with single-molecule tools.

2.3 Single-molecule fluorescence

Single-molecule fluorescence spectroscopy offers a versatile toolbox to investigate the conformations and dynamics of disordered proteins across many solution conditions (Fig. 2.2 and 2.3). The major advantages of single-molecule methods compared to classical ensemble approaches are: i) the possibility of resolving distinct conformational ensembles, allowing for distinguishing structural conformational changes occurring with solvent conditions or upon binding with ligands; ii) the ability of accessing protein dynamics contextually with information on the conformational ensemble; iii) the particularly low concentration regime (picomolar
range), which enables access to the monomeric form of the protein, even for aggregation and oligomerization prone sequences.

The two major single-molecule fluorescence approaches that are commonly applied to IDPs rely on Förster Resonance Energy Transfer (FRET) and Fluorescence Correlation Spectroscopy (FCS).

---

**Figure 2.2. Single-molecule fluorescence spectroscopy.** (A) In FRET, donor excitation can either result in a photon emission from the donor itself or a non-radiative energy transfer to the acceptor with consequent emission of an acceptor photon without a direct excitation of the acceptor. (B) The efficiency of the energy transfer process depends on the interdye distance $r$ to the power of 6 and on the characteristic Förster radius, which corresponds to the distance at which the efficiency is equal to 50% (compare with Eq. 2.1). (C) In fluorescence correlation spectroscopy, the amplitude of the correlation reports about the inverse of the number of molecules $N$ in the observation volume, whereas the shift in decay represents the characteristic time of decorrelation. In typical confocal experiments of freely diffusing molecules, one of the sources of decorrelation is the diffusion of the molecule in and out of the confocal volume and the characteristic time is the diffusion time $\tau_D$ (compare with Eq. 2.5 and 2.6). (D) Other timescales that can be studied in FCS experiments are related to the excitation of the fluorophore (antibunching), the rotational diffusion of the molecule and intrachain dynamics, and the photophysics of the triplet state.
FRET provides a molecular spectroscopic ruler to measure distances on the length scale of nanometers. FRET experiments involve the attachment of two fluorophores to the protein of interest and require that the emission spectra of one fluorophore (the donor) overlaps with the absorption spectrum of the other fluorophore (the acceptor). When this condition is satisfied, the excitation of the donor fluorophore can lead to the direct emission of a donor photon or to a non-radiative energy transfer toward the acceptor and the consequent emission of an acceptor photon. (Fig. 2.2A) The efficiency of this energy transfer is given by:

\[ E(r) = \frac{R_0^6}{R_0^6 + r^6} \]  

where \( r \) is the distance between the donor and the acceptor fluorophore and \( R_0 \) is the Förster radius. The Förster radius sets the distance at which the transfer efficiency is equal to 50% and depends on the spectroscopic properties of the dyes, their reciprocal orientation, and the refractive index of the solution. (Fig. 2.2B)

Single-molecule detection is achieved by limiting the number of molecules in the volume of observation. This can be directly realized by sparsely immobilizing molecules on the surface, such that the molecules are statistically sufficiently separated from each other, or by studying freely diffusing molecules and diluting the sample such that statistically only one molecule is observed at a time. Typical experiments make use of either total internal reflection fluorescence (TIRF) or confocal microscope setups.

In TIRF experiments, the total internal reflection enables a thin layer of the solution (< 100 nm from the coverslip) to be illuminated by an evanescent wave and the detection of excited fluorescent molecules is performed via a camera-based setup. The illumination of a small layer of solution allows studying interactions in presence of a high concentration of fluorescence...
proteins (e.g. ligands), since only the molecules close to the surface will be excited; however, it requires close proximity (normally, immobilization) of the sample to the surface. The camera-based setup enables the simultaneous detection of multiple molecules, though time resolution of fast events is limited by the camera frame rate (commonly in the tens of milliseconds timescale).

In single-molecule confocal setups, a laser beam is coupled into a microscope objective with a high numerical aperture, which focuses the beam into a diffraction limited spot within the sample. Emitted photons are detected through the same objective, filtered through a small pinhole (between 30 and 150 nm), and finally separated and detected on single-photon avalanche photodiodes. This type of setup is often coupled with fast electronics for single photon detection, allowing for molecules and fluorophore photophysics to be studied over a broad range of timescales, from picoseconds to minutes. Confocal microscopy allows for both studying freely-diffusing and immobilized molecules. However, for immobilized molecules it requires focusing on each single molecule individually, making the acquisition time for multiple immobilized molecules significantly longer compared to the one obtained by TIRF microscopy; on the positive side, this approach commonly allows for much higher time resolution on the fast timescales, enabling access to fluorescence lifetimes and anisotropy.

In both TIRF and confocal single-molecule FRET, the transfer efficiency $E$ is quantified by counting the number of donor and acceptor photons, $n_D$ and $n_A$, according to

$$E = n_A / (n_A + n_D)$$  \hspace{1cm} \text{Eq. 2.2}
If fluorescence lifetime is accessible, transfer efficiency can be also evaluated by comparing the lifetime of the donor in presence and in absence of the acceptor:

\[ E = 1 - \frac{\tau_{DA}}{\tau_D} \quad \text{Eq. 2.3} \]

where \( \tau_D \) is the fluorescence lifetime of the donor and \( \tau_{DA} \) is the fluorescence lifetime of the donor in the presence of the acceptor.

FCS provides an alternative set of approaches to investigate protein conformations (the overall size) and dynamics. The concept of fluorescence correlation was originally introduced in the 1970s by Elson, Magde, and Webb \(^{33-35} \) and relies on studying fluctuations in the intensity of fluorescence signals as a reporter for protein diffusion and concentration. This can be understood easily by considering a sufficiently dilute solution, where the passing of fluorescent molecules through the detection volume leads to significant changes in the detected intensity compared to the average background. This can be quantified by studying the correlation of the fluorescence fluctuations \(^{36,37} \):

\[ G(\tau) = \frac{\langle F(t)F(t+\tau) \rangle}{\langle F(t) \rangle^2} \quad \text{Eq. 2.4} \]

where \( F(t) \) represents the fluorescence intensity at the time \( t \), \( <> \) represents the temporal average over all times \( t \), \( \tau \) is the lag time at which the correlation \( G(\tau) \) is computed. Under the
assumption of a Gaussian 3D profile, the correlation can directly be linked to the diffusion time and concentration of molecules:

\[ G_{\text{Diff}}(\tau) = \frac{1}{N} (1 + \frac{\tau}{\tau_{\text{Diff}}})^{-1} (1 + \frac{\tau}{\tau_{\text{Diff}}} \cdot \frac{r_0^2}{z_0^2})^{-0.5} \]  \hspace{1cm} \text{Eq. 2.5}

where \( N \) is equivalent to the average number of molecules in the detection volume, \( \tau_{\text{Diff}} \) is the average time of molecules diffusing through the detection volume and \( r_0 \) and \( z_0 \) are the lateral and axial radial distances of the confocal volume, respectively (Fig. 2.2C). While the amplitude of the correlation directly reports on the average number of molecules in the detection volume, \( \tau_{\text{Diff}} \) reports about the diffusion coefficient of the molecule and can be related to its hydrodynamic Stokes-radius, \( R_H \):

\[ R_H = 2 K_B T \tau_{\text{Diff}} / 3 \pi \eta r_0^2 \]  \hspace{1cm} \text{Eq. 2.6}

where \( K_B \) is the Boltzman constant, \( T \) is the absolute temperature, and \( \eta \) is the viscosity of the solution at a given \( T \).

FCS measurements are commonly performed in a confocal setup and therefore are compatible with single-molecule FRET experiments. Indeed, FCS is not only limited to measure diffusion and concentration of molecules: depending on the source causing fluctuations in fluorescence intensity, FCS can provide access to static quenching and other photophysical properties of the dyes as well as dynamics within the molecule of interest (Fig. 2.2D).
Figure 2.3: Experimental methods for the study of disordered proteins. Single-molecule fluorescence offers a versatile set of tools for the investigation of conformations and dynamics in IDPs and IDR s. FCS provides access to the diffusion time $\tau_{\text{diff}}$ and corresponding hydrodynamic radius $R_H$ (see Eq. 6). Single-molecule FRET provides access to the mean transfer efficiency, whereas the fluorescence lifetime provides constraints for estimating the shape of the distribution. The root mean square interdye distance or $R_H$ can be used to study the coil-to-globule transition while modulating interactions (e.g. denaturant or temperature). Dependence of the interdye distance with the sequence separation between the dyes provides access to the scaling exponent. Nanosecond-FCS provides access to fast dynamics, from nanoseconds to milliseconds and enables disentangling the contribution of solvent and internal friction, either by explicitly studying the viscosity of the solution or by monitoring dynamics as function of sequence separation. Burst Variance Analysis provides an alternative method to quantify slow dynamics in the microsecond to millisecond timescale. Photon trajectories provide access to kinetics of folding and binding of IDPs and transition path-times. Recurrence analysis and microfluidics enable further investigation of kinetics associated with conformational changes in different solvents or upon binding ligands. The experimental quantities can be further compared with simulations or polymer models.
### 2.4 Accessing the disordered state

In 1999, the labs of DeGrado and Hochstrasser proved the feasibility of single-molecule FRET experiments on proteins, investigating the denaturation of immobilized \(^{38}\) and freely diffusing \(^{39}\) GCN-4 fragments. Contextually, Deniz et al. \(^{40}\) demonstrated how this approach enables distinguishing and separating properties associated with the folded and unfolded states. Specifically, their measurements of the chymotrypsin inhibitor 2 identified changes of the unfolded state as a function of the denaturant, which the authors speculated could be due to loss of secondary structure or increase of solubility of the disordered state. Schuler et al. \(^{41}\) observed analogous changes for the Cold shock protein of *Thermotoga Maritima* and, using polyproline sequences as control measurements, ruled out that the observations were due to photophysical effects and indeed represented changes in the conformations of the unfolded state. These works set the stage for investigating the properties of the denatured state of foldable proteins, which represents another occurrence of “disorder” in proteins.

In the case of denatured proteins, the mean transfer efficiency associated with the distribution of the unfolded state represents a mean value of the transfer efficiencies associated with the different configurations explored by the chain: because chain dynamics are significantly faster (nanosecond timescale) than the burst detection (millisecond) in single-molecule experiments, the different transfer efficiencies are averaged out \(^{31}\). The same applies to disordered proteins. The distance distribution associated with the denatured or disordered protein is often sufficiently well-described by simple polymer models (such as the Gaussian chain \(^{42}\), Worm-like chain \(^{43}\), and Self-Avoiding Walk distributions \(^{29,44,45}\)), or can be estimated from simulations \(^{28,46–48}\).
The work of Sherman and Haran \(^{49}\) on the unfolded state of protein L introduced an important explanation for the conformational change of the disordered state when increasing denaturant, linking the observed phenomenon to the coil-to-globule transition in polymers \(^{49-51}\). Indeed, depending on the balance of the monomer-monomer and monomer-solvent interactions, polymers can adopt conformations of a compact globule (when monomer-monomer interactions dominate) or expanded conformations (when solvent-monomer interactions are favored) \(^{42}\). Varying the strength of monomer and solvent interactions leads to a transition between these two extreme cases and results in the so-called coil-to-globule transition \(^{52}\). Examples of interactions include excluded volume interactions (the physical occupancy of space of the residues), electrostatic interactions, and hydrophobic interactions. A particular case occurs when attractive and repulsive interactions in the system (protein and solvent) cancel each other: this particular condition is referred to as the “theta state” or “theta solvent condition” and is a key reference state in polymer physics. Altogether, the coil-to-globule framework provides a physical explanation for the change in conformations of the unfolded (either denatured or intrinsically disordered) state in presence of denaturant, where the denaturant acts as a better solvent and favors more expanded conformations.

This hypothesis has been further tested by Hofmann et al. \(^{53}\). Here, they harnessed the ability of measuring interdye distances within a protein by attaching fluorophores at different positions. This approach allows directly estimating the scaling exponent \(^{54}\) of the disordered state under different solvent conditions, for both folded and intrinsically disordered proteins. The scaling exponent relates the root-mean-square interdye distance \(< r^2 >^{0.5}\) to the degree of polymerization of the molecule (in this case the number of peptide bonds in the protein) and is directly related to the three fundamental states identified in the coil-to-globule transition. \(< r^2 >^{0.5}\) is proportional
to $N^{1/3}$ for the case of a globule, to $N^{1/2}$ for the case of an ideal chain, and to $N^{0.588}$ for the case of a chain dominated by excluded volume effects. Hofmann et al.\textsuperscript{53} found that in high denaturant different sequences follow the scaling exponent expected for a chain in a good solvent, whereas under native conditions different scaling exponents are observed depending on the sequence properties (charge, hydrophobicity, etc.). (Fig. 2.4A) This observation was consistent with predictions from simulations\textsuperscript{55,56} and with previous estimates of the scaling exponent of folded proteins in denaturant\textsuperscript{57} and has been further corroborated in subsequent works exploring the role of denaturant on unfolded proteins\textsuperscript{58–60}. FRET measurements have been also compared with FCS measurements and dynamic light scattering to confirm the expansion of the disordered state of proteins with increasing denaturant concentration\textsuperscript{49,53,61,62}.

The information from single-molecule FRET is often integrated with that provided by other approaches, which can help in defining the properties of the disordered ensemble. This includes measurements of the radius of gyration via SAXS\textsuperscript{45,59,61,63–67}, local conformations via NMR\textsuperscript{65–68}, and simulations\textsuperscript{28,46–48,63,65,67,69}.
(A) **Scaling exponents**

<table>
<thead>
<tr>
<th>R_g (nm)</th>
<th>10</th>
<th>7</th>
<th>5</th>
<th>3</th>
<th>2</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>hCyp</td>
<td>10</td>
<td>100</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
</tr>
<tr>
<td>CspTm</td>
<td>10</td>
<td>100</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
<td>1000</td>
</tr>
</tbody>
</table>

inter-dye distance

(B) **Temperature-induced collapse**

<table>
<thead>
<tr>
<th>Transfer Efficiency</th>
<th>Csp M34</th>
<th>Integrase</th>
<th>λ repressor</th>
<th>ProTuN</th>
<th>ProTuC</th>
</tr>
</thead>
<tbody>
<tr>
<td>T (K)</td>
<td>290</td>
<td>310</td>
<td>330</td>
<td>350</td>
<td>350</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>R_g (nm)</th>
<th>0.4</th>
<th>2.0</th>
<th>3.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>T (K)</td>
<td>290</td>
<td>310</td>
<td>330</td>
</tr>
</tbody>
</table>

(C) **Conformational changes upon binding**

<table>
<thead>
<tr>
<th>Fraction Population</th>
<th>SDS (mM)</th>
<th>m</th>
<th>n</th>
<th>l</th>
</tr>
</thead>
<tbody>
<tr>
<td>m</td>
<td>10</td>
<td>100</td>
<td>1000</td>
<td>1000</td>
</tr>
<tr>
<td>n</td>
<td>10</td>
<td>100</td>
<td>1000</td>
<td>1000</td>
</tr>
<tr>
<td>l</td>
<td>10</td>
<td>100</td>
<td>1000</td>
<td>1000</td>
</tr>
</tbody>
</table>

(D) **Highly dynamic complexes**

<table>
<thead>
<tr>
<th>g_M</th>
<th>g_A, g_B, g_C, g_T</th>
<th>τ (µs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td></td>
<td>0.2</td>
</tr>
<tr>
<td>1.2</td>
<td></td>
<td>0.2</td>
</tr>
<tr>
<td>0.8</td>
<td></td>
<td>0.2</td>
</tr>
<tr>
<td>1.2</td>
<td></td>
<td>0.2</td>
</tr>
</tbody>
</table>

(E) **Transition path-times in IDP interactions**

\[ \ln L = \frac{1}{t_{DC} (\mu s)} \]

(F) **Phase boundaries and diffusion in condensates**

Normalized amplitude

<table>
<thead>
<tr>
<th>T_D start</th>
<th>T_D deplet</th>
<th>T_D set</th>
</tr>
</thead>
</table>

(G) **in cell measurements**

Normalized events

![Cell images](image1.png)  
![Cell images](image2.png)
Figure 2.4. Examples of single-molecule fluorescence measurements. (A) Measurements of the interdye distance (represented in terms of apparent radius of gyration) as function of interdye sequence separation for the unfolded state of two proteins, human CyclophilinA (hCyp) and Cold shock protein from Thermatoga maritima (CspTm), reveal changes in the scaling exponent from low to high values with increasing denaturant concentration (adapted from ref. 53). (B) Temperature induced collapse of five different disordered regions, including a destabilized mutant of Cold shock protein, the N-terminal disordered tail of HIV-Integrase, λ repressor, and the N- and C-terminal halves of Prothymosin alpha (adapted from ref. 69). (C) Conformational changes of alpha synuclein upon binding with SDS vesicles. Single-molecule FRET enables resolving different intermediates and bound states as a function of SDS concentration (adapted from ref. 122). (D) Nanosecond FRET-FCS reveals fast dynamics in the complex between prothymosin alpha and histone H1. Autocorrelation of donor-donor (DD) and acceptor-acceptor (AA) exhibit a correlated decay, whereas the cross-correlation (AD) shows an anticorrelated increase in the signal. A relaxation time is extracted by globally fitting the three correlations. (compare with Eq. 2.7). The rapid increase in all the correlations at a few nanoseconds is related to the excitation rate of the fluorophore and reflects the fact that a single fluorophore can emit only a photon at a time: this is commonly referred to as antibunching. (E) Confocal measurements of tethered molecules provide access to kinetics in the interaction of TAD and NCBD, including quantification of the time spent in short-lived states such as the transition path-time, as a function of NaCl concentration (adapted from ref.127). (F) FCS measurements provide quantification of concentrations inside and outside of phase separated solutions as well as access to the diffusion time of the fluorescent molecules in both phases (adapted from ref. 158). Concentrations of labeled components can be estimated from the amplitude of the correlation, whereas diffusion time is measured from the decay time of the correlation (compare Eq. 2.5). (G) In cell single-molecule FRET measurements of Prothymosin alpha (yellow curve) reveal modulation of conformational changes compared to in vitro measurements (blue curve), which are amplified by inducing osmotic stress and increasing the concentration of molecules in the cell (red curve). Adapted from ref. 137

2.5 Effect of sequence on chain conformations

The large spectrum of scaling exponents observed under native conditions for different unfolded and disordered proteins 53,64,70,71 reflects the contribution of sequence composition on the conformations of disordered ensembles. Müller-Späh et al. 72 used single-molecule FRET to probe the conformations of unfolded proteins, showing the degree of compaction and expansion of the interdye distance depends on the charged residue content in the protein, in good agreement with prediction from polymer theories 73–75. Indeed, the theory of polyelectrolytes (polymers whose monomers carry all the same type of charge) and polyampholytes (polymers carrying both positive and negative charges) indicate that the dimension of charged polymers depends on the contribution of the total net charge of the protein (which favors the expansion of the chain) and the total number of charged residues of the chain (which represents local compaction along the
chain due to the positive and negative residues in the chain). These observations matched independent computational work from the Pappu group, which further demonstrated how the net charge, total number of charged residues, and charge patterning can modulate the conformations of IDRs, favoring extended, compact, and more exotic conformations such as tadpoles and hairpins. Advanced polymer models have been further developed to capture sequence specific effects of charge distribution, revealing how the pattern of charges in polyampholyte sequences can even lead to coexistence of two distinct ensembles of conformations. Importantly, parameters such as net charge and total charge are influenced by the ionization of charged groups, which may be affected by the local density of charged residues. Polyampholyte and polyelectrolyte theories capture also the conformational response to salt, including effects of charge screening and counterion condensation.

While charged residues are an important determinant of IDPs/IDRs conformations, other amino acids can play a fundamental role on protein conformations. For example, a comparative analysis of the hydrodynamic radii of different IDPs identified the content in proline residues as an essential factor modulating the disordered conformational ensemble. SAXS measurements and simulations of Ash1 further revealed that proline residues can favor expanded configurations even when the charge content of the protein is changed by phosphorylation and more compacted conformations are expected to be populated. Understanding the effect of sequence composition is essential to rationalize the conformational response to solution conditions and how this impacts the interaction with ligands. This is particularly important in the case of linkers, where sequence composition can modulate the local effective concentration of interacting regions.
2.6 Effect of temperature

Temperature modulation provides a wealth of information regarding the entropic and enthalpic contributions controlling molecular conformations and interactions. This is often realized using custom built temperature-controlled chambers, where the sample temperature can be directly controlled and calibrated against a well-known standard (e.g., the temperature-dependent lifetime of rhodamine B). Single-molecule FRET experiments on IDRs and unfolded proteins as a function of temperature have revealed a surprising result for the temperature dependence of chain dimensions. With increasing temperature, the disordered state conformations first undergo compaction before starting to expand again (Fig. 2.4B). Similar results have been reported also for different IDPs by dynamic light scattering and small angle x-ray measurements. At first glance, these observations appear counterintuitive. Based on the classical dependence of the coil-to-globule observed in polymers, one would expect that a temperature increase would result in an expansion of the chain due to decrease of the strength of the interactions at play. The strengthening of interactions with increasing temperature points to a contribution of the “hydrophobic effect”, but the extent of collapse is found to be amplified for hydrophilic charged sequences, not hydrophobic ones. These observations can be rationalized by accounting for the temperature-dependent solvation free energies of the sequence amino acids. The same temperature-induced collapse occurs when studying the unfolded state upon hot and cold denaturation, showing that the cold-denatured and hot-denatured state of a protein represent the same denatured state and their response to temperature follows the temperature trend of solvation free energies.
2.7 Accessing chain dynamics

While FRET is often used to determine conformational changes, it can also be used to investigate chain dynamics. Fast dynamics can be accessed by studying the autocorrelation of the donor and acceptor photons or the crosscorrelation of donor and acceptor. Use of zero-mode waveguides can shorten the acquisition time (down to tens of minutes) and enable measurements of very fast dynamics (low-nanoseconds). When studying fast dynamics with FCS, the correlation function in Eq. 2.5 extends to:

\[
G(\tau) = (1 - c_{ab}^{ij} \frac{\tau}{\tau_{ab}})(1 + c_{CD}^{ij} \frac{\tau}{\tau_{CD}})(1 + c_{T}^{ij} \frac{\tau}{\tau_{T}})G_{\text{Diff}}(\tau), \quad i = j \quad \text{Eq. 2.7a}
\]

\[
G(\tau) = (1 - c_{ab}^{ij} \frac{\tau}{\tau_{ab}})(1 - c_{CD}^{ij} \frac{\tau}{\tau_{CD}})(1 + c_{T}^{ij} \frac{\tau}{\tau_{T}})G_{\text{Diff}}(\tau), \quad i \neq j \quad \text{Eq. 2.7b}
\]

where \(i\) and \(j\) identify either the donor or acceptor photon emission, \(c_{ab}^{ij}\) and \(\tau_{ab}^{ij}\) represent the amplitude and relaxation time of the antibunching component, \(c_{CD}^{ij}\) and \(\tau_{CD}^{ij}\) capture the contribution in amplitude and relaxation time of the triplet state, and \(c_{T}^{ij}\) and \(\tau_{T}^{ij}\) report about the amplitude and relaxation time of chain dynamics. Note that the relaxation timescale \(\tau_{CD}\) is the same across all correlations, whereas the sign associated with the chain dynamics amplitude \(c_{CD}^{ij}\) depends on whether this is an autocorrelation \((i = j)\) or a crosscorrelation \((i \neq j)\). The reason for change of sign in the crosscorrelation is intuitively understood by considering that FRET, as measured by comparing donor and acceptor emission, is inherently anticorrelated: an increase in the acceptor emission comes at the cost of a decrease in donor signal and vice versa (see Fig. 2.4D).
This approach has enabled quantifying the timescale at which disordered proteins sample the distribution associated with their conformational ensemble, which is often in the order of tens or hundreds of nanoseconds. The measured quantity is the reconfiguration time, which is the time that it takes to the specific interdye distance to lose memory of the previous configuration. Chain dynamics are consistent with the expected behavior of simple polymer models \(^{102-104}\), where the overall reconfiguration time of the chain is given by the sum of two contributions, one dependent on the solvent and therefore on viscosity, and one independent of the solvent and is usually referred to as internal friction. Internal friction may arise from different molecular sources, including dihedral angle rotation \(^{105-107}\) and transient contact formation \(^{108}\). ns-FRET FCS provides an effective tool to quantify solvent and internal friction contributions. One method relies on studying the viscosity dependence of the reconfiguration time titrating a viscogen that slows down the dynamics of the chain: a linear extrapolation to zero viscosity provides a value that reports on internal friction contributions. Alternatively, chain dynamics across different segment lengths of the disordered region can provide access to the same quantities: in this case, internal friction causes deviations from the behavior of ideal chain, where the reconfiguration is expected to scale with the length of the segment \(^{104}\). Comparison with all-atom simulations \(^{109}\), NMR \(^{66,110}\), and Neutron Scattering \(^{111,112}\) experiments supports this model and the importance of internal friction effects.

Photo-induced Electron Transfer (PET) FCS provides a complementary point of view \(^{109,113-115}\). Here the autocorrelation decay reports about quenching of the fluorophore due to static contact formation with quenching residues (tryptophan, tyrosine, and histidines) or synthetic moieties. In this case, the rates of forming and breaking the static complex, \(k_{on}\) and \(k_{off}\), define the amplitude \(c_{CD} = k_{off}/k_{on}\) and the relaxation time \(\tau_{CD} = 1/(k_{off} + k_{off})\) associated with chain
dynamics (Eq. 2.7a). Since the contact formation time measures the time required for two specific residues to come into contact, this time is usually longer than the reconfiguration time and the same polymer models mentioned above can be used to connect the two quantities. As expected based on the polymer models\textsuperscript{103}, contact formation is also impacted by internal friction\textsuperscript{109}.

While in the majority of cases, both FRET- and PET-FCS detect dynamics on the nanosecond time scale, this approach can indeed identify dynamics up to the diffusion time of the molecule (or longer, for immobilized molecules). Analysis of dynamics in the microsecond timescale can also be achieved by using Burst Variance Analysis (BVA), which studies how the variance of transfer efficiency varies along the burst duration\textsuperscript{116}, or Hidden Markov analysis of the photon trajectory\textsuperscript{117,118}.

### 2.8 Accessing ligand interactions

Single-molecule fluorescence spectroscopy can provide a direct readout for the interaction between disordered proteins and their ligands, whether these are small solutes (e.g., ions), another protein, or nucleic acids. Importantly, FRET enables the contextual study of conformations and dynamics, allowing discerning whether the interaction leads to folding upon binding\textsuperscript{119–122} (Fig. 2.4C), diffusion of the disordered protein on the folded binding partner\textsuperscript{123}, or formation of a dynamic complex in which both components remain disordered (as in the case of prothymosin alpha and histone H1\textsuperscript{124,125} (Fig. 2.4D)). Study of surface-tethered molecules enables quantification of the on- and off-rates of association via the analysis of the dwelling times in the bound and unbound state\textsuperscript{120,126}, as well as of transition path times\textsuperscript{121,127} (Fig. 2.4E). Alternatives for studying kinetics of interactions are offered by microfluidic mixing and recurrence analysis
Microfluidic mixing provides common deadtimes of ~ 1 ms, comparable with classic stop flow experiments. Recurrence analysis relies on extremely dilute solutions, so dilute that there is a significant probability in short times to see the same molecule coming back (“reoccurring”) in the confocal volume more than observing a new molecule. This provides a direct tool to identify subpopulations in broad distribution histograms.

2.9 IDPs in complex environments

The single-molecule fluorescence toolbox is constantly growing and advancements in the last decade have demonstrated the feasibility and relevance of investigating IDPs in a complex environment. Indeed, the lack of a stable 3D-structure makes disordered proteins prone to changes due to the surrounding milieu. We have already discussed how salt, temperature, and interactions with other molecules can alter IDPs conformations and dynamics. This becomes even more relevant in the context of the intracellular medium. Indeed, the cell milieu is occupied by a significant fraction of components (proteins, nucleic acids, metabolites), which is sufficiently high to limit the available volume that each component can explore. This phenomenon is commonly referred to as crowding and impacts many aspects of proteins and nucleic acids, including folding, dynamics, interactions, and in the case of disordered proteins, also their conformations. Recent applications on both proteins and nucleic acids have now highlighted the advantages of single-molecule fluorescence spectroscopy in quantifying such contributions on a molecule of interest. Indeed, single-molecule FRET can provide a direct readout for the conformational and kinetic changes within crowded solutions as well as probe the possibility of inducing “folded” conformations in equilibrium with the disordered state. One important finding is that disordered proteins sense crowding differently depending...
on their degree of expansion \textsuperscript{133}. IDRs that are more expanded are more sensitive to changes in the crowded solution. This can be quantified and rationalized in terms of their scaling exponent. IDRs with a scaling exponent larger than 0.5 will collapse with increasing concentration of crowding agents, whereas IDRs with a scaling exponent equal or smaller than 0.5 will have less or negligible impact of the crowded milieu, as expected based on the theory of polymer mixtures \textsuperscript{136}. A discussion of the polymer and colloidal theories that applies to crowding are reported in \textsuperscript{71}. The trend observed \textit{in vitro} with polymeric crowded solutions is also observed in cells, where the same proteins undergoes compaction when decreasing the available volume in the cell, e.g. by inducing osmotic stress \textsuperscript{137}. The emerging role of membrane-less compartments in the cellular environment has pushed to investigate the driving forces controlling their assembly. In particular, IDRs have been largely investigated in the context of biomolecular condensates. Recent experiments have showcased the possibility of using: i) FCS to determine concentrations and dynamics inside the dilute and dense phase of condensates \textsuperscript{138–142}; ii) single-molecule FRET to quantify the conformational changes of disordered proteins within the dense phase \textsuperscript{143,144}. (Fig. 2.4 F-G)

### 2.10 Summary points

- Single-molecule fluorescence spectroscopy provides a versatile toolbox to investigate the biophysical properties of disordered proteins in isolation, when bound to another macromolecule, or within crowded environments.

- Measuring interdye distances \textit{via} FRET enables studying contribution of the sequence composition and temperature effects on the dimensions of the disordered ensemble, providing direct access to physical parameters such as the scaling exponent.
Nanosecond FRET- and PET-FCS provides access to fast chain dynamics and enables resolving solvent and internal friction contributions, whereas FCS captures conformational changes of the whole protein in terms of hydrodynamic radius.

Integration of single-molecule fluorescence with orthogonal techniques such as SAXS, NMR, DLS, Neutron Scattering, and simulations provide further constraints to capture the complexity of conformational ensembles along different length- and time-scales.

Advancements in single-molecule spectroscopy and the growing understanding of the physical principles regulating disordered proteins pave the way to applying these methodologies to quantitatively understand the mechanisms of function and dysfunction.

**Acknowledgements.** We thank Alessandro Borgia, Hagen Hofmann, Alex Holehouse, Dmitrii Makarov, Tanja Mittag, Daniel Nettels, Rohit Pappu, and Ben Schuler, for many stimulating discussions over the years on the topics of single-molecule spectroscopy and intrinsically disordered proteins.

**Declaration of Interests.** The authors declare no competing interest.

**Funding.** The contribution of J.C., M.D.S.B., and A.S. to this work is supported by the NIH National Institute on Aging (Soranno, R01AG062837) and NIH National Institute on Allergic and Infectious Diseases (Soranno, R01AI163142).
2.11 References.


41. Schuler, B., Lipman, E. A. & Eaton, W. A. Probing the free-energy surface for protein


55. Tran, H. T. & Pappu, R. V. Toward an accurate theoretical framework for describing


80. Samanta, H. S., Chakraborty, D. & Thirumalai, D. Charge fluctuation effects on the shape of


103. Echeverria, I., Makarov, D. E. & Papoian, G. A. Concerted dihedral rotations give rise to


131. Tyagi, S. et al. Continuous throughput and long-term observation of single-molecule FRET


Chapter 3

The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA

This chapter is adapted from:


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.
3.1. 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.

3.2. 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) ¹. 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 ¹-³. While recent progress in vaccine development has been remarkable, the emergence of novel coronaviruses in human populations represents a continuing threat⁴. 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.

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. Here we performed a high-resolution structural and biophysical characterization of the SARS-CoV-2 nucleocapsid (N) protein, the protein responsible for genome packaging. A large fraction of N protein is predicted to be intrinsically disordered, which constitutes a major barrier to conventional structural characterization. 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. 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. 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.

3.3. Results

Coronavirus nucleocapsid proteins are multi-domain RNA binding proteins that play a critical role in many aspects of the viral life cycle \(^{12,28}\). The SARS-CoV-2 N protein shares substantial sequence conservation with other coronavirus nucleocapsid proteins (Fig. S3.1-3.5). 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. 3.1A-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.
and while the dimerization domain facilitates the formation of well-defined stoichiometric dimers, RNA-independent higher-order oligomerization is also expected to occur. Importantly, protein-protein and protein-RNA interaction sites have been mapped to all three disordered regions.

Despite recent structures of the RBD (Fig. 3.1B) and dimerization domains (Fig. 3.1C) from SARS-CoV-2, the solution-state conformational behavior of the full-length protein remains elusive. 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.

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. 3.1A and Table S3.1). 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 S3.2). 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.
3.3.1. 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 $0.65 \pm 0.03$ (Fig. 3.2A and Fig. S3.6). 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. S3.7A). To properly quantify the timescale associated with these fast structural rearrangements, we leveraged nanoseconds FCS. As expected for a dynamic population $^{45,46}$, the cross-correlation of acceptor-donor photons for the NTD is anticorrelated (Fig. 3.2B and S3.12). A global fit of the donor-donor, acceptor-acceptor, and acceptor-donor correlations yields a reconfiguration time $\tau_r = 170 \pm 30$ ns. This is longer than reconfiguration times observed for other proteins with a similar persistence length and charge content $^{46-49}$, 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 $^{50}$. 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 $r_{1-68}$ of $48 \pm 2$ Å. When using the recently proposed self-avoiding walk (SAW) model (Zheng et al., 2018) (see SI), we compute a value of $r_{1-68} 47 \pm 2$ Å. This corresponds to values of persistence length (see SI) equal to $4.5 \pm 0.4$ Å and $4.3 \pm 0.4$ Å for the Gaussian and SAW distribution, respectively, which are similar to values reported for another unfolded protein.
under native conditions ⁴⁶⁻⁴⁸,⁵¹. 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 $r_{1-68}$ is $46 \pm 2$ 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. S3.8 and Table S3.2). 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. S3.6 and S3.8). 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. 3.2C and Fig. S3.6, S3.8) and this contribution to the expansion of the chain can be described using an empirical binding model ⁵²⁻⁵⁶. 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; $\Delta G_0 = (5 \pm 1) \text{ kcal mol}^{-1}$) are presented in Fig. 3.2C. A comparative fit of the histograms assuming two overlapping populations yields a consistent result in terms of RBD stability and protein conformations (Fig. S3.9). 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. S3.9 and Table S3.6).

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). We also performed simulations of the NTD in isolation.
**Figure 3.2:** 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 S1). 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. S3.7, which accounts for denaturant binding (see Table S3.2) and unfolding of the folded RBD. The dashed line represents the estimate of folded RBD across different denaturant concentrations based on Eq. S3.8. Error bars represent propagation ± 0.03 systematic error in measured transfer efficiencies (see S1). **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 39). 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. 3.1B, center). **H)** The C-terminal half of transient helicity in H2 encodes an arginine-rich surface. We observed good agreement between simulation and experiment for the equivalent inter-residue distance (Fig. 3.2D). 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 38. 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. 3.2E). In particular, the basic beta-strand extension from the RBD (Fig. 3.1B) 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. 3.2G). 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. 3.2F, 3.2H). These features provide molecular insight into previously reported functional observations (see Discussion).
3.3.2. 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 $0.55 \pm 0.03$ and $0.75 \pm 0.03$, respectively (Fig. 3.3A). 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. 3.3C). 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. S3.7A). ns-FCS confirms fast dynamics across the measured distribution of transfer efficiencies, with a characteristic reconfiguration time $\tau_r$ of $120 \pm 20$ ns (Fig. 3.3B and S3.12). This reconfiguration time is compatible with high internal friction effects, as observed for other unstructured proteins, but may also account for the drag of the surrounding domains. The root-mean-square interdye distance corresponding to the low transfer efficiency population $r_{172-245}$ is equal to $55 \pm 2$ Å ($l_p = 5.4 \pm 0.4$ Å) when assuming a Gaussian chain distribution and $54 \pm 2$ Å ($l_p = 5.2 \pm 0.4$ Å) when using a SAW model (see SI). The one corresponding to the high transfer efficiency population is equal to $42 \pm 2$ Å when assuming a Gaussian Chain distribution or $45 \pm 2$ Å using the SAW model (with a corresponding $l_p = 3.2 \pm 0.3$ Å and $l_p = 3.6 \pm 0.3$ 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. S3.6, S3.8). These results correspond to an almost linear expansion of the chain in response to denaturant (see Fig. 3.3C).

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 S3.2). smFRET measurements of this truncated version revealed a single population that undergoes a strong compaction with decreasing GdmCl concentration (Fig. S3.6, S3.8). 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. S3.10). 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. 3.3D). 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. 3.3E). 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. 3.3F). 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).
Figure 3.3: 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. S3.6, 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. 3.3F). 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 orients 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.

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. 3.3F). 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).

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

Finally, we again applied single-molecule FRET (Fig. 3.4A) and nsFCS (Fig. 3.4B) 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 $0.59 \pm 0.03$ (Fig. 3.4A) and the denaturant dependence follows the expected trend for a disordered region, with a shift of the transfer efficiency toward lower values (Fig. 3.4C, and Fig. S3.6 and S3.8), 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. S3.8) reveals a very similar distribution, with almost identical mean transfer efficiency but a narrower width (Fig. S3.6), 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. S3.7A) appears to support a dynamic nature underlying the CTD FL population, nsFCS reveals a flat acceptor-donor cross-correlation on the nanosecond timescale (Fig. 3.4B). 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 $^{62}$, 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 \(^{63,64}\). 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. S3.12). 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 \(\tau_r\) 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 \(r_{363-419} = 51 ± 2 \text{ Å} \quad (l_p = 6.1 ± 0.5 \text{ Å})\) for a Gaussian chain distribution and \(r_{363-419} = 48 ± 1 \text{ Å} \quad (l_p = 5.4 ± 0.4 \text{ Å})\) 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 \(^{65}\).

We again obtained good agreement between all-atom Monte Carlo simulations and experiments (Fig. 3.4D). 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. 3.4E). 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. 3.4F). 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. 3.4G). 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. 3.4H). 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.
Figure 3.4: 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. S3.6, 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).
3.3.4 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. Many of the proteins that have been shown to drive phase separation in vitro are RNA binding proteins with intrinsically disordered regions. Moreover, multivalency is the key molecular feature that determines if a biomolecule can undergo higher-order assembly. 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.

In line with this expectation, we observed robust droplet formation with homopolymeric RNA (Fig. 3.5A-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. 3.5C-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. 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. S3.13](#)). 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. 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 20–27.

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.
Figure 3.5: 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.

3.3.5 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 \(^7^7\).

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 \(^7^8–^8^0\). 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. 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. 3.6A). 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. 3.6B-E). Phase separation gives rise to a single large spherical cluster with multiple polymers and binders (Fig. 3.6D, 3.6H-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. 3.6A). 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.85–87.

We performed identical simulations to those in Fig. 3.6C-D using the same system with polymers that now possess a single high-affinity binding site (Fig. 3.6E). Under these conditions we did not observe large phase separated droplets (Fig. 3.6F). Instead, each individual polymer undergoes collapse to form a single-polymer condensate (Fig. 3.6E). 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. 3.6 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. 3.6G).
Figure 3.6: 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.
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. S3.18). 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 \(^88,89\).

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. 3.6L). 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. 3.6J). 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 \(^90\). 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.

### 3.4 Discussion

The nucleocapsid (N) protein from SARS-CoV-2 is a multivalent RNA binding protein critical for viral replication and genome packaging \(^{11,12}\). 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.
3.4.1 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. To assess the formation of oligomeric species, we use a combination of NativePAGE, crosslinking and FCS experiments (see Fig. S3.14 and SI). These experiments also verified that under the conditions used for single-molecule experiments the protein exists only as a monomer.

3.4.2 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. 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.
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. 3.2H and 3.3F). 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 31,97,98. 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 99,100. Recent NMR data also shows excellent agreement with our results, identifying a transient helix that shows 1:1 overlap with H3 24. 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 101,102.

Transient helix H4 (in the LINK, Fig. 3.3F) was previously predicted bioinformatically and identified as a conserved feature across different coronaviruses, in agreement with our own secondary structure predictions (Fig. S3.19) 31. 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 103. 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 20,26.

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 20. Transient helix H6 is an amphipathic
helix with a highly hydrophobic face (Fig. 3.4H). Recent hydrogen-deuterium exchange mass spectrometry also identified H6 \[^{43}\]. Residues in this region have previously been identified as mediating M-protein binding in other coronaviruses, such that we propose H6 underlies that interaction \[^{21,104-106}\]. 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 \[^{107,108}\].

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.SI/3/5).

3.4.3 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 \textit{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 \textit{via in cell} experiments \[^{20-27}\]. Taken together, these results demonstrate that N protein \textit{can} undergo \textit{bona fide} phase separation, and that N protein condensates \textit{can} form in cells. Nevertheless, the complexity introduced by multidimensional linkage effects \textit{in vivo} could substantially influence the phase behavior and composition of condensates observed in the cell \[^{80,83,109}\]. Of note, the regime we have identified in which phase separation occurs (Fig. 3.5) 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 \textsuperscript{110–116}. 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 \textsuperscript{24,117}. 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 \textsuperscript{77}. 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. 3.7). 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\(^{118-120}\). 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)\(^{117,121-123}\).

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\(^ {22}\). 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 \(^71,82\). The ability of multivalent disordered proteins to drive RNA compaction has been observed previously in various contexts \(^14,124\). Furthermore, genome compaction by RNA binding protein has been proposed and observed in other viruses \(^120,125,126\), 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 \(^14,127\). 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 \(^73,117\).

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 \textit{in vivo} \(^5,23,72,79,128\). In particular, G3BP1 – an essential stress-granule protein that undergoes phase separation – was recently shown to co-localize with overexpressed N protein \(^24,73,79,129,130\). 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 \(^114,130\).
Our model is also in good empirical agreement with recent observations made for other viruses\textsuperscript{131}. 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 \textsuperscript{132–134}. 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. 3.6L). With this in mind, from a biotechnology standpoint, phase separation may be a convenient readout for \textit{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 \textsuperscript{72,82,135,136}. 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\textsuperscript{20}.

### 3.4.4 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.22

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.
Figure 3.7: 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.
3.5 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.

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/hiddenlab/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.

**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.
3.6 Methods

3.6.1 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/) \(^{59,137}\) 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 S3.7.

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 \(^{60,140,141}\). Post-simulation analysis was performed with Enspara \(^{142}\). For additional simulation details see the supplementary information.

3.6.2 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. 3.6 were run for $2 \times 10^9$ Monte Carlo steps, with four independent replicas. Bead interaction strengths were defined as shown in Fig. 3.6A. For additional simulation details see SI.

### 3.6.3 Protein Expression, purification, and labeling

SARS-CoV-2 Nucleocapsid protein (NCBI Reference Sequence: YP_009724397.2) including an N term extension containing His$_9$-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 His$_9$-SARS-CoV-2 Nucleocapsid pGEX vector to create the N protein constructs (SI Table S3.1) 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-His$_9$-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).
3.6.4 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).
3.7 Supplementary Information

3.7.1 Sequence Analysis

Disorder prediction was performed using IUPred2.0, with additional analysis and sequence parsing done with localCIDER and protfasta, respectively.\textsuperscript{29,144,145}

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 MSDNGPQNQR NAPRITFGGP SDSTGSNQNG ERSGARSKQR RPQGLP\textsuperscript{NNTA} 51 52 SWFTALTQHG KEDLKFPG\textsuperscript{GQ} GVPINTNSSP DDQIGYYRRA TRRIRG\textsuperscript{GDGK} 101 MKDLSPR\textsuperscript{WYF} YYLGTGPEAG LPYGANKD\textsuperscript{GI}IWVATEGALN TPKDH\textsuperscript{GTRN} 151 PANNAAI\textsuperscript{VLQ} LPQG\textsuperscript{TTLPKG F\textsuperscript{E}AEGSRGGS QASSRSSSRS RNSSRNSTPG 201 SSRGTSPARM AGNGGDAALA LLLL\textsuperscript{DRLNQ}L ESKMSGK\textsuperscript{GQQ} QQGQ\textsuperscript{TVTKKS} 251 AAEASK\textsuperscript{KPRQ KRTAT\textsuperscript{KAYNV} TQA\textsuperscript{GRRGPE Q}TQ\textsuperscript{GNFDQE LIRQGTDYKH 301 WPQIAQFAPS ASAFFG\textsuperscript{MSRI GMEVTPSGTW LTY}TGAIKLD DKDP\textsuperscript{NKFDQV} 351 ILLNKHIDAY KT\textsuperscript{P}PTEPKK DKKK\textsuperscript{ADETQ ALPQRQKKQQ TVT}LPAADL 401 DDFS\textsuperscript{KQLQQS MSSADSTQA}\textsubscript{A}

A complete list of constructs is presented in Table S3.1.

3.7.2 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.\textsuperscript{138} All simulations were performed at 330 K and at 15 mM NaCl, as have been used previously in various systems.\textsuperscript{72,139,146,147} 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/)\(^{148}\). 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 \(^{59}\). 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 \(^{149}\). 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 S3.5, and Table S3.7 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. 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. 3.1A) but in contrast to their behavior in the dimeric structure (Fig. 3.1C). 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. 3.1A. 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. We chose this strategy because at the time, no dimerization structure existed, a situation that has since resolved itself. 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.

### 3.7.3 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. 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. 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. Conformations were stored every 20 ps.

The FAST algorithm was used to enhance conformational sampling and quickly explore the dominant motions of nucleoprotein. 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. The FAST dataset was clustered using a k-centers algorithm based on RMSD between frames using backbone heavy atoms (C, Ca, Cβ, N, O) to generate 1421 discrete states, which were then launched on the distributed computing platform Folding@home.

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). 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.

3.7.4 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. 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.
3.7.5 Coarse-Grained Polymer Simulations

Coarse-grained simulations were performed using the PIMMS software package. 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. 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 $60 \times 10^9$ 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. S3.18).

### 3.7.6 Extended Discussion on Coarse-Grained Simulations

For simulations of homopolymeric polymers as shown in Fig. 3.6C,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. 3.6C 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.
3.7.7 Plasmid Construct Design.

SARS-CoV2 Nucleocapsid protein (NCBI Reference Sequence: YP_009724397.2) including an N term extension containing His$_6$-HRV 3C protease site –

CATCATCACCATCATCATCACCACCTCGAAGTTCTGTTCATCAGCGCCATAGCGCTCGTACGGT

ATCATCGATGATGATGTCGCTGGAGCTTGGAGCTGCTGACGATGGGCATCGCTGCTGACGATGGGC

TACATCGATGATGATGTCGCTGGAGCTTGGAGCTGCTGACGATGGGCATCGCTGCTGACGATGGGC

TACATCGATGATGATGTCGCTGGAGCTTGGAGCTGCTGACGATGGGCATCGCTGCTGACGATGGGC

TACATCGATGATGATGTCGCTGGAGCTTGGAGCTGCTGACGATGGGCATCGCTGCTGACGATGGGC

TACATCGATGATGATGTCGCTGGAGCTTGGAGCTGCTGACGATGGGCATCGCTGCTGACGATGGGC

TACATCGATGATGATGTCGCTGGAGCTTGGAGCTGCTGACGATGGGCATCGCTGCTGACGATGGGC

TACATCGATGATGATGTCGCTGGAGCTTGGAGCTGCTGACGATGGGCATCGCTGCTGACGATGGGC

TACATCGATGATGATGTCGCTGGAGCTTGGAGCTGCTGACGATGGGCATCGCTGCTGACGATGGGC

TACATCGATGATGATGTCGCTGGAGCTTGGAGCTGCTGACGATGGGCATCGCTGCTGACGATGGGC

TACATCGATGATGATGTCGCTGGAGCTTGGAGCTGCTGACGATGGGCATCGCTGCTGACGATGGGC

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

GST-LEVLFQGPLGSHHHHHHHHHELVLFQGPMSDNSGPQNRNAPRITFGGGSPTSGSN

QNGERSGARSKQRRPQPGLPNNTASWTALTVHQKEFLKFPQGRQPQVNTNNSPDQIGY

YRRATRRRIRGGDKMKSPPWYFYYLGTPGAEAPLYGANHDKGIIWATEGALNTPKD

HIGTRNPNAANAVLQLPQGTLPKQYAEAGRSRGGSGQASSRSSSSSRSRSSRNSSTPGGSSGT

SPARMAGNNGDAALALLLLDLRLNQLESKMSGKQQQQGTQTLPKSAAEASKKPRQKR

TAKAYNVTQAFRGRGPEQTQNGFQELQGGTDPKYHWPQIAQPASAFGMSRIG

MEVTPSGTWLTYGAIKLDKKDPNFKQDVILLNKHIDAYKTFPPTEPKKDKKKKADETQ

ALPQRQKQKQQTTLPAADLDDFSKQLQQSMSSADSTQA
Site-directed mutagenesis was performed on the His₉-SARS-CoV2 Nucleocapsid pGEX vector to create the N protein constructs (Table S3.1). All cloning and site-directed mutagenesis steps were performed by Genewiz and sequences were verified using sanger sequencing.

### 3.7.8 Protein Expression and Purification

Both GST-His₉-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 µg/mL) to OD₆₀₀ ~ 0.6 and induced with 0.2 mM IPTG for 12 hours at 16°C. Harvested cells were lysed with sonication at 4°C in lysis buffer (50 mM Tris pH 8, 500 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 an HisTrap FF column (GE Healthcare) in buffer A (50 mM Tris pH 8, 500 mM NaCl, 10% glycerol, 20 mM imidazole, 5 mM βME). 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 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 βME, 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 μg/mL) to OD$_{600}$ ~ 0.6 and induced with 0.2 mM IPTG for 3 hours at 37°C. Harvested cells were lysed with sonication at 4°C 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 βME and incubated at 4°C 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 βME) 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 βME, 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 µg/mL) to OD600 ~ 0.6 and induced with 0.2 mM IPTG for 3 hours at 37°C. Harvested cells were lysed with sonication at 4°C in lysis buffer (50 mM MES pH 6, 500 mM NaCl, 10% glycerol, 5 mM βME, 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 βME). 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 βME, 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 spectrosopically in 50 mM MES (pH 6.0), 300 mM NaCl, 10% (v/v) glycerol using an extinction coefficient = 120M⁻¹cm⁻¹

3.7.9 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. 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. 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. 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.

### 3.7.10 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).

3.7.11 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) 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).

3.7.12 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.
3.7.13 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 $E=n_A/\left(n_A+n_D\right)$, where $n_D$ and $n_A$ 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. 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 $\gamma_{PIE}$ 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, $R_0 = 5.4 \text{ nm}$ \(^{164}\). 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 $\pm 0.03$, 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 \(^{64,165}\). Standard deviation of the transfer efficiency for multiple replicates of the same experimental conditions commonly results in a standard deviation equal or less than $\pm 0.01$. 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.

3.7.14 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 $(t_A(R) - t_{A0})/t_{D0}$\textsuperscript{166}. 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 $\kappa^2 = 2/3$.

### 3.7.15 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\textsuperscript{167}. 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 $g(\tau) = 1 + \frac{1}{N} \left(1 + \frac{\tau}{\tau_D}\right)^{-1} \left(1 + \frac{\tau}{\alpha^2 \tau_D}\right)^{-1/2}$, where $N$ is the average number of molecules in the confocal volume, $\tau_D$ is the diffusion time along the xy
plane, $\alpha$ is the eccentricity of the three dimensional Gaussian observational volume. $\tau_D = \omega_{xy}^2 / 4 D$, where $D$ is the 3D translational diffusion coefficient and $\omega_{xy}$ 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 $\alpha = \omega_z / \omega_{xy}$.

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). The overall equation that we fit to the FCS traces is then

$$g(\tau) = 1 + (g_{diff}(\tau) - 1)(1 + cT1 \text{Exp}[\frac{-\tau}{\tau_{T1}}])(1 + cT2 \text{Exp}[\frac{-\tau}{\tau_{T2}}])$$

where $\tau_{T1}$, $\tau_{T2}$, $c_{T1}$, and $c_{T1}$, denotes the characteristic times and amplitudes of the contributions of two triplet states to $g(\tau)$. Parameters $\tau_{diff}$, $\tau_{T1}$, $\tau_{T2}$, $c_{T1}$, $c_{T2}$, and $N$ were fitted by least square nonlinear regression analysis for each concentration of unlabeled protein tested (Fig. S3.14 A-B), while $\alpha$ was fixed at a value of 6 determined independently from analysis of fluorescence intensity profiles of fluorescent nanobeads.

Making use of the definition of $\tau_{diff}$ and the Stokes-Einstein equation, we have, for each concentration of unlabeled protein $(\tau_{diff} / \tau_{diff0}) = (R_{h} / R_{h0})$, where $\tau_{diff0}$ and $R_{h0}$ are the diffusion time and hydrodynamic radius in the absence of unlabeled protein, respectively. Error bars in Fig. S3.14 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.
3.7.16 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 \(^\text{62,169}\). 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:

\[
g_{ij}(\tau) = 1 + \frac{1}{N}(1 - c_{AB} \exp[-(\tau - \tau_0)/\tau_{AB}])(1 + c_{CD} \exp[-(\tau - \tau_0)/\tau_{CD}])(1 + c_{T} \exp[-(\tau - \tau_0)/\tau_{T}])
\]

Eq S3.1

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 \(^\text{170}\). 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.
3.7.17 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 \(^{171}\). 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 \(^{172}\). 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 \(\nu\) for the SAW model.

Estimates of these parameters are obtained by numerically solving:

\[
\langle E \rangle = \int_0^{l_c} P(R) E(R) \, dr \quad \text{Eq. S3.2}
\]

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:

\[
E(R) = \frac{R^6}{R_0^6 + R^6} \quad \text{Eq. S3.3}
\]

The Gaussian chain distribution is given by:

\[
P_{FJC}(R, r) = 4\pi R^2 \left(\frac{3}{2\pi^2}\right)^{3/2} \exp\left(-\frac{3R^2}{2r^2}\right) \quad \text{Eq. S3.4}
\]

The SAW model can be expressed as:
\begin{equation}
P_{SAW}(R, \nu) = A_1 \frac{4\pi}{b_0 N^\nu} \left( \frac{R}{b_0 N^\nu} \right)^{2+g} \text{Exp} \left[ -A_2 \left( \frac{R}{b_0 N^\nu} \right)^{\delta} \right]
\end{equation}
Eq. S3.5

where \( A_1 = \frac{\delta}{4\pi} \frac{\Gamma(5+g/\delta)}{\Gamma(3+g/\delta)} \), \( A_2 = \left( \frac{\Gamma(5+g/\delta)}{\Gamma(3+g/\delta)} \right)^{\delta/2} \), \( g = (\gamma - 1)/\nu \), \( \delta = \frac{1}{\Gamma(1-\nu)} \), \( \gamma = 1.1615 \), \( \Gamma \) is the Euler Gamma Function, \( b_0 = 0.55 \) nm is an empirical prefactor, \( N \) is the number of residues between the fluorophores, and \( \nu \) 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 \( r_{m,n} = |m-n|^{0.5/\text{dye}} |m-n+2|^{0.5/\text{dye}} \) with \( /\text{dye} = 4.5 \). Finally, the persistence length is computed using the Gaussian conversion \( r^2 = 2l/l_p^8 \).

### 3.7.18 Binding of Denaturant and Folding.

As in previous works, we model the chain expansion with the denaturant in terms of a simple binding model:

\[ r(c) = r_0 \left( 1 + \rho \frac{Kc}{1+Kc} \right) \]

Eq. S3.6

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 unfolded $f_u$ accordingly to:

$$r(c) = (r_{0f}f_f + r_{0u}f_u)\left(1 + \rho \frac{Kc}{1+Kc}\right)$$ \text{ Eq. S3.7}$$

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

$$f_f = \frac{\text{Exp}[-m(c-c_{1/2})/RT]}{1+\text{Exp}[-m(c-c_{1/2})/RT]}$$ \text{ Eq. S3.8}$$

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 parameter $\Delta G_0$ can be computed as $\Delta G_0 = m c_{1/2}$.

### 3.7.19 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. S3.7). 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. S3.6). 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:

\[
f_u = \frac{1}{1 + K^{u-i} + K^{i-f}}, \quad f_i = \frac{1}{1 + (K^{u-i})^{-1} + (K^{i-f})^{-1}}, \quad f_f = \frac{1}{1 + (K^{u-i})^{-1} + (K^{i-f})^{-1}}
\]

Eq. S3.9

where \( K^{u-i} \) and \( K^{i-f} \) are

\[
K^{u-i} = \exp[-\frac{m^{u-i}(c - c_{1/2}^{u-i})}{RT}] ; \quad K^{i-f} = \exp[-\frac{m^{i-f}(c - c_{1/2}^{i-f})}{RT}]
\]

Eq. S3.10

Fitted values to the model are reported in Table S3.2. Importantly, the observed values confirm in large measure the inferred stability measured via the NTD. The small discrepancy in the overall stability observed (Fig. S3.9) 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.

3.7.20 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 \cite{173,174} (see Table S3.3), where the increasing concentration of ions screens the interaction between oppositely charged residues (see Fig. S3.11).

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. S3.11). 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. S3.8), 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. S3.10). For comparison no change is observed when studying the NTD FL under the same conditions (Fig. S3.10). These observations for the LINK-ΔDimer mimic what was previously observed in the case of the Cold Shock Protein from *Thermotoga Maritima* \(^{173}\) 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. S3.11), 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.

### 3.7.21 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 \(^{173,174}\), 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 \( \alpha \) is the ratio between \( l \) and \( l_0 \), with \( l \) being a rescaled segment that accounts for excluded volume and electrostatic interactions.

\( \alpha \) is computed according to the equation proposed by Higgs and Joanny\textsuperscript{173,174}:

\[
\alpha^5 - \alpha^3 = \frac{4}{3} \left( \frac{3}{2\pi} \right)^{1.5} N^{0.5} v^* 
\]

Eq. S3.13

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

\[
v^* b^3 = v b^3 + \frac{4\pi l_b (f-g)^2}{k^2} - \frac{\pi l_b^2 (f-g)^2}{k} \]

Eq. S3.14

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.

### 3.7.22 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\textsuperscript{31,34,44}. Indeed, NativePAGE experiments reveal the existence of multiple bands (Fig. S3.14 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. S3.14 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 S3.14 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.

### 3.7.23 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.
3.7.24 **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.

3.7.25 **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. 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. 176.
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 $F(x)$:

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

where

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

where $x$ denotes poly(rU) concentration and $A$, $\beta$, $\sigma$ and $\mu$ are parameters fitted through weighted minimum least squares for each protein concentration (solid lines in Fig. 3.5 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. 3.5 C-D (solid lines). Contour lines were computed at 1, 10, 20, 50 and 100 times the limiting value employed (\( A_{340 \text{ nm,1mm}} = 0.005 \)).
### 3.8 Supporting Tables

#### Table S3.1. N protein constructs

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
<th>Start Position (WT)</th>
<th>End Position (WT)</th>
<th>Labeling Positions</th>
</tr>
</thead>
<tbody>
<tr>
<td>WildType (WT)</td>
<td>GPMSDNGPQNQRNAPRITFGGPPSDSTGSNQNGERSGARSKQRRPQGPLPNNTASWFTALTQHGKEDLKFPFRGQGVPINTNSSPDDQIQGYYRRATARRIRGGDGKMKDLSPRWYFYYLGTGPEAGLPHYGANKDGIIWVATEGALNTPKDHIGTRNPANNAIVLQLPQGTTLPKGFYAEGSRGGSQA</td>
<td>1</td>
<td>419</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>GSNQNGERSGARSKQRRPQGPLPNNTASWFTALTQHGKEDLKFPFRGQGVPINTNSSPDDQIQGYYRRATARRIRGGDGKMKDLSPRWYFYYLGTGPEAGLPHYGANKDGIIWVATEGALNTPKDHIGTRNPANNAIVLQLPQGTTLPKGFYAEGSRGGSQA</td>
<td>1</td>
<td>419</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>AGNGGDAAALALLLDDLRNLQLESKMSGKGQQQQQGQTGTKSAAAEASKKPRQKRTATKAYNVTQAFGRGPEQTQGNGDQELIRQGTDDYKHWQPQIAQFAPSASSAFFGMSRIEMEGTVPSGTWTLYTGAIKLDDKDPNFKDQVILLNKHIDAYKTFPPTEPKKDKKKKKADETQALPQRQKKQQT</td>
<td>1</td>
<td>419</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>VTLLPAADLDDFSKQLQQSMSSADSTAAGNGGDAAALALLLDDLRNLQLESKMSGKGQQQQQGQTGTKSAAAEASKKPRQKRTATKAYNVTQAFGRGPEQTQGNGDQELIRQGTDDYKHWQPQIAQFAPSASSAFFGMSRIEMEGTVPSGTWTLYTGAIKLDDKDPNFKDQVILLNKHIDAYKTFPPTEPKKDKKKKKADETQALPQRQKKQQT</td>
<td>1</td>
<td>419</td>
<td>-</td>
</tr>
<tr>
<td>NTD FL</td>
<td>GPSCSDNGPQNQRNAPRITFGGPPSDSTGSNQNGERSGARSKQRRPQGPLPNNTASWFTALTQHGKEDLKFPFCQGQVPINTNSSPDDQIQGYYRRATARRIRGGDGKMKDLSPRWYFYYLGTGPEAGLPHYGANKDGIIWVATEGALNTPKDHIGTRNPANNAIVLQLPQGTTLPKGFYAEGSRGGSQA</td>
<td>1</td>
<td>419</td>
<td>M1C-R68C</td>
</tr>
<tr>
<td></td>
<td>AGNGGDAAALALLLDDLRNLQLESKMSGKGQQQQQGQTGTKSAAAEASKKPRQKRTATKAYNVTQAFGRGPEQTQGNGDQELIRQGTDDYKHWQPQIAQFAPSASSAFFGMSRIEMEGTVPSGTWTLYTGAIKLDDKDPNFKDQVILLNKHIDAYKTFPPTEPKKDKKKKKADETQALPQRQKKQQT</td>
<td>1</td>
<td>419</td>
<td>M1C-R68C</td>
</tr>
<tr>
<td></td>
<td>VTLLPAADLDDFSKQLQQSMSSADSTAGNGGDAAALALLLDDLRNLQLESKMSGKGQQQQQGQTGTKSAAAEASKKPRQKRTATKAYNVTQAFGRGPEQTQGNGDQELIRQGTDDYKHWQPQIAQFAPSASSAFFGMSRIEMEGTVPSGTWTLYTGAIKLDDKDPNFKDQVILLNKHIDAYKTFPPTEPKKDKKKKKADETQALPQRQKKQQT</td>
<td>1</td>
<td>419</td>
<td>M1C-R68C</td>
</tr>
<tr>
<td>NTD RBD</td>
<td>QA</td>
<td>1</td>
<td>173</td>
<td>M1C-R68C</td>
</tr>
<tr>
<td>---------------------------------------------</td>
<td>-------------------------</td>
<td>----</td>
<td>-----</td>
<td>------------</td>
</tr>
<tr>
<td>GP CSD NGPQNQRNAPRITFGGGPSDSTG</td>
<td>SNQNGERSGARS KQR RR PQ QLPNNTAS WFTALTQHGKE DLFKPF CGQGV PINT N SPPPDDIQIGYYRRATRRIRRGGD GKM KD LSPRWYFYYLGTGPEAGLPYG ANKD IIWVATEGALNTPK DHIGTRNP ANNA IVLQLPQGTTLPKGFCAEGSRGGQA SSSRSSRRNSSRNSTPGSSRGTSPARM AGNGGDAAALALLLLLLDLRN QLES KMS GKGQQQQQGTVTKSAAEASKPRQ KRTAKAYNVT QA FGRRGPEQTQGNF GDQELIRQGT D YKHWPQIAQFAPSAS AFFGMSRIGMEVTPSWLT YTGAI KL DDKDPFNFKDVILLNKHIDAY KTFPPT EPKKDKK KKADE TQALPQRQKKQQT VTLPPAADLD DFSKQLQQSM SSS ADST QA</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RBD FL</th>
<th>QA</th>
<th>1</th>
<th>419</th>
<th>R68C-Y172C</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPMSDNGPQNQRNAPRITFGGGPSDST</td>
<td>GSNQNGERSGARS KQR RR PQ QLPNNTA WFTALTQHGKE DLFKPF CGQGV PINT N SPPPDDIQIGYYRRATRRIRRGGD GKM KD LSPRWYFYYLGTGPEAGLPYG ANKD IIWVATEGALNTPK DHIGTRNP ANNA AIVLQLPQGTTLPKGFCAEGSRGGQA SSSRSSRRNSSRNSTPGSSRGTSPARM AGNGGDAAALALLLLLLDLRN QLES KMS GKGQQQQQGVCVT KSSAAEASKPRQ KRTAKAYNVT QA FGRRGPEQTQGNF GDQELIRQGT D YKHWPQIAQFAPSAS AFFGMSRIGMEVTPSWLT YTGAI KL DDKDPFNFKDVILLNKHIDAY KTFPPT EPKKDKK KKADE TQALPQRQKKQQT VTLPPAADLD DFSKQLQQSM SSS ADST QA</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LINK FL</th>
<th>QA</th>
<th>1</th>
<th>417</th>
<th>Y172C-T245C</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPMSDNGPQNQRNAPRITFGGGPSDST</td>
<td>GSNQNGERSGARS KQR RR PQ QLPNNTA WFTALTQHGKE DLFKPF CGQGV PINT N SPPPDDIQIGYYRRATRRIRRGGD GKM KD LSPRWYFYYLGTGPEAGLPYG ANKD IIWVATEGALNTPK DHIGTRNP ANNA AIVLQLPQGTTLPKGFCAEGSRGGQA SSSRSSRRNSSRNSTPGSSRGTSPARM AGNGGDAAALALLLLLLDLRN QLES KMS GKGQQQQQGVCVT KSSAAEASKPRQ KRTAKAYNVT QA FGRRGPEQTQGNF GDQELIRQGT D YKHWPQIAQFAPSAS AFFGMSRIGMEVTPSWLT YTGAI KL DDKDPFNFKDVILLNKHIDAY KTFPPT EPKKDKK KKADE TQALPQRQKKQQT VTLPPAADLD DFSKQLQQSM SSS ADST QA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LINK-ΔDimer</td>
<td>VTLLPAADLDDFSKQLQQSMSSADST QA</td>
<td>CTD FL</td>
<td>GPMSDNCPQQRNPQRNPRTFFGGPSDST GSNQNGERSGARSKQQRPQQLPNNTA SWFTALTQHGKEDLKFPFRQGGVPINT NSSPDDQIYGYYRRATRIRRGGDGKMK DLSPRWYFYYLGTGPEAGLPGYANKE GIIWVATEGANLPKDHIGTRNPANNA AIVLQLPQGTTLPKGFCAEGRGGSQA SSSSSRSRNSRNSNTPSPGSRTSPARM AGNGGDAALALLLLDRLNQLKLESKMS GKGQQQQGQC 1</td>
<td>247</td>
</tr>
<tr>
<td>CTD</td>
<td>GPMSDNCPQQRNPQRNPRTFFGGPSDST GSNQNGERSGARSKQQRPQQLPNNTA SWFTALTQHGKEDLKFPFRQGGVPINT NSSPDDQIYGYYRRATRIRRGGDGKMK DLSPRWYFYYLGTGPEAGLPGYANKE GIIWVATEGANLPKDHIGTRNPANNA AIVLQLPQGTTLPKGFCAEGRGGSQA SSSSSRSRNSRNSNTPSPGSRTSPARM AGNGGDAALALLLLDRLNQLKLESKMS GKGQQQQGQC 363</td>
<td>419</td>
<td>F363C-A419C</td>
<td></td>
</tr>
</tbody>
</table>
Table S3.2. Fit parameters to denaturant binding models.

<table>
<thead>
<tr>
<th></th>
<th>$\rho$</th>
<th>$K$ (M$^{-1}$)</th>
<th>$R_0$ (Å)</th>
<th>$m$ (kcal mol$^{-1}$ M$^{-1}$)</th>
<th>$c_{1/2}$ (M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTD-FL (2 state)</td>
<td>1.4 ± 0.3</td>
<td>0.34 ± 0.04</td>
<td>47 ± 2 (fixed $r_{0u}$)</td>
<td>4.0 ± 0.8</td>
<td>1.3 ± 0.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>34 ± 3 ($r_{0u}$)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NTD-RBD</td>
<td>1.5 ± 0.2</td>
<td>0.30 ± 0.03</td>
<td>46 ± 2 (fixed $r_{0u}$)</td>
<td>5.7 ± 1.5</td>
<td>1.50 ± 0.09</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>36 ± 2 ($r_{0u}$)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NTD-FL (3 state)</td>
<td>0.97 ± 0.09</td>
<td>0.34 ± 0.04</td>
<td>47 ± 2 (fixed $r_{0u}$)</td>
<td>Fixed based on RBD fit</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>39.8 ± 0.9 ($r_{0u}$)</td>
<td>Fixed based on RBD fit</td>
<td></td>
</tr>
<tr>
<td>NTD-RBD (3 state)</td>
<td>1.28 ± 0.09</td>
<td>0.32 ± 0.03</td>
<td>47 ± 2 (fixed $r_{0u}$)</td>
<td>Fixed based on RBD fit</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>38.3 ± 0.6 ($r_{0u}$)</td>
<td>Fixed based on RBD fit</td>
<td></td>
</tr>
<tr>
<td>RBD-FL</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>6.6 ± 0.5 (m$^{-1}$)</td>
<td>1.68 ± 0.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(c$^{-1/2}$ U-I)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8.1 ± 0.5 (m$^{1/2}$ F)</td>
<td>1.64 ± 0.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(c$^{1/2}$ f F)</td>
</tr>
<tr>
<td>LINK-FL</td>
<td>0.9 ± 0.3</td>
<td>0.09 ± 0.04</td>
<td>55 ± 2 (fixed)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LINK-ADimer</td>
<td>0.98 ± 0.07</td>
<td>0.64 ± 0.14</td>
<td>42 ± 2 (fixed)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>0.18 ± 0.04</td>
<td>49.0 ± 1.4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CTD-FL</td>
<td>0.54 ± 0.05</td>
<td>0.23 ± 0.04</td>
<td>51 ± 2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CTD</td>
<td>0.62 ± 0.07</td>
<td>0.42 ± 0.11</td>
<td>47.1 ± 1.4 (fixed)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table S3.3. Fit parameters of Higgs & Joanny theory

<table>
<thead>
<tr>
<th></th>
<th>v</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTD-FL</td>
<td>3.47 ± 0.05</td>
</tr>
<tr>
<td>LINK-FL</td>
<td>4.2 ± 0.2</td>
</tr>
<tr>
<td>CTD-FL</td>
<td>7.5 ± 0.5</td>
</tr>
</tbody>
</table>

Table S3.4. Scaling exponents

<table>
<thead>
<tr>
<th></th>
<th>( v_{\text{SAW}} )</th>
<th>( v_{\text{simulation}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTD-FL</td>
<td>0.510 ± 0.009</td>
<td>0.52</td>
</tr>
<tr>
<td>NTD-RBD</td>
<td>0.500 ± 0.009</td>
<td></td>
</tr>
<tr>
<td>LINK-FL</td>
<td>0.530 ± 0.008</td>
<td>0.58</td>
</tr>
<tr>
<td>LINK-( \Delta ) DIMER</td>
<td>0.464 ± 0.009</td>
<td></td>
</tr>
<tr>
<td>CTD-FL</td>
<td>0.542 ± 0.007</td>
<td>0.49</td>
</tr>
<tr>
<td>CTD</td>
<td>0.534 ± 0.008</td>
<td></td>
</tr>
</tbody>
</table>

Table S3.5. All-atom simulation summary

<table>
<thead>
<tr>
<th>System</th>
<th>No. sims</th>
<th>Total steps per sim (M)</th>
<th>Prod. steps per sim.(M)</th>
<th>Config. output</th>
<th>Ensemble size</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTD-RBD</td>
<td>400</td>
<td>24</td>
<td>20</td>
<td>20,000</td>
<td>399,000</td>
</tr>
<tr>
<td>RBD-LINK-DIM</td>
<td>29</td>
<td>66</td>
<td>60</td>
<td>20,000</td>
<td>82,113</td>
</tr>
<tr>
<td>DIM-CTD</td>
<td>200</td>
<td>24</td>
<td>20</td>
<td>20,000</td>
<td>200,000</td>
</tr>
<tr>
<td></td>
<td>NTD</td>
<td>LINK</td>
<td>CTD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>30</td>
<td>40</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>71</td>
<td>101</td>
<td>71</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>66</td>
<td>80</td>
<td>66</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>30,000</td>
<td>30,000</td>
<td>30,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>64,000</td>
<td>66,660</td>
<td>64,000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table S3.6. Interdye distances**

<table>
<thead>
<tr>
<th></th>
<th>( R_{\text{Gauss}} ) (Å)</th>
<th>( R_{\text{SAW}} ) (Å)</th>
<th>( R_{\text{simulation}} ) (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTD-FL</td>
<td>48 ± 2</td>
<td>47 ± 2</td>
<td>53</td>
</tr>
<tr>
<td>NTD-RBD</td>
<td>46 ± 2</td>
<td>45 ± 2</td>
<td></td>
</tr>
<tr>
<td>LINK-FL low E</td>
<td>55 ± 2</td>
<td>54 ± 2</td>
<td>59</td>
</tr>
<tr>
<td>LINK-FL high E</td>
<td>42 ± 2</td>
<td>45 ± 2</td>
<td>59</td>
</tr>
<tr>
<td>LINK-ΔDIMER</td>
<td>40 ± 2</td>
<td>40 ± 2</td>
<td></td>
</tr>
<tr>
<td>CTD-FL</td>
<td>51 ± 2</td>
<td>48.7 ± 1.4</td>
<td>46</td>
</tr>
<tr>
<td>CTD</td>
<td>49 ± 2</td>
<td>47.1 ± 1.4</td>
<td></td>
</tr>
</tbody>
</table>

**Table S3.7. Sequences used in simulations**

<table>
<thead>
<tr>
<th>Name (start – end)</th>
<th>Sequence (IDRs shown in bold and red)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTD-RBD (1 – 173)</td>
<td>MSDNGPQNQRNAPRITFGGPDSTGSNQNGERSGARSKQRPPQGLPNNTASWFTALTQHGKEDLKFPQGVPINTNSSPDQIQY YRRAATRIRRGDGGKMKDLSPRWYFYLYLGTGPEAGLPYGANKDGIIWVATEGALNTPDHGTRNPANNAIIVLQLPQGTTLPKGFYA</td>
</tr>
<tr>
<td>RBD-linker-DIM (50 – 366)</td>
<td>ASWFTALTQHGKEDLKFPQGVPINTNSSPDQIQY YRRAATRIRRGDGGKMKDLSPRWYFYLYLGTGPEAGLPYGANKDGIIWVATEGALNTPDHGTRNPANNAIIVLQLPQGTTLPKGFYAEGSRGGSQASSRSSRSRSNNSSNSTPSSRGTSPARMAHGGDAALALLLDRLNQLESKMSGKQQQGQTTKKSAAEASKPRQKRATTKQYNTQAFGRGPEQTQGNFDQDLDLRQGTDYKHWPQIAQFAPSA SAGGMSRIGEVTPSWTLTYHGAIKLDKDPQFKDNVILNNK HIDAYKTFPPT</td>
</tr>
<tr>
<td>DIM-CTD</td>
<td>TKKSAAEASKPRQKRATTKAYNTQAFGRGPEQTQGNFDQEQE</td>
</tr>
</tbody>
</table>

139
<table>
<thead>
<tr>
<th>Region</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>(247 – 419)</td>
<td>LIRQGTDYKHWPQIAQFAPSAAFFGMSRIGMEVTPSGTWLTYGAIKLDDKDPNFKDQVIIKNKHIDAYKTFPPTEPKKDKKKKADETQALPQRQKKQQTCTLPAADLDDFSKQLQQSMSSADSTQA</td>
</tr>
<tr>
<td>NTD (1 – 49)</td>
<td>MSDNGPQONQRNAPRITFGGPDSTGNSAQNGERSGARSKQRRPQGLPNNT</td>
</tr>
<tr>
<td>Linker (175 – 245)</td>
<td>EGSRGGSQASSRSSRRSRSRNSTPGSSRGTSPARMAANGGGDAALALLLDLRNLQLESKMSGKGQQQGQT</td>
</tr>
<tr>
<td>CTD (365 – 419)</td>
<td>PTEPKKDKKKKADETQALPQRQKKQQTCTLPAADLDDFSKQLQQSMSSADSTQA</td>
</tr>
</tbody>
</table>

### 3.9 Supporting Figures

![NDT-IDR](image)

**Fig. S3.1.** Sequence alignment of the coronavirus N-terminal domain (NTD)
Fig. S3.2. Sequence alignment of the coronavirus RNA binding domain (RBD)

Fig. S3.3. Sequence alignment of the coronavirus linker (LINK)
Fig. S3.4. Sequence alignment of the coronavirus dimerization domain

Fig. S3.5. Sequence alignment of the coronavirus C-terminal domain (CTD)
Fig. S3.6. 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. S3.7. 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. 3.2A, 3.3A, 3.4A). 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. S3.12). Data in panels C and E are presented as mean ± standard deviation.
Fig. S3.8. 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. S3.9. 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. S3.8) 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. S3.8) 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. S3.7 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. S3.10. 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. S3.11. 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. S3.11-12) 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. S3.11-12) 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. S3.11-12) 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 mean 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. S3.12. 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). Addition of GdmCl (e.g., 0.26 M) causes a decrease in the transfer efficiency distribution width 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. S3.13. 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. S3.14. 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. S3.15. 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 \cite{146,177}).
Fig. S3.16. 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. S3.17. Distributions for the radius of gyration ($R_g$) of 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).
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. S3.19. 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 \[^{178}\]. 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 \[^{24,26,43}\]. 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.
3.7 References


51. Soranno, A., Zosel, F. & Hofmann, H. Internal friction in an intrinsically disordered


64. 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


78. Posey, A. E., Holehouse, A. S. & Pappu, R. V. Chapter One - Phase Separation of


92. 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


171. Schäfer, L. *Excluded Volume Effects in Polymer Solutions: as Explained by the...*


Chapter 4.

The dimerization domain of SARS CoV 2 Nucleocapsid protein is partially disordered as a monomer and forms a high affinity dynamic complex.

This chapter is adapted from:

Jasmine Cubuk, Debjit Roy, J. Jeremias Incicco, Kathleen B. Hall, Alex S. Holehouse, Melissa D. Stuchell-Brereton, Andrea Soranno, In submission with Biophysical Reports

Author Contributions. J.C. performed and analyzed experiments. DR and JJI provided research tools. KH, MDSB, and AS supervised the work. JC, MDSB, and AS designed the experiments. JC, KH, ASH, MDSB, and AS have written the manuscript.
4.1 Abstract

The SARS-CoV-2 Nucleocapsid (N) is a 419 amino acids protein that drives the compaction and packaging of the viral genome. This compaction is aided not only by protein:RNA interactions, but also by protein:protein interactions that contribute to increasing the valence of the nucleocapsid protein. Here, we focused on quantifying the mechanisms that lead to dimer formation. Single-molecule Förster Resonance Energy Transfer enabled investigating the conformations of the dimerization domain in the context of the full-length protein as well as the corresponding energetics. Under monomeric conditions, we observed significantly expanded configurations of the dimerization domain (compared to the folded dimer structure), which are consistent with a dynamic conformational ensemble. Addition of unlabeled protein stabilized a folded dimer configuration with a high mean transfer efficiency, in agreement with predictions based on known structures. Dimerization has a high affinity of approximately 12 nM and is driven by enthalpic interactions between the two protein subunits. Interestingly, the dimer structure retains some of the conformational heterogeneity of the monomeric units and addition of denaturant reveals that the dimer domain can significantly expand before being completely destabilized. Our findings suggest that the inherent flexibility of the monomer form is required to adopt the specific fold of the dimer domain, where the two subunits interlock one with each other. We proposed that the retained flexibility of the dimer form may favor the capture and interactions with RNA and its temperature dependence may explain some previous observations for the phase separation propensity of the N protein.
4.2 Introduction

The SARS-CoV-2 Nucleocapsid (N) protein is responsible for the packaging of the 30 kbases-long viral RNA into small viral particles of about 100 nm\(^1\). The N protein is a flexible and dynamic protein with two structured domains, the RNA binding domain (RBD) and the dimerization domain, flanked by three disordered regions, the N- and C-terminal tails and the linker connecting the RBD and the dimerization domain\(^2\). Previous investigation of the SARS-CoV N protein pointed to oligomerization as a scaffolding mechanism that favors a dense organization of the RNA genome\(^3-5\). While oligomerization appears to be key to the function of SARS-CoV-2 N, little is known about the mechanism regulating the assembly of oligomers and how oligomerization impacts the flexible regions of the protein, starting from the dimer assembly. Recent Analytical UltraCentrifugation (AUC) experiments have revealed formation of stable dimers at concentrations as low as tens of nanomolar\(^6\). However, AUC measurements at low nanomolar concentration are challenging, providing only a rough estimate of the dimerization constant. Furthermore, the folded structure of the dimer reveals an N- and a C-terminal 3\(_{10}\) helix that enclose a series of short helical segments and two larger beta-strands interspersed by disordered regions\(^7\) (Fig. 4.1). A complex interface arises from the intertwining of the two molecules and pairing of the two beta-sheets from each protein subunit\(^3\). No structural details are known regarding the conformational changes needed in the monomeric structure for the formation of such a stable dimer. The limited interface and number of contacts within the dimer suggests that the same structure is most likely not stable in the monomer. Therefore, the dimer must form through some degree of unfolding (unless the region is already disordered) and refolding of the monomeric unit.
Here, we set out to investigate the mechanism of assembly of the dimer and its impact on the other domains of the protein using confocal single-molecule Förster Resonance Energy Transfer (FRET). This approach enables us to monitor the conformations of the full-length protein at sufficiently low concentrations (≤ 100 pM) to access the monomeric form and to characterize the thermodynamics of dimer formation and the associated conformational changes across the full-length protein. We complemented our observations with all-atom simulations that corroborate the experimental findings and provide atomistic details of the system.

4.3 METHODS

4.3.1 Single-molecule FRET. The full-length protein variants containing cysteine residues for labeling have been expressed in E.coli, purified, and sequentially labeled with Alexa Fluor 488 and 594. Single-molecule measurements have been performed using a modified MT200 instrument (Picoquant, Germany). For all measurements, unless otherwise stated, we used 100 pM of labeled protein (as estimated from serial dilution of a sample of known concentration, based on absorption at 280 nm). Pulsed Interleaved Excitation\(^8,^9\) (PIE) has been used to distinguish the donor-acceptor labeled molecules from donor-donor and acceptor-acceptor species. All measured data were analyzed using the Mathematica package Fretica (https://schuler.bioc.uzh.ch/programs/) to extract mean transfer efficiencies, lifetimes, and labeling stoichiometry ratio, as previously described\(^12,^10,^11\).

Extended Methods are described in the Supplementary Information.
4.4 RESULTS

To enable the study of N protein dimerization and associated conformational changes, we have designed a full-length construct that contains cysteine residues in position 245 and 363 (Fig. 4.1) These two residues are expected to be in close proximity upon dimerization based on the known crystal structure³. Simulations of transfer efficiency distribution using an AlphaFold¹²,¹³ prediction of the dimer (the chosen labeling positions extend beyond the residues in the available PDB structures) and FRETraj¹⁴ (to account for photon statistics) provide an estimate of mean transfer efficiency for this dye pair of approximately 0.90 ± 0.06 (Fig. S4.1). Such close proximity is due to the specific fold of the dimer domain and we expect the monomer form to adopt more expanded configurations. We will refer to this construct as DDₚ₇.₉₈.
Figure 4.1. Domain architecture of SARS-CoV-2 N protein and structural features of the dimerization domain. A. The SARS CoV 2 N protein contains 5 distinct domains: the N-terminal Domain (NTD), the RNA Binding Domain (RBD), a linker domain (LINK), the Dimerization Domain (DD), and the C-terminal Domain (CTD). B. Structure of the Dimerization Domain in dimer form (PDB: 6WZO) with the two units highlighted in distinct colors: teal and black. C. 2D-topology of the Dimerization Domain when in a stable complex. Letters indicate the main alpha-helices, $\eta_1$ and $\eta_2$ indicate $\beta_{10}$ helices, and $\beta_1$ and $\beta_2$ represent beta strands.

4.4.1 The Dimerization Domain is partially disordered in the monomer form. As a first step, we investigate the conformations of the DD$_{FL}$ at a concentration of 100 pM. To confirm the protein is monomeric at this concentration, we mixed equimolar concentrations of two single-labeled (F363C) protein preparations that were labeled with either donor or acceptor
and observed no appearance of a population with 1:1 donor:acceptor stoichiometry at 100 pM total concentration (50 pM of each species). Under the buffer conditions used in this work (50 mM Tris, pH 7.4, 150 mM KCl), contributions from dimer formation in the stoichiometric plot are observed only above 600 pM of total protein concentration (Fig. S4.1). After confirming that the protein is monomeric, we investigated the conformations of DD\textsubscript{FL} in its monomer form. The histogram of transfer efficiencies reveals a peak centered at 0.567 ± 0.005, which is clearly at variance with the value of ~ 0.9 expected from the configuration in the folded structure of the dimer (Fig. S4.1).

To further investigate the existence of stable configurations, we tested the effects of denaturant on the protein and we observed that addition of Guanidinium Chloride (GdmCl) leads to a constant shift of the mean transfer efficiency toward lower values, as expected for a protein region that is at least partially (if not completely) unstructured (Fig. 4.2). This is also confirmed from the investigation of the dependence of the lifetime and transfer efficiency, which provides evidence of dynamics on the nanosecond-microsecond timescale and no significant restrictions to the sampled distance distributions (Fig. S4.2). Therefore, we interpret our results assuming the interdye distance is described by a Gaussian chain distribution with a root-mean-square interdye distance of 5.70 ± 0.02 nm. Using the semi-empirical SAW-ν distribution recently proposed by Zheng et al., we obtain a similar root-mean-square distance, 5.4 ± 0.2 nm and a scaling exponent ν of 0.48 ± 0.09. We want to emphasize that this does not imply the lack of structure, but that compensatory effects between local structure formation and chain dynamics can give rise to similar statistics of an equivalent completely disordered chain.
**Figure 4.2. Dimer domain conformations.** A. Representative histograms of transfer efficiencies for the DD1L construct at 100 pM in aqueous buffer conditions (50 mM Tris, 150 mM KCl) and with increasing denaturant concentrations. B. Representative stoichiometry ratio vs. transfer efficiency plots for the corresponding histogram of transfer efficiencies. Stoichiometry ratio of 0.5 indicates 1:1 labeled material, indicating monomeric sample under these conditions. C. Distribution of transfer efficiencies as a function of denaturant concentration. D. Root-mean-squared interdye distances extracted using a Gaussian chain distribution as a function of denaturant concentration. In panels C and D, the line represents a fit to the model in Eq. S4.7, which accounts for denaturant binding.

### 4.4.2 Dimer formation induces folding of the DD domain.

We then proceeded to evaluate the protein conformational changes occurring upon dimerization. By titrating increasing concentration of unlabeled protein we observed the stabilization of a second population at higher transfer efficiency, whose mean transfer efficiency (0.84 ± 0.02) is in good agreement with the expected value based on the folded structure (Fig. 4.3A and S4.1). We then used singular value decomposition (SVD) to quantify the variation of the signal upon binding\textsuperscript{16–18}. The advantage of SVD is that it provides a model free tool to interpret the data without requiring a complex assignment of the distribution of transfer efficiencies to specific conformations of the monomer.
and dimer. To this end, the measured signal is represented by a matrix $\mathbf{H}$, where each line in the matrix refers to a histogram collected at distinct concentrations of unlabeled protein. Singular value decomposition of the matrix $\mathbf{H}$ is given by

$$\mathbf{H} = \mathbf{U} \mathbf{S} \mathbf{V}^T$$  \hspace{1cm} \text{Eq. 4.1}

where $\mathbf{S}$ is the diagonal matrix of the singular values, $\mathbf{U}$ is the orthonormal matrix of the transfer efficiency distributions of each singular value, and $\mathbf{V}^T$ is a matrix containing the amplitude information associated with each concentration of unlabeled protein. SVD can be used to distinguish the contribution of signal and signal changes from experimental noise by comparing the amplitude of the singular values. As shown in Fig. 4.3B, two major singular values are identified in this titration, whereas all the others contribute to a significantly lesser extent to the total signal. Inspection of the amplitudes further reveals a sigmoidal trend on a logarithmic scale of the concentrations, which reflects the dimerization isotherm of the protein as associated to changes in the first and second singular vector (Fig. 4.3D). A global fit of the amplitude curves to a binding model that accounts for dimerization of labeled and unlabeled species (Supplementary Information) results in a value of the dimerization dissociation constant of $6 \pm 2$ nM.
**Figure 4.3. Stabilization of the dimer conformations upon binding: conformational changes and dimerization affinity.** A. Normalized histograms of transfer efficiencies from 100 pM labeled N protein to a total concentration of 5 μM N protein. B. Singular values from the SVD analysis reveal two relevant components. C. Basis vectors for each significant singular value. Representative basis vector for component one (top) and representative basis vector for component two (middle). D. Reconstruction of histograms using the singular value (S), basis vector (U), and amplitudes (V) (bottom). E. Binding isotherms of dimerization of N protein from the SVD analysis of the amplitudes associated with the first (top) and second (bottom) components. Lines represent fit to Eqs. S4.9 and S4.10, which accounts for the binding of the unlabeled N protein to the labeled N protein in solution.

### 4.4.3 Denaturant effect on dimer stability

We further investigate the stability of the dimer structure in the presence of GdmCl. We chose conditions such that DD_{FL} is complexed in a stable dimer configuration (in presence of 1 μM unlabeled protein) and then added increasing GdmCl concentrations (up to 1.5 M GdmCl, where the complex is completely destabilized) (Fig. 4.4). We found that with increasing concentration of denaturant (from 0 to 1.3 M GdmCl), the protein adopts more expanded conformations in both the dimer and monomer conformations, with the mean transfer efficiencies of the two states shifting from 0.84 ± 0.02 to 0.70 ± 0.01 (dimer conformation) and 0.567 ± 0.005 to 0.292 ± 0.005 (monomer conformation). Contrasting
these observations with the corresponding lifetime information, we confirmed that the dimer population in the transfer efficiency distribution also represents a dynamic conformational ensemble (Fig. S4.2). The corresponding mean interdye distances are reported in Fig. 4.4B and S4.4. In aqueous buffer conditions, we quantified the distance to be 3.8 ± 0.1 nm, according to a Gaussian chain distribution, or 3.7 ± 0.2 nm, according to the SAW-ν model. The scaling exponent for the SAW-ν model is reduced to 0.40 ± 0.02, as implied by the more compact configuration.

The comparison between lifetime and transfer efficiency is less precise at high transfer efficiencies, making it more difficult to distinguish effects of chain dynamics from linker dynamics. However, for a completely folded protein, the mean transfer efficiency is expected to remain stable over GdmCl concentration, allowing for changes due to the refractive index. The expansion of the dimer folded conformation with increasing denaturant concentration (which exceeds the shifts due to refractive index changes and is computed by accounting for such effects) suggests that the observed chain dynamics are not simply the result of dye-linker dynamics. Instead, a certain flexibility must be encoded in the dimer structure and a certain destabilization of the structure is allowed without losing the stability of the dimer complex. We speculate that these conformational changes may occur along the N- and C-terminal portions of the dimerization domain, which can be destabilized without altering the interface between the intertwining beta sheets at the core of the dimer structure.

For the monomer, we observe a continuous expansion with increasing concentration of the denaturant, with the protein adopting a root-mean-square radius of 8.5 ± 0.2 nm at 1.3 M GdmCl, which is significantly more expanded than the configuration observed in absence of denaturant. This conformational change implies that large portions of the structured
conformations of the monomeric protein have been destabilized. In addition, the observation of the coexistence of these conformations with a structured dimer implies that, despite some structured conformations having been destabilized and their conformations being expanded, a stable dimer can still be formed. This is not unreasonable, since formation of the dimer clearly requires large conformational changes in the monomer to allow the two beta sheets to intertwine together.

Analysis of the dissociation constant reveals a linear trend with the denaturant concentration in the regime studied in our single-molecule experiments (Fig. 4.4D). It is interesting to note that the linear trend would suggest a higher fraction of dimer should still be present at 1.5 M GdmCl, but we do not observe such a population. We speculate that this represents a threshold concentration over which either the dimer domain is completely destabilized or any residual structure in the monomer required for dimer formation has been destabilized (based on equilibrium data we cannot distinguish the two case scenarios). To test this hypothesis we measured the fraction of dimer at 1.75 M GdmCl at a concentration of protein 10 μM and at 2 M GdmCl at the concentration of 45 μM. According to the linear fit we should observe a fraction bound of 0.5 and 0.5, respectively, as determined from the linear extrapolation of the $ln K_D$ versus GdmCl concentration. We found that no dimer complex is formed at any of these conditions, supporting our hypothesis that 1.5 M GdmCl is suppressing dimer formation by denaturing essential folded structures in either the monomer or the dimer.
Figure 4.4: Denaturant dependence of the dimerization domain. A. Representative histogram of transfer efficiencies for the DD\textsubscript{FL} construct in presence of 1 \(\mu\)M unlabeled protein in absence and with increasing concentrations of denaturant. B. Transfer Efficiencies as a function of denaturant for the dimer (darker gray) and monomer (lighter gray) distributions. Lines represent a fit to the Schellman weak binding model in Eq. S4.7. C. Root-mean-squared inter-dye distances as a function of denaturant for the dimer (darker gray) and monomer (lighter gray) distributions. Lines represent a fit to Eq. S4.7. D. Plot of fraction of dimer as a function of denaturant concentration. Here, dimers can be seen present up to 1.3M GdmCl. E. Plot of \(\ln(K_D)\), with \(K_D\) expressed in M units, as a function of denaturant. A linear trend is observed between the dissociation constant and concentration of GdmCl.

4.4.4 Temperature effect on dimer stability. We further investigated the temperature dependence of the dimer stability to quantify the enthalpic and entropic contributions at play and the corresponding conformational changes. To this end, we used a temperature-controlled cuvette, analogous to previous studies\textsuperscript{22–24}. To limit any contributions from the pH dependence on temperature, we performed all temperature measurements in the HEPES buffer. Similar to Tris buffer, the protein adopts a transfer efficiency of 0.57 ± 0.03 in the monomer configuration and of 0.85 ± 0.03 in the dimer configuration. We started by studying the temperature response of DD\textsubscript{FL} under monomeric conditions to establish a baseline for the conformational changes of this
region. We found that the domain adopts more compact conformations as we increase temperature, in line with previous temperature dependence experiments on disordered proteins\textsuperscript{22,23,25}. We then studied the effects of temperature on the domain when the protein is a stable dimer (1 μM). We observed that from 10 °C to 30 °C, the protein remains dimeric and the mean transfer efficiency associated with the conformations of the dimer species report a small shift toward higher values, indicating a small, but measurable compaction of the protein. Starting from 37 °C, a broadening of the distribution of transfer efficiencies is observed. This broader distribution of transfer efficiencies can be disentangled into two populations that represent the monomer and dimer species. Mean transfer efficiencies and corresponding areas of each subpopulation can be used to estimate the association constant at each temperature (Fig. 4.5).

When increasing temperature to 50 °C or higher, the protein is completely dissociated and adopts the conformations of the monomeric form.

To analyze the enthalpic and entropic contributions of dimerization, we performed a van’t Hoff analysis on the temperature dependance of the dissociation constant (Fig. 4.5E). According to the van’t Hoff equation, if in a given range of temperatures the enthalpy of a reaction does not change, the logarithm of the dissociation constant $K_D$ is a linear function of the reciprocal of absolute temperature $T$, where the slope reports about the enthalpy $\Delta H^0$ and the intercept reports about the entropy $\Delta S^0$ at standard conditions (1 M) of binding:

$$\ln K_D = \frac{\Delta H^0}{R} \frac{1}{T} - \frac{\Delta S^0}{R}$$

\textbf{Eq. 4.2a}

where $R$ is the gas constant.
We observed a linear decrease of $ln K_D$ as a function of $1/T$, in the range of temperatures from 37 °C to 47 °C. We found that the dimer formation is exothermic, with a $\Delta H^\circ$ of $-99 \pm 6$ kcal/mol and a $\Delta S^\circ$ of $0.29 \pm 0.02$ kcal/mol K.

However, when extending the plot to lower temperatures and including the $K_D$ determined at 23 °C, we clearly observe a deviation from linearity that suggests a non-negligible change in the heat capacity, and fitting with the non-linear form of the van’t Hoff equation:

$$ln K_D = \frac{\Delta H^\circ}{R} \frac{1}{T} - \frac{\Delta S^\circ}{R} + \Delta C_p (1 - \frac{T}{T_r})$$  \hspace{1cm} \text{Eq. 4.2b}$$

where $\Delta H^\circ$ and $\Delta S^\circ$ are the enthalpy and entropy at the reference temperature, $\Delta C_p$ is the heat capacity, $T$ is temperature, and $T_r$ is the reference temperature. Corresponding values of the fit are reported in Supplementary Fig. 4.4 and Table 4.5. Analysis of dimer fractions as estimated from the extrapolation at lower temperatures with Eq. 4.2a and Eq. 4.2b reveals a small discrepancy with respect to the experimentally determined fractions at lower temperatures. We attribute this discrepancy to the contribution of the effect of temperature on the conformations of the dimer domain, which may result in a variation of enthalpic and entropic contributions.
Figure 4.5: Temperature dependence of the dimerization domain. A. Normalized histograms of transfer efficiencies for the DD$_{1}$L construct at 100pM (top) and in presence of 1 μM unlabeled protein (bottom), ranging from 10 °C to 63 °C. B. Corresponding 2D Temperature vs transfer efficiency plots for 100pM (top) and with the addition of 1 μM unlabeled protein (bottom). C. Dependence of transfer efficiency on temperature for 100 pM total protein concentration (gray) compared to 1 μM total protein concentration (black). D. Dependence of root mean-squared interdye distance on temperature for 100pM total protein concentration (gray) compared with 1 μM total protein concentration (black). E. Fraction of labeled protein in dimers (formed with unlabeled protein) as a function of temperature. The line is computed assuming the linear dependence of ln(K$_D$) as a function of 1/T (K$^{-1}$) shown in panel F, according to Eqs. S4.9-11. F. Plot of the ln(K$_D$), with K$_D$ expressed in M units, as a function of 1/T (K$^{-1}$). The linear fit reports about the enthalpy (slope) and entropy (intercept) of dimer dissociation according to Eq. 4.2a. Values obtained from the fitting procedures are reported in Supplementary Table 4.5.

4.4.5 Does dimerization affect other domains? Finally, we tested whether formation of the dimer complex induces conformational changes in the other N protein domains (Fig. S4.5). To this end, we compared the conformations of the N-terminal domain (NTD), Linker, and C-terminal domain (CTD), with and without saturating concentrations of unlabeled proteins. At 150 mM KCl, the conformations of each region of the protein matches our previous observations, where all the three populations exhibit a narrow distribution of transfer efficiencies. Upon addition of 1 μM of unlabeled protein, we observed no significant shift in the NTD (with mean transfer efficiency shifting from 0.60 ± 0.01 to 0.595 ± 0.003 ) or RBD (with mean transfer efficiency shifting from 0.818 ± 0.003 to 0.828 ± 0.004 ). We detected a small
expansion of the LINK (with mean transfer efficiency shifting from 0.570 ± 0.003 to 0.520 ± 0.004) and a significant expansion of the CTD (mean transfer efficiency shifting from 0.64 ± 0.02 to 0.570 ± 0.001). The conformational changes of the Linker and of CTD are consistent with the suppression of previously identified interactions between the DD domain and surrounding disordered regions.

4.5 Discussion

Our experiments and simulations provide new insights on the conformational properties of the dimerization domain in the context of the full-length protein, both in its monomer and dimer forms, and on the energetics associated with the complex formation.

An approximate estimate of the dissociation constant was previously determined using Analytical Ultracentrifugation, where appearance of the monomer population was observed between 30 nM and 1 nM, resulting in an approximate \( K_D \) of roughly 30 nM for unlabeled protein and 2 nM for labeled protein. Our single molecule approach enabled us to follow the transition from monomer to dimer with a fine-tuned titration, starting from 100 pM (labeled concentration) and increasing concentration up to 10 \( \mu \)M (using unlabeled protein). We modeled the data accounting for the formation of dimers between labeled and unlabeled proteins (labeled-labeled, unlabeled-unlabeled, and labeled-unlabeled) and corresponding impact on the association constants (Supplementary Information). Following this procedure, our experiments yield a dissociation constant between labeled and unlabeled molecules \( K_{D,L/U} \) of 6 ± 2 nM and a dissociation constant between unlabeled (or labeled) molecules \( K_{D,U/U} \) of 11 ± 3 nM. Our results are in reasonable agreement with the previous estimates reported by the Schuck.
group\textsuperscript{37}. The good agreement between the methods supports that we are capturing the same phenomenon, besides small differences that can be introduced by solution conditions and specific labeling positions of the protein.

At the same time, our experiments enabled access to conformational changes in the monomer and dimer forms of N protein. We found that in the monomer form, the dimerization domain of the N protein exhibits a significant degree of flexibility and is very sensitive to solution conditions (as probed by denaturant and temperature). We reason that this flexibility is required for facilitating the formation of the specific fold of the dimer domain, where beta sheets belonging to the two proteins are required to intertwine one with each other. The dimer form, while much more compact, also exhibits a dynamic behavior and can be modulated by solution conditions. In particular, denaturants can significantly expand the conformations of the complex.

We speculate that the high stability of the complex and its inherent flexibility can be harnessed by the nucleocapsid for increasing the ability of the nucleocapsid protein to recruit and trap the viral RNA. Indeed, a flexible chain ensures a larger captured radius than a conventional folded domain\textsuperscript{28}. At the same time, dimerization ensures a higher valence of interaction, which can increase the effective affinity for the nucleic acid. We note that our results are likely to apply to a series of analogous dimerization domains in other coronaviruses\textsuperscript{7,29} (SARS-CoV\textsuperscript{4}, hCoV-NL63\textsuperscript{30}, MERS-CoV\textsuperscript{31}, MHV-A59\textsuperscript{32}, IBV\textsuperscript{33}), since they share the same fold. While similar with respect to the structure adopted by the dimer domain, the sequence composition differs across viruses and may modulate thermal stability, affinity, as well as the conformational heterogeneity of monomeric and dimeric forms. Understanding how the sequence modulates the energetics and conformational properties of the dimer may provide insights on future emerging coronavirus and help designing drugs that target dimer stability (limiting the valence of the N
protein) or its conformational flexibility (limiting the capture radius and adaptable interface for binding RNA).

Our experiments also provide a possible molecular mechanism of the lower critical solution temperature (LCST) previously identified in phase separation experiments with the N protein. While the exact critical temperature was not identified, N protein was shown to undergo phase separation in isolation starting from a temperature of 45 °C at protein concentrations between 1 and 4 μM. This range of temperature coincides with our temperature-dependent suppression of protein dimerization at the same protein concentration and is consistent with differential scanning fluorimetry measurements. We reason that unfolding of the dimer domain and the large flexibility and disorder of the monomer can augment the multivalence of N protein and favor its phase separation in absence of RNA.

Interaction with RNA complicates this scenario, introducing a new set of interactions between the protein and the nucleic acid (e.g. through the RNA binding domain), which can be realized through both disordered and structured domains. While it was suggested that binding of RNA could destabilize the dimer structure, we do not observe any alteration of the binding fraction when titrating non-specific single-stranded RNA as well as specific double-stranded RNA (Fig. S4.4). In light of our experiments, the expansion of the dimerization domain upon binding of RNA, as observed in cryo-EM measurements, may reflect the inherent flexibility of this dimer complex and its adaptability for favoring RNA binding. Future work is required to carefully evaluate how the multivalence of RNA interactions rewire the protein-RNA phase diagram, its temperature dependence, and how these interactions control condensation of single and multiple RNA chains.
4.6 Conclusions. Here, we have completed the characterization of the conformational properties of the N protein by investigating the structural ensemble of the dimerization domain. We have found that the dimerization domain forms a high affinity complex (starting from low nanomolar concentrations) that retains part of the flexibility intrinsic to the monomeric form. Our results pave the way to constructing quantitative models of the protein-protein and protein-RNA interactions at play when the N protein condenses viral genomic RNA or undergoes phase separation.

Acknowledgements. We thank Ben Schuler and Daniel Nettels for developing and maintaining the Fretica package used for the analysis of the single-molecule data. We thank Upasana L. Mallimadugula for her valuable input on figures.
4.7 Supporting Information.

4.7.1 Protein expression, purification, and labeling. GST-His$_9$-SARS-CoV2 WT (cysless), NTD-FL, RBD-FL, LINK-FL, and CTD-FL were expressed and purified as previously described$^2$.

GST-His$_9$-SARS-CoV2 DD$_{FL}$ Nucleocapsid protein was expressed recombinantly in Gold BL21(DE3) cells (Agilent). 6 L cultures were grown in LB medium with carbenicillin (100 µg/mL) to OD$_{600}$ ~ 0.8 and induced with 0.25 mM IPTG for 3 hours at 37°C. Harvested cells were lysed with sonication at 4°C 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 1.5 hr) and the pellet was resuspended in 50 mM Tris pH 8, 300 mM NaCl, 10% glycerol, 6 M Urea, 5 mM βME and incubated at 4°C 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, 300 mM NaCl, 10% glycerol, 20 mM imidazole, 5 mM βME) containing 6 M Urea. The column was then washed with buffer A allowing the protein to refold on the column. The column was then washed with high salt buffer (Buffer A +2M NaCl), followed by Buffer A. 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 βME, buffer B: buffer A + 1 M NaCl) over 120 min. Purified N protein was
analyzed using SDS-PAGE. Protein concentrations of stock solutions were determined spectroscopically in 50 mM Tris (pH 8.0), 200-500 mM NaCl, 10% (v/v) glycerol using extinction coefficients of 42530 M⁻¹ cm⁻¹(FL).

All Nucleocapsid variants were labeled with Alexa Fluor 488 maleimide (Molecular Probes) under denaturing conditions in buffer A (50 mM Tris pH 8, 6M 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(+5mM BME) 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 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).

4.7.2 Experimental setup and procedure for single-molecule fluorescence experiments. Single-molecule experiments were conducted on a Picoquant MT200 instrument (Picoquant, Germany). Pulsed Interleaved Excitation (PIE) was realized by alternating the pulses from 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). The excitation rate was set to 20 MHz such that a delay of approximately 25 ns occurs between each of the two laser pulses.
Laser beams were focused through a 60x1.2 UPlanSApo Superapochromat water immersion objective (Olympus, Japan) and emitted photons were collected through the same objective. Emitted photons were further filtered through a dichroic mirror (ZT568rpc, Chroma, USA), a long pass filter (HQ500LP, Chroma Technology), and the confocal pinhole (100 μm diameter) before being selected according to polarization and emission wavelength. Donor and acceptor photons were first separated by a dichroic mirror (585DCXR, Chroma) and further refined by using band pass filters, ET525/50m or HQ642/80m (Chroma Technology), respectively. Finally, the filtered photons were focused on SPAD detectors (Excelitas, USA) and their arrival time was recorded with a HydraHarp 400 TCSPC module (PicoQuant, Germany). All FRET experiments used a donor excitation with a laser power of 100 μW (measured at the back aperture of the objective) and an acceptor excitation adjusted to match the total emission intensity after donor excitation (between 50 and 70 μW).

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. Measurements were performed in uncoated polymer coverslip cuvettes (Ibidi, Wisconsin, USA) and pegylated coverslips (for temperature dependence). 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).

4.7.3 Construction of transfer efficiency histograms. Photons were time-binned in bins of 1 ms. Contiguous bins with at least one above 15 photons were merged into a single burst and bursts with at least 20 photons were selected for further analysis. 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

$$E = n_A / (n_A + n_D)$$

Eq. S4.1

where $n_A$ and $n_D$ are the numbers of donor and acceptor photons, respectively.

Transfer efficiencies were corrected for background, acceptor direct excitation, channel crosstalk, differences in detector efficiencies, and quantum yields of the dyes.

The labeling stoichiometry ratio $S$ was computed accordingly to:

$$S = I_D / (I_D + \gamma_{PIE} I_A)$$

Eq. S4.2

where $I_D$ and $I_A$ represent the total intensities observed after donor and acceptor excitation, $\gamma_{PIE}$ is a correction factor accounting for differences in donor and acceptor 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).

### 4.7.4 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.

4.7.5 Root-mean-square interdye distance. Conversion of mean transfer efficiencies to an interdye distance were performed after verifying fast rearrangement of the ensemble based on the comparison of lifetime vs transfer efficiency. Under this assumption, the conversion is obtained by integrating the distance dependence of the transfer efficiency

\[ E(R) = \frac{R_0^6}{R_0^6 + R^6} \quad \text{Eq. S4.3} \]

on the distance distribution according to:

\[ \bar{E} = \int_0^\infty E(R) P(R) \, dR \quad \text{Eq. S4.4} \]

where \( R_0 \) is the Förster radius.

For a Gaussian chain, the distribution is given by:

\[ P(R) = 4\pi R^2 \left( \frac{3}{2\pi r^2} \right)^{3/2} \exp\left( \frac{-3R^2}{2r^2} \right) \quad \text{Eq. S4.5} \]

where \( r \) represents the root-mean-square interdye distance of the protein and is the only fitting parameter.

For the SAW-\( \nu \) model can be expressed as:

\[ P_{SAW}(R, \nu) = A_1 \frac{4\pi}{b_0N^\nu} \left( \frac{R}{b_0N^\nu} \right)^{2+\nu} \exp\left[ -A_2 \left( \frac{R}{b_0N^\nu} \right)^\delta \right] \quad \text{Eq. S4.6} \]
where \( A_1 = \frac{6}{4\pi} \frac{\Gamma(5+g/\delta)}{\Gamma(3+g/\delta)} \frac{\Gamma(\delta)}{\Gamma(\delta)} \), \( A_2 = \left( \frac{\Gamma(5+g/\delta)}{\Gamma(3+g/\delta)} \right)^\frac{5}{3} \), \( g = \frac{(y-1)}{\nu} \delta = \frac{1}{(1-v)} \), \( \gamma = 1.1615 \), \( \Gamma \) is the Euler Gamma Function, \( b_0 = 0.55 \) nm is an empirical prefactor \(^{15}\), \( N \) is the number of residues between the fluorophores, and \( \nu \) is the scaling exponent. Here \( \nu \) is the only fitting parameter.

4.7.6 Binding of Denaturant. Similar to previous works\(^2,35\), we model expansion of the disordered chain with denaturant with an empirical model based on the denaturant dependence of the Schellman weak-binding model\(^{36}\):

\[
 r(c) = r_0 (1 + \rho \frac{Kc}{1+Kc})
\]

Eq. S4.7

where \( r_0 \) is the root mean square interdye distance in the absence of denaturant, \( \rho \) is a term that represents the chain expansion in the presence of denaturant compared to the absence \( (r_0) \), \( K \) is the binding constant and \( c \) is the concentration of denaturants. This semi-empirical model is used only for comparison between different titrations, without aiming to extract significant quantities. A more precise description would require the use of a coil-to-globule model, which provides access to the free energy change in the protein chain and can be directly related to the Schellman model\(^{37}\).

4.7.7 Fluorescence lifetime. Fluorescence lifetimes were estimated from bursts using a maximum likelihood method\(^{38,39}\) and implemented in the Fretica package. The dynamic regime was computed according to:

\[
 \tau_D/\tau_{D0} = 1 - \bar{E} + \frac{\sigma^2}{1-E}
\]

Eq. S4.8a
\[
\frac{(\tau_A - \tau_{A0})}{\tau_{D0}} = 1 - \bar{E} + \frac{\sigma^2}{E}
\]

Eq. S4.8b

where \(\tau_D\) and \(\tau_{D0}\) are the donor lifetimes in presence and absence of the acceptor, \(\tau_D\) and \(\tau_{D0}\) are the acceptor lifetimes in presence and absence of the donor, \(\bar{E}\) is the mean value of the transfer efficiency according to Eq. S4.4 and \(\sigma\) represents the corresponding variance, \(\sigma^2 = \int E^2(r)P(r)\,dr - \bar{E}^2\). In absence of fluctuations over a distribution of distance (i.e., \(\sigma^2 = 0\)), the equations reduces to the well known static regime equation where:

\[
\frac{\tau_D}{\tau_{D0}} = \frac{(\tau_A - \tau_{A0})}{\tau_{D0}} = 1 - E
\]

Eq. S4.8c

4.7.8 Modeling dimerization. Binding experiments have been analyzed accounting for the possibility of forming dimers of labeled molecules \(DD_L\) with other labeled molecules \((DD_L:DD_L)\), labeled molecules with other unlabeled molecules \(DD_U\) \((DD_L:DD_U)\), as well as between unlabeled molecules \((DD_U:DD_U)\). To this end, we imposed a mass balance such that:

\[
\begin{bmatrix}
[DD_L]^{tot}
\end{bmatrix} = \begin{bmatrix}
[DD_L] \\
[DD_U]
\end{bmatrix} + 2 \begin{bmatrix}
DD_L: DD_L \\
DD_L: DD_U
\end{bmatrix}
\]

Eq. S4.9a

\[
\begin{bmatrix}
[DD_U]^{tot}
\end{bmatrix} = \begin{bmatrix}
[DD_U] \\
[DD_L]
\end{bmatrix} + 2 \begin{bmatrix}
DD_U: DD_U \\
DD_L: DD_U
\end{bmatrix}
\]

Eq. S4.9b

For each dimer we can write the corresponding dissociation constant as:

\[
K^{LL}_{DL} = \frac{[DD_L]^2}{[DD_L:DD_L]}
\]

Eq. S4.10a
\[ K_{DU} = \frac{[DD_D]^2}{[DD_D:DD_D]} \quad \text{Eq. S4.10b} \]

\[ K_{DL} = \frac{[DD_L][DD_D]}{[DD_L:DD_D]} \quad \text{Eq. S4.10c} \]

Finally, we accounted for the fact that if expressed in the microscopic rate constants of binding and unbinding the \(K_{DU} = 1/2\) \(K_{DL} = 1/2\), under the assumptions that the rates of unlabeled and labeled are equal\(^{26}\).

Fraction of labeled protein that forms dimers with unlabeled protein (coinciding with the FRET population at 0.5 stoichiometry ratio) is given by:

\[ f_{\text{Dimer}}^{LU} = \frac{\left[DD_D\right]_{\text{tot}}}{2\left[DD_L\right]_{\text{tot}}^2 + \left[DD_D\right]_{\text{tot}}^2} \sqrt{K_{DU} + 4\left[DD_L\right]_{\text{tot}} + \left[DD_D\right]_{\text{tot}}^2} - \sqrt{K_{DL} + 4\left[DD_L\right]_{\text{tot}} + \left[DD_D\right]_{\text{tot}}^2} \quad \text{Eq. S4.11} \]
4.8 Supporting Figures

Supplementary Figure S4.1. Estimated transfer efficiency based on the dimer structure. A. Structure estimated using AlphaFold with labeling positions 245 and 363. Dye-clouds based on available excluded volume for the dyes computed using FRETraj. B. Corresponding distribution of transfer efficiencies generated using FRETTraj.
 Supplementary Figure S4.2. Labeling Stoichiometry Ratio histograms of DD_{FL} measured at 100 pM, 300 pM, and 600 pM total concentrations of nucleocapsid protein. Single-labeled F363C Alexa-488 was mixed with Single-labeled F363C Alexa-594 at equimolar concentrations. We do not observe any population at 0.5 stoichiometry under 600 pM, which would represent the formation of a dimer.
Supplementary Figure S4.3. Lifetime vs Transfer efficiency of the DD$_{yL}$ at 100pM (top) and with the addition of 1 μM unlabeled N protein (bottom). In both cases, the populations sit near the dynamic line (green for donor, red for acceptor), as opposed to falling on the static line (gray), indicating a rigid configuration.
**Supplementary Figure S4.4. Effects on solution conditions on dimerization.** A. Root mean-squared interdye distance for the DDFL as a function of denaturant. Grey points and fit represent the distance distributions for the DDFL at 100 pM total protein concentration (Fig 4.2). Teal points and fit represent the distance distributions for the DDFL at 1 μM total protein concentration (100 pM labeled + 1 μM unlabeled) (Fig. 4.4). B. Histogram of transfer efficiencies for 1.75 M GdmCl with the addition of 10 μM unlabeled (top) and 2 M GdmCl with the addition of 45 μM unlabeled (bottom). Extrapolation from the Van’t Hoff dependence of ln(K_D) predicts a fraction of dimer of ~0.5; however, no dimer conformation is observed in either scenario. C. Histogram of transfer efficiencies at 37 °C (100 pM labeled + 1 μM unlabeled) with the addition of high concentrations of specific double-stranded RNA (top, SL5B) and non-specific single-stranded RNA (bottom, (rU)10). D-F. Fit of the ln(K_D) as a function of the reciprocal of temperature (D) according to Eq. 4.2b (dashed line) compared to the fit of Eq. 4.2b to the fraction of the dimer population (E) (solid line). Overlap of the experimental trends and corresponding fit in panel F.
Supplementary Figure S4.5. Dimerization alters the conformations of the Linker domain and CTD, but not NTD or RBD. Transfer efficiency distribution for each labeled segment in the context of the full-length protein.
4.9 Supporting Tables

**Supplementary Table 4.1.** Sequence of Wildtype (WT) Nucleocapsid Protein. Labeling positions are reported as bold and underlined residues. Highlighted regions delineate folded domains.

<table>
<thead>
<tr>
<th>Start Position (WT)</th>
<th>End Position (WT)</th>
<th>Labeling Positions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 MSDNGPQNQR NAPRITFGGP SDSTGSNQNG ERSGARSKQR RPQGLPNNTA</td>
<td>419</td>
<td>M1C-R68C</td>
</tr>
</tbody>
</table>

**Supplementary Table 2.** N protein constructs

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTD&lt;sub&gt;FL&lt;/sub&gt;</td>
<td>GPCSDNGPQNQRNAPRITFGGPDSTG SNQNGERSGARSKQRORPQGLPNNTA WFTALTQHGKEDLKFPCRQGQGVPINTNS SPDDQIGYRRATRIRRGGDGKMKL SPRWYFYLYLGTEAGPLYGANKDGII WVATEGALNTPKDHIGTRNPANNAI VLQLPGQTTLPKGTYAEGRGGSQASS RSSRSRNRNSSNTPGSSRTGSPARMA GNGGDAALALLLLDLRLNQLESKMSG KQGQQQGQQTVTKKSAAEASKKRQPQK RTATKAYNTQAFGRGPEQTQGNGF DQELIRQGTQDHWPQIAQFPSASAF FGMSRIGMEVTPSTWLYTGAIKLD DKDPNFQDVILLNKHIDAYKTFPPTE PKKDKKKKADETQALPQKQKQQTIV TLLPAADLDDFSQNLQQQSMSSADSTQ</td>
</tr>
<tr>
<td>RBD&lt;sub&gt;FL&lt;/sub&gt;</td>
<td>GPMSDNGPQNQRNAPRITFGGPSDSTG SNQNGERSGARSKQRRPQGLPNNTAS WFTALTQHGKEDLKFPCGQGVQYGPINTNS SPDDQIQGYYRRATRRIRGGDGMKDL SPRWYFYYLGTGPEAGLPYGANRDGII WVATEGALNTPKDHIQTRNPANNAI VQLPQQTTLPGFCAEGSRGGSQASS RSSSRSRNSRNSRSTPGSSRGTSAPARMA GNGGDAALALLLLDRNLQLESKMSG KQGQQQQGQTVKKSAAEASKKPRQK RTATKAYNVTAQFGRGPEQTQGNGF DQELIRQGTDYKHWPQIAQFAPSASAF FGMSRIGMEVTPGTWLQYTAIQKLD DKDPNFKDQVILLNKHIDAYKTTFPTE PKKDDKAKADETQALPQRQKKQQTQTV TLLPAADLDFFSKQLQQSMSSSADSTQ A</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>LINK&lt;sub&gt;FL&lt;/sub&gt;</td>
<td>GPMSDNGPQNQRNAPRITFGGPSDSTG SNQNGERSGARSKQRRPQGLPNNTAS WFTALTQHGKEDLKFPCGQGVQYGPINTNS SPDDQIQGYYRRATRRIRGGDGMKDL SPRWYFYYLGTGPEAGLPYGANRDGII WVATEGALNTPKDHIQTRNPANNAI VQLPQQTTLPGFCAEGSRGGSQASS RSSSRSRNSRNSRSTPGSSRGTSAPARMA GNGGDAALALLLLDRNLQLESKMSG KQGQQQQGQCVTKKSAAEASKKPRQK RTATKAYNVTAQFGRGPEQTQGNGF DQELIRQGTDYKHWPQIAQFAPSASAF FGMSRIGMEVTPGTWLQYTAIQKLD DKDPNFKDQVILLNKHIDAYKTTFPTE PKKDDKAKADETQALPQRQKKQQTQTV TLLPAADLDFFSKQLQQSMSSSADSTQ A</td>
</tr>
<tr>
<td>DD&lt;sub&gt;FL&lt;/sub&gt;</td>
<td>GPMSDNGPQNQRNAPRITFGGPSDSTG SNQNGERSGARSKQRRPQGLPNNTAS WFTALTQHGKEDLKFPCGQGVQYGPINTNS SPDDQIQGYYRRATRRIRGGDGMKDL SPRWYFYYLGTGPEAGLPYGANRDGII WVATEGALNTPKDHIQTRNPANNAI VQLPQQTTLPGFCAEGSRGGSQASS RSSSRSRNSRNSRSTPGSSRGTSAPARMA GNGGDAALALLLLDRNLQLESKMSG KQGQQQQGQCVTKKSAAEASKKPRQK RTATKAYNVTAQFGRGPEQTQGNGF DQELIRQGTDYKHWPQIAQFAPSASAF FGMSRIGMEVTPGTWLQYTAIQKLD DKDPNFKDQVILLNKHIDAYKTTFPTE PKKDDKAKADETQALPQRQKKQQTQTV TLLPAADLDFFSKQLQQSMSSSADSTQ A</td>
</tr>
<tr>
<td>Protein</td>
<td>Sequence</td>
</tr>
<tr>
<td>---------</td>
<td>----------</td>
</tr>
<tr>
<td>CTD&lt;sub&gt;FL&lt;/sub&gt;</td>
<td>GNGGDAALALLLDRLNQLESKMSG KGQQQQQGQCVTKKSAAEASKKPRQK RTATKAYNVTQAFGRRGPEQTQGNFG DQELIRQTGTDYKHWPQIAQFAPSASAF FGMSRIGMEVTPSGTWLTYTGAIKLD DKDPNFKDQVILLNKHIDAYKTCPPTEPKKDKKKKKADETQALPQRQKKQKQT</td>
</tr>
<tr>
<td>Single-Labeled&lt;sub&gt;FL&lt;/sub&gt;</td>
<td>GPMSDNGPQNQRNAPRITFGGPDSTG SNQNGERSGARSQRRPQQLPNNTAS WFTALTTQHKGKEKLFPRGPQGVPIINTNS SPDDQIGYRRATRIRRGGDGKMKL SPRWYFYLYLGTGPEAGLPYGANKDGII WATEGALNTPKDHIGTRPANNAAI VLQLPGTTLPKGYAEGRSGGSQASS RSSSSRSRSSRNSTPGSSRTSPARMA GNGGDAALALLLDRLNQLESKMSG KGQQQQQGQCVTKKSAAEASKKPRQK RTATKAYNVTQAFGRRGPEQTQGNFG DQELIRQTGTDYKHWPQIAQFAPSASAF FGMSRIGMEVTPSGTWLTYTGAIKLD DKDPNFKDQVILLNKHIDAYKTCPPTEPKKDKKKKKADETQALPQRQKKQKQT</td>
</tr>
</tbody>
</table>
Supplementary Table 4.3. FRETTraj parameters for Alexa 488 and Alexa 594.

<table>
<thead>
<tr>
<th>FRETTraj parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>{&quot;Position&quot;:</td>
</tr>
<tr>
<td>{&quot;Alexa488&quot;:</td>
</tr>
<tr>
<td>{&quot;attach_id&quot;: 6,</td>
</tr>
<tr>
<td>&quot;linker_length&quot;: 12.5,</td>
</tr>
<tr>
<td>&quot;linker_width&quot;: 4.5,</td>
</tr>
<tr>
<td>&quot;dye_radius1&quot;: 5,</td>
</tr>
<tr>
<td>&quot;dye_radius2&quot;: 6,</td>
</tr>
<tr>
<td>&quot;dye_radius3&quot;: 3,</td>
</tr>
<tr>
<td>&quot;cv_fraction&quot;: 0.0,</td>
</tr>
<tr>
<td>&quot;cv_thickness&quot;: 6,</td>
</tr>
<tr>
<td>&quot;use_LabelLib&quot;: False},</td>
</tr>
<tr>
<td>&quot;Alexa594&quot;:</td>
</tr>
<tr>
<td>{&quot;attach_id&quot;: 936,</td>
</tr>
<tr>
<td>&quot;linker_length&quot;: 17,</td>
</tr>
<tr>
<td>&quot;linker_width&quot;: 3,</td>
</tr>
<tr>
<td>&quot;dye_radius1&quot;: 8,</td>
</tr>
<tr>
<td>&quot;dye_radius2&quot;: 5.7,</td>
</tr>
<tr>
<td>&quot;dye_radius3&quot;: 3,</td>
</tr>
<tr>
<td>&quot;cv_fraction&quot;: 0.0,</td>
</tr>
<tr>
<td>&quot;cv_thickness&quot;: 6,</td>
</tr>
<tr>
<td>&quot;use_LabelLib&quot;: False},</td>
</tr>
<tr>
<td>},</td>
</tr>
<tr>
<td>&quot;Distance&quot;: {&quot;Alexa488-Alexa594&quot;:</td>
</tr>
<tr>
<td>{&quot;R0&quot;: 54,</td>
</tr>
<tr>
<td>&quot;n_dist&quot;: 1000000}}</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Supplementary Table 4.4. RNAs used in this study

<table>
<thead>
<tr>
<th>RNA</th>
<th>Genomic position</th>
<th>nucleotides</th>
<th>Sequence 5’-3’</th>
<th>Origin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poly(rU)</td>
<td>-</td>
<td>10</td>
<td></td>
<td>IDT</td>
</tr>
<tr>
<td>SL5B</td>
<td>228 - 252</td>
<td>30</td>
<td>GGGCAUACCUA GUUUCGUCC GUGUGGCC</td>
<td>in vitro transcribed</td>
</tr>
</tbody>
</table>
**Supplementary Table 4.5.** Thermodynamic parameters fit from temperature dependence with 41 °C as reference temperature.

<table>
<thead>
<tr>
<th>van’t Hoff</th>
<th>ΔH° (kcal/mol)</th>
<th>ΔS° (kcal/mol K)</th>
<th>ΔC_p (kcal/ K)</th>
<th>ΔG° at 41 °C (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>fraction dimer</td>
<td>- 103 ± 5</td>
<td>- 0.30 ± 0.02</td>
<td>3 ± 2</td>
<td>-10 ± 7</td>
</tr>
<tr>
<td>ln(K_D)</td>
<td>- 95 ± 4</td>
<td>- 0.27 ± 0.01</td>
<td>6.4 ± 0.6</td>
<td>-10 ± 6</td>
</tr>
</tbody>
</table>
4.9 References


Chapter 5

The disordered N-terminal tail of SARS CoV-2

Nucleocapsid protein forms a dynamic complex with RNA


5.1 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.
5.2 Introduction

The SARS-CoV-2 virus is a positive-sense single-stranded RNA coronavirus with a genome of nearly 30000 nucleotides\(^1\). This large genome is packaged into small viral particles of \(~100\) nm diameter\(^2\). 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”\(^3,4\) formed by the SARS-CoV-2 N protein inside the virion are at variance with previously proposed helical structures seen in other coronaviruses\(^5,6\) 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\(^7\). Independent experiments from many labs (including ours) have demonstrated that N protein can undergo phase separation with nucleic acid, both \emph{in vitro} and in living cells\(^8\textendash16\). Phase separation can be favored by specific RNA sequence motifs\(^10\) and altered, in cells, by interactions with small molecules\(^17\). 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\(^7\). 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\textsuperscript{18} 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\textsuperscript{19}.

**Figure 5.1. 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.

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. 5.1 and Supplementary Tables 5.1 and 5.2) 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\textsuperscript{8,9,20} in a RNA dose-dependent manner\textsuperscript{9}, suggesting that localization is also mediated by its interaction with nucleic acid.
Single-molecule Förster Resonance Energy Transfer (FRET)\textsuperscript{21–23} 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\textsuperscript{24}), 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\textsuperscript{25}, while others found it dispersed over the viral genome, comprising both single-stranded and double-stranded regions\textsuperscript{16,26}. Given the lack of consensus in the literature, we have opted for “nonspecific” poly(rU)_\text{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\textsuperscript{27} (Supplementary Table 5.3).

5.3 Materials and Methods

5.3.1 Protein expression and purification. GST-His9-SARS-CoV2 NTD-RBD\textsubscript{L} and NTD\textsubscript{L}-RBD Nucleocapsid constructs were expressed recombinantly in Gold BL21(DE3) cells
(Agilent). 4 L cultures were grown in LB medium with carbenicillin (100 μg/mL) to OD₆₀₀ ~0.8 and induced with 0.25 mM IPTG for 3 hours at 37 °C. Harvested cells were lysed with sonication at 4 °C 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-RBD₄ and NTD₄-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 RBD₄ Nucleocapsid construct was expressed recombinantly in Gold BL21(DE3) cells (Agilent). 4 L cultures were grown in LB medium with carbenicillin (100 μg/mL) to OD₆₀₀ ~ 0.6 and induced with 0.3 mM IPTG for 3 hours at 37 °C. Harvested cells were lysed with sonication at 4 °C 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.

**5.3.2 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).

5.3.3 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 °C in a rotating mixer, the solution was recovered and gel debris removed by centrifugation. RNA was precipitated overnight at -20 °C 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.

5.3.4 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).
5.3.5 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:

\[
\overline{E} - \overline{E}_f = (\overline{E}_b - \overline{E}_f) \frac{K_A[RNA]_{tot}}{1 + K_A[RNA]_{tot}}
\]  

Eq. 5.1

where \(\overline{E}_f\) and \(\overline{E}_b\) 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:

\[
f_b = \frac{K_A[RNA]_{tot}}{1 + K_A[RNA]_{tot}}
\]

Eq. 5.2a

for the 1:1 binding cases, and:

\[
f_{b1} = \frac{K_{A1}[RNA]_{tot}}{1 + K_{A1}[RNA]_{tot} + K_{A1}K_{A2}[RNA]_{tot}}
\]

Eq. 5.2b

\[
f_{b2} = \frac{K_{A1}K_{A2}[RNA]_{tot}^2}{1 + K_{A1}[RNA]_{tot} + K_{A1}K_{A2}[RNA]_{tot}^2}
\]

Eq. 5.2c
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:

\[
f_b = \frac{K_{\text{int}} [\text{poly(rU)}]}{1 + K_{\text{int}} [\text{poly(rU)}]} \tag{Eq. 5.3}
\]

where \( K_{\text{int}} \) is the intrinsic association constant.

**5.3.6 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).

**5.3.7 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.
5.3.8 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. 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), 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 (Δ31-33). For assessing the role of each site, also we performed simulations of the P13L mutation alone versus the Δ31-33 alone. All simulations of the Omicron constructs were performed with and without(rU).

All simulations were run with multiple independent repeats using a 30 nm³ 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. 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. 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 5.4.

Simulations were analyzed using SOURSOP (https://soursop.readthedocs.io/) and MDTraj. 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.

5.3.9 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.

5.4 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. 5.1). 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. 5.1). 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.

5.4.1 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$^7$ 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. 5.2) that is compatible with previous observations of the completely folded RBD in the context of the NTD-RBD and full-length protein$^7$. 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$^7$. 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 $\Delta G_{UI} = 2.8 \pm 0.1$ kcal mol$^{-1}$ and $c_{UI,1/2} = 1.26 \pm 0.03$ M and $\Delta G_{IF} = 7.6 \pm 0.4$ kcal
mol$^{-1}$ and $c_{\text{IF}}^{1/2} = 1.21 \pm 0.01$ M (Fig. 5.2 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 5.5).

Figure 5.2. 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. S5.6 and S5.7.

5.4.2 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. 5.3)

**Figure 5.3. poly(rU) binding to RBD and NTD-RBD.**

A. Representative distributions of transfer efficiencies at different concentrations of poly(rU) for RBD. Distributions are fitted to a single Gaussian distribution.

B. Representative distributions of transfer efficiencies at different concentrations of poly(rU) for NTD-RBD. Distributions are fitted to a single Gaussian distribution.

C. Representative distributions of transfer efficiencies at different concentrations of poly(rU) for NTD-RBD. Distributions are fitted to two Gaussian distributions.

D. Variations in the mean transfer efficiency of RBD upon binding poly(rU).

E. Variations in the mean transfer efficiency of NTD-RBD upon binding poly(rU).

F. Fraction bound of NTD-RBD as a function of poly(rU) concentration. Solid lines represent the fit to the binding equations Eq. 5.3. Best fit values of $K_{int}$ are shown in Supplementary Table 5.6.

A plot of the deviation in mean transfer efficiency as a function of nucleic acid concentration reveals a 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\textsubscript{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 $K_{\text{int}}$ of $(6 \pm 2) \times 10^{-2}$ μM\textsuperscript{-1} (Fig. 5.3, Supplementary Table 5.6) 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 apparent weaker binding affinities because of the reduction in productive binding configurations for short oligonucleotides. When repeating the same titration, for (rU)\textsubscript{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. 5.4 and Supplementary Fig. 5.1). 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\textsuperscript{-1} to $(1.2 \pm 0.3)$ μM\textsuperscript{-1} (Supplementary Table 5.7).
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.

$$K_a = K_{int}(n - M + 1) \quad \text{Eq. 5.4}$$

Indeed, measured association constants follow a linear trend and fit to Eq. 5.4 results in an intrinsic association constant $K_{int} = (4.5 \pm 0.5) \times 10^{-2} \, \mu M^{-1}$ 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:

$$K_A = K_{int,M}(n - M + 1) + 2 \sum_{j=1}^{M-1} K_{int,j} \quad \text{for } M < n \quad \text{Eq. 5.5a}$$

$$K_A = K_{int}(M - n + 1) + 2 \sum_{j=1}^{M-1} K_{int,j} \quad \text{for } M \geq n \quad \text{Eq. 5.5b}$$

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 $K_{int} = (5.2 \pm 0.4) \times 10^{-2} \, \mu M^{-1}$ and a contact site size $M = 23 \pm 2$. Note that $K_{int}$ is within error of the
value determined with Eq. 5.4 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 change in 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. 5.5. The value falls between the estimates obtained with Eq. 5.4 and Eq. 5.5. 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.

5.4.3 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 \(K_{\text{int}} = (2.0 \pm 0.4) \mu\text{M}^{-1}\).

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\(^7\). 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 $\mu$M of poly(rU) (Fig. 5.3). 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. 5.3). These quantities can be further analyzed to extract binding affinity for the nucleic acid, $K_{int} = (4.0 \pm 0.3) \mu$M$^{-1}$, 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 $\sim$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. 5.4). 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. 5.4 and 5.2. A linear fit using Eq. 5.4 for RNAs with length between 20 and 40 nucleotides results in a \(K_{\text{int}} = (4.2 \pm 0.4) \mu\text{M}^{-1}\) and \(M = 21 \pm 1\) nucleotides. A complete fit of the dataset using Eq. 5.5 results in an intrinsic association constant \(K_{\text{int}} = (4.3 \pm 0.2) \mu\text{M}^{-1}\) and \(M = 25 \pm 2\) 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 \(\text{NTD}_L\)-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 \(\text{NTD-RBD} K_{\text{int}}\). 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 the absence of RNA is flexible and dynamic$^7$. 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. 5.2). This indicates that the NTD remains largely dynamic and contacts must occur only across a small set of nucleotides.
Figure 5.4. Length dependence of poly(rU) binding to NTD-RBD and RBD. A-B. Representative histograms of NTD-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, $E-E_0$, for RBD$_L$ (C) and NTD$_L$-RBD (D) for all nucleotide lengths. Compare with single titrations in Supplementary Fig. 5.1 for replicates and errors associated with each point. Solid lines are fit to Eq. 5.1. 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)$_n$.

5.4.4 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. 5.5A)\textsuperscript{28}. Mpipi was specifically developed with intrinsically disordered regions in mind\textsuperscript{28}. 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\textsuperscript{28,33,34}.

We first simulated RBD with (rU)\textsubscript{10} to identify residues on the folded domain that contribute to ssRNA binding (Fig. 5.5B). 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’-UCUCUAACG-3’.\textsuperscript{31} This result suggests that our simulations, at least qualitatively, are able to recapitulate experimentally measured protein:RNA interactions (Fig. 5.5B).

Having first performed simulations of (rU)\textsubscript{10} with the RBD, we next performed simulations of NTD-RBD and (rU)\textsubscript{10}. In addition to the previously observed RBD interactions with (rU)\textsubscript{10}, we now observed additional interactions between the disordered NTD and (rU)\textsubscript{10} (Fig. 5.5B, Supplementary Fig. 5.3). The NTD remains fully disordered in the bound state of NTD-RBD:(rU)\textsubscript{10} (Supplementary Fig. 5.4) and the pattern of RBD – (rU)\textsubscript{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. 5.3). This region maps to the β-extension previously identified as engaging in RNA interactions \textsuperscript{31}. Within the NTD, residues 30-50 contain five positively charged amino acids (four arginines and one lysine) and interact directly with (rU)\textsubscript{10}, in good agreement with recently published NMR experiments\textsuperscript{35} (Fig. 5.5B). 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.

**Figure 5.5. Coarse-grained simulations of the Nucleocapsid protein with ssRNA.** A. The Mipi forcefield is used to model SARS-CoV 2 N-protein interactions with ssRNA (rU)$_{10}$. 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. 5.3). 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)$_{25}$. 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.

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

236
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 5.8-5.9 and Supplementary Fig. 5.5). 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 \( 4.7 \pm 0.4 \% \), in good agreement with our experimentally measured ratio of association constants of \( 3 \pm 1 \% \) for \( K_{A,RBD}/K_{A,NTD-RBD} \) (Fig. 5.5D, Supplementary Table 5.10). Intriguingly, simulations of NTD alone with \((rU)_{25}\) revealed substantially weaker binding compared to either the RBD or NTD-RBD (Fig. 5.5D-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. 5.4). This is in contrast to simple expectations for polyelectrolyte condensation, where oppositely charged polymers are expected to compact upon interaction with one another\(^{36,37}\). 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. 5.5C). These trends are in qualitative agreement with the single-molecule FRET measurements (Fig. 5.4F). 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. 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.

5.4.5 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 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. 5.6 and 5.7, 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$^7$.

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. 5.6 and Supplementary Fig. 5.6-5.7. 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. 5.8). 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$^{38}$ (see Supplementary Table 5.11). 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 $(0.17 \pm 0.02) \mu M^{-1}$ for (rU)$_{20}$ and $(0.38 \pm 0.04) \mu M^{-1}$ for (rU)$_{40}$, corresponding to a weaker affinity in higher salt concentration (see Supplementary Table 5.12).

5.4.6 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\textsuperscript{10} and has been confirmed to adopt no secondary structure at room temperature\textsuperscript{39}.

We quantified binding of V21 using the NTD\textsubscript{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 \(~ 1 \mu\text{M} RNA\) concentration, which we interpret as representing the binding between one protein and one RNA strand. However, at concentrations of V21 higher than 1 \(\mu\text{M}\), 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. 5.7, Supplementary Table 5.13). Data are globally fit to a model that accounts for two distinct bound states with corresponding association constants \(K_{A1}^{V21}\) of \((6.2 \pm 0.3) \mu\text{M}^{-1}\) and \(K_{A2}^{V21}\) of \((0.15 \pm 0.10) \mu\text{M}^{-1}\). \(K_{A1}^{V21}\) is \(~ 50\%\) larger than the corresponding association constant for r(U)\textsubscript{20}, \(K_{A1}^{rU20} = (4.3 \pm 0.3) \mu\text{M}^{-1}\), whereas the mean transfer efficiency of the bound state appears only slightly smaller than that for r(U)\textsubscript{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.

**Figure 5.7. Specific ssRNA binding to NTD-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. 5.2b and 5.2c. B. Graphical representation of the SARS-CoV-2 5’ UTR based on Iserman et al.10, 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 5.13.

### 5.4.7 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 signals27, including the SL5B hairpin in the 5’ UTR and the NSP15 hairpin from the mRNA of the Nonstructural Protein 1527,40,41 (see Fig. 5.8, Supplementary Fig. 5.9, Supplementary Table 5.14). 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. 5.10).

**Figure 5.8. 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 5.8). Solid lines are fit to Eq. 5.1.

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: $K_{A^{NSP15}}^{N} = (7.8 \pm 0.7) \times 10^{-1}$ μM$^{-1}$ and $K_{A^{SL5B}}^{N} = (5.3 \pm 0.4) \times 10^{-1}$ μM$^{-1}$. These values are compatible with the one associated with the second binding mode of V21, $K_{A^{V21}}^{N}$, supporting the hypothesis that this binding mode is due to hybridization of a double-stranded RNA. Interestingly, the conformational changes of
NTD<sub>1</sub>-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)<sub>10</sub>, 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. 5.7). 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 \( K_A^{\text{Tetraloop}} = (6.7 \pm 0.8) \times 10^{-1} \mu M^{-1} \) and \( K_A^{\text{Decaloop}} = (3.4 \pm 0.5) \mu M^{-1} \), 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)<sub>10</sub>, 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 \( K_A^{\text{Tetraluge}} = (3.4 \pm 0.7) \times 10^{-1} \mu M^{-1} \), 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.

5.4.8 Omicron variant. Many mutations in the N protein occur within the disordered regions<sup>42</sup>. 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 (Supplementary Table 5.2). Residue 13 is part of a predicted short helix motif 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\textsubscript{L}-RBD (\textsuperscript{Om}\textsubscript{NTD\textsubscript{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. 5.9). 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 $K_A = (9 \pm 1) \times 10^{-1}$ 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\textsubscript{L}-RBD (Wuhan-Hu-1) and \textsuperscript{Om}NTD\textsubscript{L}-RBD (as supported by the same transfer efficiency in the bound state), but with different affinities (as indicated by the concentration dependence).
Figure 5.9. 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. 5.2a. D. Trend of the normalized binding affinity ($K_A^*$) predicted by simulations with Mpipi model for the Omicron mutant and additional variants.

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. 5.9). 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 5.15). 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.

5.5 DISCUSSION

5.5.1 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.
5.5.2 The NTD-RBD forms a dynamic complex with RNA. Our data confirm the previous observations that the NTD is a flexible and dynamic region\(^7\), 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 5.1) 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\(^7\)) 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. 5.11 and 5.12). These RBD residues comprise the positively charged \(\beta\)-extension, a flexible pair of beta strands that prior work has identified as wrapping around single-stranded RNA during binding \(^31\). Previous computational work proposed that the interplay between charged residues on the RBD surface and in the NTD can tune NTD conformational behavior\(^44\). 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 \(^{45}\). 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\(^{46}\), as for the case of prothymosin alpha and histone H1\(^{47,48}\).

5.5.3 **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\(^{10,49}\), including *in cell* crosslinked studies of
the protein to the 5’ UTR\textsuperscript{10}, 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.

5.5.4 NTD mutations alter RNA binding. A high number of mutations occur in disordered regions of the Nucleocapsid protein\textsuperscript{42}. 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 \textsuperscript{8,50}. 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.

5.6 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.

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.
5.7 Supporting Information

5.7.1 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.

5.7.2 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

$$ E = \frac{n_A}{n_A + n_D} $$

Eq. S5.1
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:

\[
S = \frac{I_D}{(I_D + \gamma_{PIE} I_A)}
\]

where \( I_D \) and \( I_A \) represent the total intensities observed after donor and acceptor excitation and \( \gamma_{PIE} \) 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).

### 5.7.3 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.
5.7.4 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 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 \( r = \langle R^2 \rangle^{1/2} \).

Estimates of this parameter is obtained by numerically solving:

\[
< E > = \int_0^\infty E(R) P(R) \, dR \quad \text{Eq. S5.3}
\]

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 \( R_0 \):

\[
E(R) = \frac{R_0^6}{R_0^6 + R^6} \quad \text{Eq. S5.4}
\]

The Gaussian chain distribution is given by:

\[
P(R) = 4\pi R^2 \left( \frac{3}{2\pi r^2} \right)^{3/2} \exp \left( \frac{-3R^2}{2r^2} \right) \quad \text{Eq. S5.5}
\]

Eqs. S5.4 and S5.5 are substituted into Eq. S5.3 and \( r \) is numerically optimized such that in integral equals the experimentally determined value for mean transfer efficiency.
### 5.7.5 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 $N \rightleftharpoons I \rightleftharpoons U$, the corresponding fraction folded, intermediate, and unfolded can be written in terms of the equilibrium constant $K_{UN}$ and $K_{NI}$ as:

\[
 f_U = \frac{1}{1 + K_{UI} + K_{UI} K_{IN}} \\
 f_I = \frac{K_{UI}}{1 + K_{UI} + K_{UI} K_{IN}} \\
 f_N = \frac{K_{UI} K_{IN}}{1 + K_{UI} + K_{UI} K_{IN}}
\]

The equilibrium constant $K_{UN}$ and $K_{NI}$ can be expressed as:

\[
 K_{UI} = \exp \left[ \frac{\Delta G_{0}^{UI}}{RT} \left( c - c^{UI} \right) / c^{UI} \right] \\
 K_{IN} = \exp \left[ \frac{\Delta G_{0}^{IN}}{RT} \left( c - c^{IN} \right) / c^{IN} \right]
\]

where $\Delta G_{0}^{UI}$ and $\Delta G_{0}^{IN}$ 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 $N \rightleftharpoons U$, the corresponding $c^{UN}$ represents the midpoint of the folding.

256
transition, in the general case with more than two states, the crossing points do not necessarily occur at the midpoint (50\%) of the transition.

5.7.6 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.:

$$K_A = \frac{\sum_i [(PR_{M})_{position_i}]}{[P][R]} = \sum_i K_{position_i}$$

Eq. S5.8

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\textsuperscript{51}).

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
\[ K_A = \frac{\sum_{i}^{(PR_{M})_{position_i}}}{[P][R_{M}]} = \sum_{i}^{K_{position_i}} = \sum_{m=i}^{M} (# \text{ positions with } m \text{ cont. nt})K_{int,m} \quad \text{Eq. S5.9} \]

**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:

\[ K_A = \frac{\sum_{i}^{(PR_{M})_{position_i}}}{[P][R_{M}]} = \sum_{i}^{K_{position_i}} = K_{int} \quad \text{for } M < n \quad \text{Eq. S5.10a} \]

\[ K_A = \frac{\sum_{i}^{(PR_{M})_{position_i}}}{[P][R_{M}]} = \sum_{i}^{K_{position_i}} = K_{int} \quad (M - n + 1) \quad \text{for } M \geq n \quad \text{Eq. S5.10b} \]

**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 \( K_{\text{int, } m} \) 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,

\[
K_A = K_{\text{int, } M} (n - M + 1) + 2 \sum_{j=1}^{M-1} K_{\text{int, } j} \text{ for } M < n \quad \text{Eq. S5.11a}
\]

\[
K_A = K_{\text{int, } M} (M - n + 1) + 2 \sum_{j=1}^{n-1} K_{\text{int, } j} \text{ for } M \geq n \quad \text{Eq. S5.11b}
\]

- the values of intrinsic association constants with stretches of nucleotides shorter than \( n \), \( K_{\text{int, } m} \), are given by

259
The terms in the equation can be conceptualized with the following reaction scheme:

\[ P + R_m \xrightarrow{K_{tg}} P \cdot R_m \xrightarrow{K_{m,n}} PR_m \]  

Eq. S5.13

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.

This equation for \( K_{\text{int},m} \) involves the assumption that the translational-rotational entropic cost of the bimolecular encounter in the proper orientation \( (\Delta G_{\text{tg}} = -RT \log K_{\text{tg}}) \) 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 \( (\Delta G_{\text{in},m} = m/n \Delta G_{\text{in}} \text{ or in terms of equilibrium constants, } K_{\text{in},m} = (K_{\text{in}})^{m/n}) \). Therefore:

\[
K_A = K_{\text{tg}} \{(K_{\text{in}})^{M/n} (n - M + 1) + 2 \sum_{j=1}^{M-1} (K_{\text{in}})^{j/n}\} \quad \text{for } M < n \quad \text{ Eq. S5.14a}
\]

\[
K_A = K_{\text{tg}} \{K_{\text{in}} (M - n + 1) + 2 \sum_{j=1}^{n-1} (K_{\text{in}})^{j/n}\} \quad \text{for } M \geq n \quad \text{ Eq. S5.14b}
\]
where the first term represents the binding to the longest available stretch of nucleotides (\(M \text{ if } M<n\), or \(n \text{ if } M>n\)) 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

\[
\sum_{j=1}^{M-1} (K_{in})^{j/n} = (K_{in})^{M/n} \left(1 - (K_{in})^{1-(M)/n} (K_{in})^{1/n-1}\right) \quad \text{Eq. S5.15a}
\]

\[
\sum_{j=1}^{n-1} (K_{in})^{j/n} = K_{in} \left(1 - (K_{in})^{1-(n)/n} (K_{in})^{1/n-1}\right) \quad \text{Eq. S5.15b}
\]

Then we have:

\[
K_A = K_{\text{int}} \left\{ (K_{in})^{M/n} [(n - M + 1) + 2 \frac{1-(K_{in})^{1-M/n}}{(K_{in})^{1/n-1}}] \right\} \text{ for } M < n \quad \text{Eq. S5.16a}
\]

\[
K_A = K_{\text{int}} (M - n + 1) + 2 \frac{1-(K_{in})^{1-(n)/n}}{(K_{in})^{1/n-1}} \text{ for } M \geq n \quad \text{Eq. S5.16b}
\]
or, in terms of free energy,

\[
K_A = K_{\text{int}} \left[ \frac{\Delta G \left( \frac{n-M}{n} \right)}{RT} \right] (n - M + 1) + \frac{2}{e^{\frac{\Delta G}{RT}} (1 - e^{\frac{\Delta G}{RT}})}\] for \( M < n \) \hspace{1cm} \text{Eq. S5.17a}

\[
K_A = K_{\text{int}} \left[ (M - n + 1) + \frac{2}{e^{\frac{\Delta G}{RT}} (1 - e^{\frac{\Delta G}{RT}})}\right] for \ M \geq n \hspace{1cm} \text{Eq. S5.17b}
\]

5.7.7 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\textsuperscript{52,53}. 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:

\[
g_{ij}(\tau) = 1 + \frac{1}{N} \left( 1 - c_{AB} \text{Exp}[-(\tau - \tau_0)/\tau_{AB}] \right) \times \\
(1 + c_{CD} \text{Exp}[-(\tau - \tau_0)/\tau_{CD}]) (1 + c_i \text{Exp}[-(\tau - \tau_0)/\tau_i]) \hspace{1cm} \text{Eq S5.18}
\]

where \( N \) is the mean number of molecules in the confocal volume and \( i \) and \( j \) indicate the type of signal (either from the \textbf{A}cceptor or \textbf{D}onor 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.\textsuperscript{54}

### 5.7.8 Coarse-grained simulations

Coarse-grained simulations were performed using the Mpipi model\textsuperscript{28}. 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\textsuperscript{55}. 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\textsuperscript{28,29}. 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. 5.5B). 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. 5.5C,D).

Having determined the fraction bound, we then calculated an apparent $K_D$ with the expression:

$$K_D = \frac{(1-f_{\text{bound}})^2}{N_A V f_{\text{bound}}}$$  \hspace{1cm} \text{Eq. S5.19}

where $f_{\text{bound}}$ 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. 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)$_n$. 

264
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\cite{57}. 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. 5.5D, F, G), despite the many assumptions made in the coarse-grained force field.

### 5.7.9 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:

$$Abs = \frac{(\alpha_{ds} + \beta_{ds}T) + (\alpha_{ss} + \beta_{ss}T)e^{-m(T-T_m)}}{1 + e^{-m(T-T_m)}}$$

Eq. S5.20
where $\alpha_{ds}$ and $\alpha_{ss}$ refer to the absorbance of the RNA in the ds and ss state at initial temperature.

$\beta_{ds}$ and $\beta_{ss}$ 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.
5.8 SUPPORTING FIGURES AND TABLES
Supplementary Figure S5.1. 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. S5.1.
Supplementary Figure S5.2. Dynamics of the disordered NTD when complex with RNA. A. Example of nanosecond-FCS (nsFCS) traces of NTD–RBD in the presence of (rU)$_n$. Donor-donor, acceptor-acceptor, and donor-acceptor correlations are shown in green, red, and orange (respectively) with the fit according to Eq. S5.18 and corresponding residuals. B. Reconfiguration times computed for the chain in the absence and in the presence of (rU)$_n$ with $n = 10, 20, 30, 40$. 

269
Supplementary Figure S5.3. 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.
Supplementary Figure S5.4. 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. 5.5C. 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.
Supplementary Figure S5.5. Simulations of N protein construct and RNA binding. A. Example simulation snapshots from the NTD-RBD + (rU)_15 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.
**Supplementary Figure S5.6.** 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)$_{20}$.

**Supplementary Figure S5.7.** 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.
Supplementary Figure S5.8. 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 5.12 (not visible because smaller than the marker for the experimental point).

Supplementary Figure S5.9. Transfer efficiency distributions for NTD$_L$-RBD and RNA hairpins. Representative histograms of NTD$_L$-RBD + hairpin RNA (hpRNA) as a function of concentration.
Supplementary Figure S5.10. 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. S5.20 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.
**Supplementary Figure S5.11. 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.
Supplementary Figure S5.12. 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. 5.11A, 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. 5.11A) 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. 5.11B).
Supplementary Table 5.1. Sequence of wild type NTD-RBD. Labeling positions are reported in red.

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
<th>Start Position (WT)</th>
<th>End Position (WT)</th>
<th>Labeling Positions (WT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTD\textsubscript{L}-RBD</td>
<td>GP\textcolor{red}{C}SDNGPQNQR\textcolor{red}{N}APRITFGGP\textcolor{red}{S}DSTGSNQNGERSGARSKQR\textcolor{red}{R}PQGLPNNTA\textcolor{red}{S}\textcolor{red}{D}STGSNQNGERSGARSKQR\textcolor{red}{R}PQGLPNNTA</td>
<td>1</td>
<td>173</td>
<td>1, 68</td>
</tr>
<tr>
<td>NTD&lt;sub&gt;L&lt;/sub&gt;-RBD Omicron (P13L, Δ31-33)</td>
<td>GP\text{CDNGPQNRNALRITFGGPGSDSTGSNQNGGARSKQRRPPQGLPNNTASWFTALTQHGKE DLKFPCGQQVSPDQDQ}YRRATRRIRGDGKMKDLS\text{PWY FYYLGTGPEAGLyGANKDGII W}VATE\text{GALNTPKDHIGTRNPANAAIVLQLPQGTLPKGFCA}</td>
<td>1</td>
<td>173</td>
<td>1, 68</td>
</tr>
<tr>
<td>-------------------------------------</td>
<td>----------------------------------------------------------------------------------</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>NTD-RBD&lt;sub&gt;L&lt;/sub&gt;</td>
<td>GP\text{MSDNGPQNRNAPRITFGGPGSDSTGSNQNGERSGARSKQRRPPQGLPNNTASWFTALTQHGKE DLKFPCGQQVSPDQDQ}YRRATRRIRGDGKMKDLS\text{PWY FYYLGTGPEAGLyGANKDGII W}VATE\text{GALNTPKDHIGTRNPANAAIVLQLPQGTLPKGFCA}</td>
<td>1</td>
<td>173</td>
<td>68,172</td>
</tr>
<tr>
<td>RBD&lt;sub&gt;L&lt;/sub&gt;</td>
<td>GP\text{GLPNNTASWFTALTQHGKE DLKFPCGQQVSPDQDQ}YRRATRRIRGDGKMKDLS\text{PWY FYYLGTGPEAGLyGANKDGII W}VATE\text{GALNTPKDHIGTRNPANAAIVLQLPQGTLPKGFCA}</td>
<td>44</td>
<td>173</td>
<td>68,172</td>
</tr>
</tbody>
</table>
Supplementary Table 5.3. RNA sequences used in this study.

<table>
<thead>
<tr>
<th>RNA</th>
<th>Genomic position</th>
<th>Nts</th>
<th>Sequence 5’-3’</th>
<th>Origin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poly(rU)</td>
<td>-</td>
<td>&lt;250</td>
<td></td>
<td>Midland Certified Reagent Company</td>
</tr>
<tr>
<td>Poly(rU)</td>
<td>-</td>
<td>10,12,15, 17,20,25, 30,35,40</td>
<td></td>
<td>IDT, Horizon Discovery</td>
</tr>
<tr>
<td>V21</td>
<td>127-148</td>
<td>21</td>
<td>UAUAUUUUAAUAACUAUAUUACU</td>
<td>IDT, Horizon Discovery</td>
</tr>
<tr>
<td>SL5B*</td>
<td>228 - 252</td>
<td>30</td>
<td>GGGCAUACCUCUAGGUUUGCUCCGGUGUGCC</td>
<td>in vitro transcribed</td>
</tr>
<tr>
<td>NSP15</td>
<td>19972-20000</td>
<td>31</td>
<td>GGGCUCACUGUCUUUUUGAAGGUGUAGAGUCC</td>
<td>in vitro transcribed</td>
</tr>
<tr>
<td>Tetraloop</td>
<td>based on NSP15</td>
<td>29</td>
<td>GGGCUCACUGUCUCCGAUGGUGAGCUC</td>
<td>in vitro transcribed</td>
</tr>
<tr>
<td>Decaloop</td>
<td>based on NSP15</td>
<td>34</td>
<td>GGGCUCACUGUCUUUUUUUUGUGAGUGAGUGAGGUGAGCUC</td>
<td>in vitro transcribed</td>
</tr>
<tr>
<td>Tetrabulge</td>
<td>based on NSP15</td>
<td>30</td>
<td>GGGCUCACUGUCUUCCGAUGGUAGAGUCC</td>
<td>in vitro transcribed</td>
</tr>
</tbody>
</table>

*The SL5B sequence is taken from the SARS-CoV genome and differs for two nucleotides from the SARS-CoV-2 genome.
Supplementary Table 5.4. Summary of simulation details.

<table>
<thead>
<tr>
<th>Simulation components</th>
<th>Box size (nm³)</th>
<th>Temp (K)</th>
<th>Steps per simulation (millions)</th>
<th>Number of independent replicas</th>
<th>Total product ion frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTD</td>
<td>30</td>
<td>298</td>
<td>300</td>
<td>30 (5x6 starting conf)</td>
<td>268,650</td>
</tr>
<tr>
<td>RBD</td>
<td>30</td>
<td>298</td>
<td>300</td>
<td>30 (5x6 starting conf)</td>
<td>268,420</td>
</tr>
<tr>
<td>NTD-RBD</td>
<td>30</td>
<td>298</td>
<td>300</td>
<td>25 (5x5 starting conf)</td>
<td>221,592</td>
</tr>
<tr>
<td>NTD + (rU)₂₅</td>
<td>30</td>
<td>298</td>
<td>300</td>
<td>30 (5x6 starting conf)</td>
<td>268,650</td>
</tr>
<tr>
<td>RBD + (rU)₁₀</td>
<td>30</td>
<td>298</td>
<td>300</td>
<td>30 (5x6 starting conf)</td>
<td>268,650</td>
</tr>
<tr>
<td>RBD + (rU)₁₂</td>
<td>30</td>
<td>298</td>
<td>300</td>
<td>30 (5x6 starting conf)</td>
<td>268,650</td>
</tr>
<tr>
<td>RBD + (rU)₁₅</td>
<td>30</td>
<td>298</td>
<td>300</td>
<td>30 (5x6 starting conf)</td>
<td>264,405</td>
</tr>
<tr>
<td>RBD + (rU)₁₇</td>
<td>30</td>
<td>298</td>
<td>300</td>
<td>30 (5x6 starting conf)</td>
<td>268,322</td>
</tr>
<tr>
<td>RBD + (rU)₂₀</td>
<td>30</td>
<td>298</td>
<td>300</td>
<td>30 (5x6 starting conf)</td>
<td>267,693</td>
</tr>
<tr>
<td>RBD + (rU)₂₅</td>
<td>30</td>
<td>298</td>
<td>300</td>
<td>30 (5x6 starting conf)</td>
<td>265,683</td>
</tr>
<tr>
<td>RBD + (rU)₃₀</td>
<td>30</td>
<td>298</td>
<td>300</td>
<td>30 (5x6 starting conf)</td>
<td>265,026</td>
</tr>
<tr>
<td>RBD + (rU)₃₅</td>
<td>30</td>
<td>298</td>
<td>300</td>
<td>30 (5x6 starting conf)</td>
<td>268,650</td>
</tr>
<tr>
<td>RBD + (rU)₄₀</td>
<td>30</td>
<td>298</td>
<td>300</td>
<td>30 (5x6 starting conf)</td>
<td>263,567</td>
</tr>
<tr>
<td>RBD + (rU)₁₈₀</td>
<td>30</td>
<td>298</td>
<td>300</td>
<td>30 (5x6 starting conf)</td>
<td>268,503</td>
</tr>
<tr>
<td>Sample Name</td>
<td>T</td>
<td>X</td>
<td>Y</td>
<td>Z</td>
<td>Conf</td>
</tr>
<tr>
<td>-------------</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>------</td>
</tr>
<tr>
<td>NTD-RBD + (rU)&lt;sub&gt;10&lt;/sub&gt;</td>
<td>30</td>
<td>298</td>
<td>300</td>
<td>30 (5x6 starting conf)</td>
<td>268,650</td>
</tr>
<tr>
<td>NTD-RBD + (rU)&lt;sub&gt;12&lt;/sub&gt;</td>
<td>30</td>
<td>298</td>
<td>300</td>
<td>30 (5x6 starting conf)</td>
<td>268,650</td>
</tr>
<tr>
<td>NTD-RBD + (rU)&lt;sub&gt;15&lt;/sub&gt;</td>
<td>30</td>
<td>298</td>
<td>300</td>
<td>30 (5x6 starting conf)</td>
<td>267,563</td>
</tr>
<tr>
<td>NTD-RBD + (rU)&lt;sub&gt;17&lt;/sub&gt;</td>
<td>30</td>
<td>298</td>
<td>300</td>
<td>30 (5x6 starting conf)</td>
<td>267,341</td>
</tr>
<tr>
<td>NTD-RBD + (rU)&lt;sub&gt;20&lt;/sub&gt;</td>
<td>30</td>
<td>298</td>
<td>300</td>
<td>30 (5x6 starting conf)</td>
<td>265,965</td>
</tr>
<tr>
<td>NTD-RBD + (rU)&lt;sub&gt;25&lt;/sub&gt;</td>
<td>30</td>
<td>298</td>
<td>300</td>
<td>30 (5x6 starting conf)</td>
<td>263,434</td>
</tr>
<tr>
<td>NTD-RBD + (rU)&lt;sub&gt;30&lt;/sub&gt;</td>
<td>30</td>
<td>298</td>
<td>300</td>
<td>30 (5x6 starting conf)</td>
<td>268,134</td>
</tr>
<tr>
<td>NTD-RBD + (rU)&lt;sub&gt;35&lt;/sub&gt;</td>
<td>30</td>
<td>298</td>
<td>300</td>
<td>30 (5x6 starting conf)</td>
<td>268,650</td>
</tr>
<tr>
<td>NTD-RBD + (rU)&lt;sub&gt;40&lt;/sub&gt;</td>
<td>30</td>
<td>298</td>
<td>300</td>
<td>30 (5x6 starting conf)</td>
<td>267,220</td>
</tr>
<tr>
<td>NTD-RBD + (rU)&lt;sub&gt;180&lt;/sub&gt;</td>
<td>30</td>
<td>298</td>
<td>300</td>
<td>30 (5x6 starting conf)</td>
<td>268,650</td>
</tr>
<tr>
<td>OmNTD-RBD (P13L,Δ&lt;sup&gt;31-33&lt;/sup&gt;) + (rU)&lt;sub&gt;25&lt;/sub&gt;</td>
<td>30</td>
<td>298</td>
<td>300</td>
<td>30 (5x6 starting conf)</td>
<td>267,333</td>
</tr>
<tr>
<td>OmNTD-RBD (P13L) + (rU)&lt;sub&gt;25&lt;/sub&gt;</td>
<td>30</td>
<td>298</td>
<td>300</td>
<td>30 (5x6 starting conf)</td>
<td>268,527</td>
</tr>
<tr>
<td>OmNTD-RBD (Δ&lt;sup&gt;31-33&lt;/sup&gt;) + (rU)&lt;sub&gt;25&lt;/sub&gt;</td>
<td>30</td>
<td>298</td>
<td>300</td>
<td>30 (5x6 starting conf)</td>
<td>266,023</td>
</tr>
</tbody>
</table>
**Supplementary Table 5.5. RBD Folding parameters.**

<table>
<thead>
<tr>
<th></th>
<th>RBD&lt;sub&gt;L&lt;/sub&gt; full-length protein</th>
<th>RBD&lt;sub&gt;L&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>c&lt;sub&gt;1/2,IF&lt;/sub&gt; (M)</td>
<td>1.68 ± 0.02</td>
<td>1.26 ± 0.03</td>
</tr>
<tr>
<td>c&lt;sub&gt;1/2,IF&lt;/sub&gt; (M)</td>
<td>1.64 ± 0.02</td>
<td>1.21 ± 0.01</td>
</tr>
<tr>
<td>ΔG&lt;sub&gt;U1&lt;/sub&gt; (kcal mol&lt;sup&gt;-1&lt;/sup&gt;M&lt;sup&gt;-1&lt;/sup&gt;)</td>
<td>6.6 ± 0.5</td>
<td>2.8 ± 0.1</td>
</tr>
<tr>
<td>ΔG&lt;sub&gt;IF&lt;/sub&gt; (kcal mol&lt;sup&gt;-1&lt;/sup&gt;M&lt;sup&gt;-1&lt;/sup&gt;)</td>
<td>8.1 ± 0.5</td>
<td>7.6 ± 0.4</td>
</tr>
</tbody>
</table>

**Supplementary Table 5.6. Intrinsic association constants**

<table>
<thead>
<tr>
<th></th>
<th>K&lt;sub&gt;int&lt;/sub&gt; for poly(rU) (µM&lt;sup&gt;-1&lt;/sup&gt;) nucleotides</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBD&lt;sub&gt;L&lt;/sub&gt;</td>
<td>(6 ± 1) x 10&lt;sup&gt;-2&lt;/sup&gt;</td>
</tr>
<tr>
<td>NTD-RBD&lt;sub&gt;L&lt;/sub&gt;</td>
<td>2.0 ± 0.4</td>
</tr>
<tr>
<td>NTD&lt;sub&gt;L&lt;/sub&gt;-RBD</td>
<td>4.0 ± 0.3</td>
</tr>
</tbody>
</table>

**Supplementary Table 5.7. RBD<sub>L</sub> and NTD<sub>L</sub>-RBD association constants for (rU)<sub>n</sub> as measured by single-molecule FRET**

<table>
<thead>
<tr>
<th>n</th>
<th>K&lt;sub&gt;A&lt;/sub&gt; (µM&lt;sup&gt;-1&lt;/sup&gt;) molecules</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RBD&lt;sub&gt;L&lt;/sub&gt;</td>
</tr>
<tr>
<td>10</td>
<td>(4 ± 3) x 10&lt;sup&gt;-2&lt;/sup&gt;</td>
</tr>
<tr>
<td>12</td>
<td>(7.9 ± 0.4) x 10&lt;sup&gt;-2&lt;/sup&gt;</td>
</tr>
<tr>
<td>15</td>
<td>(1.1 ± 0.6) x 10&lt;sup&gt;-1&lt;/sup&gt;</td>
</tr>
<tr>
<td>17</td>
<td>(1.9 ± 0.4) x 10&lt;sup&gt;-1&lt;/sup&gt;</td>
</tr>
<tr>
<td>20</td>
<td>(2.6 ± 0.7) x 10&lt;sup&gt;-1&lt;/sup&gt;</td>
</tr>
<tr>
<td>25</td>
<td>(6.0 ± 0.9) x 10&lt;sup&gt;-1&lt;/sup&gt;</td>
</tr>
<tr>
<td>30</td>
<td>(6.6 ± 1.3) x 10&lt;sup&gt;-1&lt;/sup&gt;</td>
</tr>
<tr>
<td>35</td>
<td>-</td>
</tr>
<tr>
<td>40</td>
<td>1.2 ± 0.3</td>
</tr>
</tbody>
</table>

283
### Supplementary Table 5.8. Simulation-derived association constants (K_A) in µM^1

<table>
<thead>
<tr>
<th>Construct</th>
<th>NTD; K_A (µM^1)</th>
<th>RBD; K_A (µM^1)</th>
<th>NTD-RBD; K_A (µM^1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(rU)_{10}</td>
<td>(9.5 ± 0.7) x 10^{-5}</td>
<td>(4.6 ± 0.4) x 10^{-4}</td>
<td>(1.90 ± 0.05) x 10^{-3}</td>
</tr>
<tr>
<td>(rU)_{12}</td>
<td>(2.5 ± 0.2) x 10^{-4}</td>
<td>(8.7 ± 0.8) x 10^{-4}</td>
<td>(4.2 ± 0.1) x 10^{-3}</td>
</tr>
<tr>
<td>(rU)_{15}</td>
<td>(5.6 ± 0.4) x 10^{-4}</td>
<td>(1.7 ± 0.1) x 10^{-3}</td>
<td>(1.00 ± 0.03) x 10^{-2}</td>
</tr>
<tr>
<td>(rU)_{17}</td>
<td>(8.0 ± 0.5) x 10^{-4}</td>
<td>(2.5 ± 0.2) x 10^{-3}</td>
<td>(1.8 ± 0.1) x 10^{-2}</td>
</tr>
<tr>
<td>(rU)_{20}</td>
<td>(1.10 ± 0.04) x 10^{-3}</td>
<td>(3.5 ± 0.1) x 10^{-3}</td>
<td>(4.3 ± 0.7) x 10^{-2}</td>
</tr>
<tr>
<td>(rU)_{25}</td>
<td>(2.00 ± 0.08) x 10^{-3}</td>
<td>(5.2 ± 0.1) x 10^{-3}</td>
<td>(1.1 ± 0.1) x 10^{-1}</td>
</tr>
<tr>
<td>(rU)_{30}</td>
<td>(2.7 ± 0.1) x 10^{-3}</td>
<td>(8.1 ± 0.8) x 10^{-3}</td>
<td>(2.1 ± 0.2) x 10^{-1}</td>
</tr>
<tr>
<td>(rU)_{35}</td>
<td>(4.1 ± 0.2) x 10^{-3}</td>
<td>(1.1 ± 0.1) x 10^{-2}</td>
<td>(4.0 ± 1.0) x 10^{-3}</td>
</tr>
<tr>
<td>(rU)_{40}</td>
<td>(4.7 ± 0.1) x 10^{-3}</td>
<td>(1.30 ± 0.06) x 10^{-2}</td>
<td>(6.9 ± 0.7) x 10^{-3}</td>
</tr>
<tr>
<td>(rU)_{180}</td>
<td>1.20 ± 0.02</td>
<td>3.2 ± 0.6</td>
<td>(1.0 ± 0.2) x 10^{-2}</td>
</tr>
</tbody>
</table>

### Supplementary Table 5.9. Simulation-derived dissociation constants (K_D) in µM.

<table>
<thead>
<tr>
<th>Construct</th>
<th>NTD; K_D (µM)</th>
<th>RBD; K_D (µM)</th>
<th>NTD-RBD; K_D (µM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(rU)_{10}</td>
<td>(1.10 ± 0.08) x 10^{4}</td>
<td>2200 ± 200</td>
<td>540 ± 10</td>
</tr>
<tr>
<td>(rU)_{12}</td>
<td>(4.0 ± 0.3) x 10^{3}</td>
<td>1200 ± 100</td>
<td>240 ± 9</td>
</tr>
<tr>
<td>(rU)_{15}</td>
<td>(1.8 ± 0.1) x 10^{3}</td>
<td>570 ± 30</td>
<td>98 ± 3</td>
</tr>
<tr>
<td>(rU)_{17}</td>
<td>(1.20 ± 0.08) x 10^{3}</td>
<td>400 ± 30</td>
<td>56 ± 4</td>
</tr>
<tr>
<td>(rU)_{20}</td>
<td>(9.0 ± 0.4) x 10^{2}</td>
<td>290 ± 10</td>
<td>24 ± 3</td>
</tr>
<tr>
<td>(rU)_{25}</td>
<td>(5.0 ± 0.2) x 10^{2}</td>
<td>192 ± 5</td>
<td>9.0 ± 0.6</td>
</tr>
<tr>
<td>(rU)_{30}</td>
<td>(4.0 ± 0.1) x 10^{2}</td>
<td>120 ± 10</td>
<td>4.7 ± 0.5</td>
</tr>
<tr>
<td>(rU)_{35}</td>
<td>(2.4 ± 0.1) x 10^{2}</td>
<td>92 ± 9</td>
<td>2.7 ± 0.8</td>
</tr>
<tr>
<td>(rU)_{40}</td>
<td>(2.10 ± 0.05) x 10^{2}</td>
<td>80 ± 4</td>
<td>1.5 ± 0.1</td>
</tr>
<tr>
<td>(rU)_{180}</td>
<td>0.80 ± 0.02</td>
<td>0.30 ± 0.07</td>
<td>(1.0 ± 0.1) x 10^{-2}</td>
</tr>
</tbody>
</table>
### Supplementary Table 5.10. Simulation-derived ratio of association constants $K_{A}^{*}$ defined as $(K_{A}$ of Construct + (rU)$_n$)/(K$_A$ of NTD-RBD + (rU)$_{25}$).

<table>
<thead>
<tr>
<th>Construct</th>
<th>NTD; $K_{A}^{*}$</th>
<th>RBD; $K_{A}^{*}$</th>
<th>NTD-RBD; $K_{A}^{*}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(rU)$_{10}$</td>
<td>$(8.5 \pm 0.9) \times 10^4$</td>
<td>$(4.2 \pm 0.5) \times 10^3$</td>
<td>$(2.0 \pm 0.1) \times 10^2$</td>
</tr>
<tr>
<td>(rU)$_{12}$</td>
<td>$(2.3 \pm 0.2) \times 10^3$</td>
<td>$(7.8 \pm 0.9) \times 10^3$</td>
<td>$(4.0 \pm 0.3) \times 10^2$</td>
</tr>
<tr>
<td>(rU)$_{15}$</td>
<td>$(5.0 \pm 0.5) \times 10^3$</td>
<td>$(1.6 \pm 0.1) \times 10^2$</td>
<td>$(9.0 \pm 0.7) \times 10^2$</td>
</tr>
<tr>
<td>(rU)$_{17}$</td>
<td>$(7.2 \pm 0.7) \times 10^3$</td>
<td>$(2.2 \pm 0.2) \times 10^2$</td>
<td>$(1.6 \pm 0.2) \times 10^1$</td>
</tr>
<tr>
<td>(rU)$_{20}$</td>
<td>$(1.0 \pm 0.1) \times 10^2$</td>
<td>$(3.1 \pm 0.2) \times 10^2$</td>
<td>$(3.9 \pm 0.7) \times 10^1$</td>
</tr>
<tr>
<td>(rU)$_{25}$</td>
<td>$(1.8 \pm 0.1) \times 10^2$</td>
<td>$(4.7 \pm 0.3) \times 10^2$</td>
<td>$1.0 \pm 0.1$</td>
</tr>
<tr>
<td>(rU)$_{30}$</td>
<td>$(2.4 \pm 0.2) \times 10^2$</td>
<td>$(7.3 \pm 0.9) \times 10^2$</td>
<td>$1.9 \pm 0.2$</td>
</tr>
<tr>
<td>(rU)$_{35}$</td>
<td>$(3.7 \pm 0.3) \times 10^2$</td>
<td>$(9.9 \pm 1.3) \times 10^2$</td>
<td>$3.6 \pm 0.9$</td>
</tr>
<tr>
<td>(rU)$_{40}$</td>
<td>$(4.2 \pm 0.3) \times 10^2$</td>
<td>$(1.10 \pm 0.09) \times 10^1$</td>
<td>$6.2 \pm 0.7$</td>
</tr>
<tr>
<td>(rU)$_{180}$</td>
<td>$10.7 \pm 0.7$</td>
<td>$28 \pm 6$</td>
<td>$900 \pm 200$</td>
</tr>
</tbody>
</table>

### Supplementary Table 5.11. Ion released upon binding of (rU)$_{20}$ and (rU)$_{40}$ (compare with Fig. 6 and Supplementary Fig. 5)

<table>
<thead>
<tr>
<th></th>
<th>$\alpha$ (KCl)</th>
<th>$\alpha$ (KCl+Tris HCl)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(rU)$_{20}$</td>
<td>$-3.49 \pm 0.05$</td>
<td>$-5.0 \pm 0.1$</td>
</tr>
<tr>
<td>(rU)$_{40}$</td>
<td>$-3.7 \pm 0.5$</td>
<td>$-5.0 \pm 0.7$</td>
</tr>
</tbody>
</table>
Supplementary Table 5.12. Association and dissociation constants of NTD$_L$-RBD as a function of salt concentration for (rU)$_{20}$ and (rU)$_{40}$

<table>
<thead>
<tr>
<th></th>
<th>K$_A$ (µM$^{-1}$)</th>
<th>50 mM KCl</th>
<th>110 mM KCl</th>
<th>150 mM KCl</th>
<th>175 mM KCl</th>
<th>200 mM KCl</th>
</tr>
</thead>
<tbody>
<tr>
<td>(rU)$_{20}$</td>
<td>8 ± 2</td>
<td>(5.4 ± 0.6) x 10$^{-1}$</td>
<td>(1.7 ± 0.2) x 10$^{-1}$</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>(rU)$_{40}$</td>
<td>14 ± 2</td>
<td>-</td>
<td>(3.8 ± 0.4) x 10$^{-1}$</td>
<td>(2.5 ± 0.3) x 10$^{-1}$</td>
<td>(5 ± 0.9) x 10$^{-2}$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>K$_D$ (µM)</th>
<th>50 mM KCl</th>
<th>110 mM KCl</th>
<th>150 mM KCl</th>
<th>175 mM KCl</th>
<th>200 mM KCl</th>
</tr>
</thead>
<tbody>
<tr>
<td>(rU)$_{20}$</td>
<td>0.12 ± 0.03</td>
<td>1.9 ± 0.2</td>
<td>5.8 ± 0.5</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>(rU)$_{40}$</td>
<td>0.07 ± 0.01</td>
<td>-</td>
<td>2.6 ± 0.2</td>
<td>4.0 ± 0.4</td>
<td>19 ± 4</td>
<td></td>
</tr>
</tbody>
</table>

Supplementary Table 5.13. NTD$_L$-RBD association constants for V21 binding

<table>
<thead>
<tr>
<th></th>
<th>K$_A$ (µM$^{-1}$) molecules</th>
</tr>
</thead>
<tbody>
<tr>
<td>K$_{A1}$</td>
<td>6.2 ± 0.3</td>
</tr>
<tr>
<td>K$_{A2}$</td>
<td>0.15 ± 0.10</td>
</tr>
</tbody>
</table>

Supplementary Table 5.14. NTD$_L$-RBD association constants for hairpin RNA sequences

<table>
<thead>
<tr>
<th></th>
<th>K$_A$ (µM$^{-1}$) molecules</th>
</tr>
</thead>
<tbody>
<tr>
<td>hpRNA</td>
<td>NTD$_L$-RBD</td>
</tr>
<tr>
<td>NSP15</td>
<td>(7.8 ± 0.7) x 10$^{-1}$</td>
</tr>
<tr>
<td>Tetraloop</td>
<td>(6.7 ± 0.8) x 10$^{-1}$</td>
</tr>
<tr>
<td>Tetrabulge</td>
<td>(3.4 ± 0.7) x 10$^{-1}$</td>
</tr>
<tr>
<td>Decaloop</td>
<td>3.4 ± 0.5</td>
</tr>
<tr>
<td>SL5B</td>
<td>(5.3 ± 0.4) x 10$^{-1}$</td>
</tr>
</tbody>
</table>
Supplementary Table 5.15. Simulation-derived ratio of association constants $K_A^*$ defined as

$$(K_A$ of Construct + (rU)$_n$)/(K$_A$ of NTD-RBD + (rU)$_{25}$)

<table>
<thead>
<tr>
<th>Construct</th>
<th>OmNTD-RBD (P13L,$\Delta^{31-33}$); $K_A^*$</th>
<th>NTD-RBD (P13L); $K_A^*$</th>
<th>NTD-RBD ($\Delta^{31-31}$); $K_A^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(rU)$_{25}$</td>
<td>(7 ± 2) x 10$^{-1}$</td>
<td>1.0 ± 0.2</td>
<td>(7 ± 1) x 10$^{-1}$</td>
</tr>
</tbody>
</table>
5.8 REFERENCES


53. Nettels, D., Hoffmann, A. & Schuler, B. Unfolded protein and peptide dynamics


Chapter 6

Macromolecular crowding and intrinsically disordered proteins: a polymer physics perspective.

This chapter is adapted from:


**Author contributions.** J.C. and A.S. wrote the manuscript.
6.1 Abstract

The cell is a crowded environment where a relevant fraction of the available space is occupied by proteins, nucleic acids, and metabolites. Here we present recent advancements in the understanding of crowding effects on intrinsically disordered proteins. Differently from their structured counterparts, these proteins do not adopt a stable three-dimensional structure and remain flexible and dynamic in solution. The physics of polymers and colloids provides a framework to interpret how crowding modulates conformations, dynamics, and interactions of disordered proteins. Flory-Huggins models enable rationalizing the different degree of compaction induced by crowding agents in terms of depletion interactions. The same interactions modulate the diffusion of the disordered proteins in a crowded milieu and the association and dissociation rates when interacting with a ligand. All together, this theoretical framework provides new insights on the interpretation of the effects of the cellular environment on disordered proteins.
6.2 Introduction

At variance with the typical conditions at which several biochemical experiments are performed, the cellular environment is not a dilute aqueous buffer solution. Quite the opposite, the cellular milieu is a solution containing a significant concentration of different components (proteins, nucleic acids, metabolites), such that between 10% to 25% of the volume\(^1\) is considered not accessible. Given that 75% of the volume in the cell remains available, an occupied volume fraction of 25% may seem like a small number. To understand its impact on a molecule of interest, it is useful to put it into context using a simple model, where the “crowding agents” are represented by hard spheres. This model provides an easy reference point to estimate how the solution is densely occupied: the maximum packing fraction for randomly distributed spheres approaches 50% of volume fraction\(^6\) and at 25% the average distance between two spheres is about 1 nm for a solution of 6-nm radius spheres (see Fig. 6.1A). This distance is significantly less than the size of each other sphere; therefore, the addition of a sphere of equivalent size is likely to be impacted by the presence of the others in solution.

![Volume Fraction](Image)

**Figure 6.1. Volume Fraction** A. Examples of three solutions with different volume fractions of randomly distributed hard spheres (radius of 6 nm), from 5% to 25%. In these solutions, the average distance between the two surfaces of a given sphere and the closest one is approximately 4.2, 1.8, and 1.1 nm, respectively. B. Example of “excluded volume” effects: the center of mass of the gray sphere (black dot) cannot enter any of the circles surrounding the black spheres.
From a physico-chemical point of view, reducing the space accessible to a specific molecule of interest introduces non-ideal contributions to the chemical potential of the solution that has important implications on reaction equilibria and kinetics. This phenomenon is commonly referred to as “macromolecular crowding” or simply “crowding”. Though early examples of crowding effects were already observed on protein solubility, association, and enzymatic activity, its relevance to the biochemistry of proteins and nucleic acids was rationalized in foundational works by Minton. It is important to note that the term “crowding” refers to a purely excluded volume effect (Fig. 6.1B), where the accessible space in the solution is reduced by “crowding agents”. However, no molecule is completely inert and non-specific interactions can arise between the molecule of interest and the crowded milieu. To this end, recent work has started to investigate the contribution of such interactions in the crowded cellular environment. A large body of work, both experimental and theoretical, has provided quantitative insights on how crowding impacts folding and binding of structured proteins. Here, we aim to provide a succinct yet detailed description of how crowding can impact disordered proteins and how, viewing the phenomenon through the lens of polymer physics, can help in rationalizing the mechanisms at play.

In the last twenty years disordered proteins have emerged as a class of proteins that, despite the lack of a well-defined structure, play key-roles in many biological processes, from transcription and translation to signaling and cellular assemblies, as well as in several diseases. Compared to structured proteins, intrinsically disordered proteins (IDPs) and regions (IDRs) require a paradigm shift in understanding how intra- and inter- molecular forces regulate conformations, dynamics, binding, and ultimately, their function. It is intuitive that their
lack of a stable conformation makes the conformations and dynamics of a disordered protein very susceptible to the surrounding environment. The finding that biomolecular condensates are often driven by IDRs interactions, where condensation leads to a significant increase in the local concentration of IDRs, poses a further question on how disordered proteins can act as polymeric crowders.

In the following, we will first describe the elements that control the conformations and dynamics of IDRs in dilute solutions, which provide the baseline for understanding the impact of crowding, and we will then discuss how crowding agents can impact conformations, diffusion, and binding of these proteins.

6.3 Conformations and dynamics of IDRs in dilute solutions.

To understand the effects of crowding on IDRs, it is important to first consider the conformations and dynamics adopted by IDRs in a dilute solution. IDRs conformations are commonly reported in terms of the end-to-end distance (which represents an average measure of the distance between the first and the last residue of the chain), radius of gyration (which represents an average measure of the distance of each residue from the center of mass of the protein), or hydrodynamic radius (which reports about the diffusion of the overall molecule). Since disordered proteins adopt a heterogeneous ensemble of conformations, these quantities are described in terms of probability distributions and their mean-square values. Many experimental approaches, such as single-molecule FRET$^{51-63}$, Small-Angle X-ray Scattering$^{54,64-67}$ and Nuclear Magnetic Resonance$^{68,69}$, enables quantifying the mean-square realizations, whereas the corresponding distributions are either inferred using polymer models (Gaussian chain, Worm-like chain, Self-Avoiding Walk)$^{70}$ or quantified via simulations (all-atom or coarse-grained)$^{71-73}$. 

297
In absence of structured elements, the conformations of IDRs are dictated by the interplay between residue-residue and residue-solvent interactions, with attractive residue-residue interactions leading to compaction and favorable residue-solvent interactions or repulsive residue-residue interactions leading to expansion. Polymer physics offers a precise framework to classify IDRs of different length and composition in terms of these interactions\textsuperscript{74}. When attractive residue-residue interactions dominate over residue-solvent interactions, the protein is in a “poor solvent” regime and adopts compact conformations typically described as a “globule”. Viceversa, when the favorable interaction with the solvent dominates over the attractive residue-residue interactions, the protein is in a “good solvent” and adopts expanded configurations. The limiting case where attractive and repulsive interactions within the chain and with the solvent counterbalance each other is commonly referred to as the “theta solvent” condition (Fig. 6.2 A-C). Polymer physics assigns to each of these conditions a very specific scaling exponent $\nu$\textsuperscript{75} that links the mean-square end-to-end distance $R$ (or one of the other relevant dimensions) with the sequence length $N$ (i.e. the number of amino acids):

$$R \propto N^{\nu} \quad \text{Eq.6.1}$$

Specifically, in a “theta solvent” condition the exponent $\nu$ is equal to 0.5, whereas in the case of a good solvent is equal to 0.588\textsuperscript{76} and for a globule $\nu$ is equal to 1/3. The exponent value for the globule reflects the proportional relation between the volume of the globule and the number of monomers within the globule (Fig. 6.2 A).

Due to finite-size effects\textsuperscript{77,78}, a continuum of values of the scaling exponent can be observed, through experimental measurements (such as single-molecule FRET and SAXS) and
computational simulations, which reflects the interactions encoded by the protein sequence [Fig. 6.2F]. IDR−ons are commonly enriched in hydrophilic and charged residues\cite{79,80}. Hydrophilic residues favor expanded conformations, whereas hydrophobic residues contribute to chain compaction\cite{60}. Charged residues contribute in two distinct ways [Fig. 6.2D-E]. The first way is via the polyelectrolyte effect, which depends on the net charge of the sequence and represents the overall repulsion of identical charges within the chain, leading to an expansion of the conformations\cite{81-83}. The second way is via the polyampholyte effect, which depends on the total charge of the chain (sum of positive and negative charges) and represents a correction to the polyelectrolyte contribution, due to the local attraction of positive and net charges\cite{79}. The balance of these contributions can be easily understood in terms of the Higgs and Joanny theory\cite{79}, where the chain dimensions are described as the results of the contribution of a mean “effective excluded volume” $w^* b^3$ per residue:

$$w^* b^3 = w b^3 + \frac{4\pi l_B^2 (f-g)^2}{\kappa^2} - \frac{\pi l_B^2 (f+g)^2}{\kappa},$$

Eq. 6.2

where $w b^3$ is the excluded volume term in absence of electrostatics, $b$ is the bond length, $f$ and $g$ are the fraction of positive and negative charges, $l_B$ is the Bjerrum length (which sets the strength of the electrostatic interaction) and $\kappa$ is the inverse of the Debye screening length (which accounts for ion screening of charged residues). Note $w^*$ and $w$ represent numerical factors that rescale the volume of the bond length $b$ to account for differences in the sequence composition as well as for proportionality factors.
It is intuitive to imagine that a larger excluded volume results in more expanded configurations, whereas a smaller excluded volume leads to more compact conformations. The term “effective excluded volume” is introduced to describe the resulting effect between the physical excluded volume (steric repulsion) and the contribution of other interactions. As such, the effective excluded volume is a positive quantity in a good solvent, equal to zero in a theta solvent, and negative in a poor solvent. A zero effective excluded volume does not mean that there are no intrachain interactions, but that the interactions are counterbalancing each other. Similarly, a negative effective excluded volume indicates that attractive intrachain interactions are stronger than the repulsive contribution of the physical excluded volume.

As described by Eq. 2, the effective excluded volume \( \omega b^3 \) is the result of three terms. The first term, \( \omega b^3 \), represents the mean physical excluded volume (i.e. the mean volume occupied) by the residues in the protein of interest. The second term is the polyelectrolyte term, which depends on the mean square net charge per residue \((f - g)^2\). This term is always positive or equal to zero and therefore contributes to an increase of the effective excluded volume. The third term describes the polyampholyte contributions and depends on the total fraction of charges in the chain \((f + g)^2\). The sign of this contribution is negative and therefore diminishes the effective excluded volume. In this respect, it can lead to a cancellation or even overcompensation of the polyelectrolyte term. The balance among these elements decides whether electrostatic interactions between residues contribute to expansion or collapse of the chain.

Some charged residues can also interact with hydrophobic residues via cation-pi interactions, which can further contribute to the modulation of IDR's dimensions by compensating the polyelectrolyte repulsion.
A clear role is played by the sequence patterning of residues, where well-distributed sequences have been found to adopt very different conformations than the one in which specific types of residues (e.g., charged or aromatic) are segregated (Fig 6.2 G-I). Recent polymer theories have started to address the role of patterning in disordered proteins and various factors have been proposed to quantify and compare the degree of patterning of the sequence.

Besides charged residues, two notable cases are the ones of proline and glycine residues, which represent two opposite extremes for what concerns the dihedral angles sampled within the Ramachandran plot. Interestingly, glycine does not only introduce a high degree of flexibility, but also strong backbone-backbone interactions, which ultimately favors chain compaction. Instead, proline residues have been found to buffer chain conformations, e.g., by counteracting the effect of charged residues on chain conformations.

Investigation of protein dynamics by nanosecond-FRET fluorescence correlation spectroscopy (FCS), photoinduced electron transfer (PET)-FCS, NMR, and simulations have revealed an intimate relation between protein conformations and dynamics. While one would expect IDRs dynamics to be enslaved by the solvent viscosity, independent evidence from experiment and simulations supports the occurrence of internal interactions that dominate intrachain diffusion. These contributions are commonly referred as “internal friction”, where the term “internal” refers to the origin of the interactions and explicits their non-solvent mediated nature. Various sources have been identified for “internal friction” effects, including contribution of dihedral angle hopping and transient contact formation between residues, in line with prediction for polymer models. These internal friction effects can become dominant on the dynamics of IDRs, with contributions that can be even more than ten times the one observed for the solvent-enslaved component. Intriguingly, due to their nature,
internal friction effects correlate with protein conformations: expanded chains in good solvent approaches the limit of a chain dominated by solvent viscosity, whereas more compact chains near theta state exhibit transient contacts and are dominated by internal friction. With these elements in mind, we can now discuss the impact of crowding on disordered proteins.

**Figure 6.2. Conformations of disordered proteins in dilute solutions.**

A. Dependence of the end-to-end distance with sequence length according to Eq. 6.1 for poor, theta, and good solvent. B. Corresponding distance distribution for a disordered protein of length 200 amino acids, assuming a common prefactor \( R_0 = 0.55 \text{ nm} \). C. Coil-to-globule transition of the end-to-end distance distribution, normalized by the dimension of the chain at the theta condition, as function of intrachain interactions: favorable intrachain interactions increase from right to left and favorable solvent interaction from left to right. D. Schematic depiction of the components determining the effective excluded volume: the physical excluded volume, the repulsive contribution of the polyelectrolyte term (net fractional charge of the chain), and the attractive contribution of the polyampholyte term (total fractional charge of the chain) (see Eq. 6.2). E. Boundaries for polyampholyte and polyelectrolyte effects as determined by Das and Pappu using simulation of disordered proteins containing different fractions of positive and negative charges. F. Scaling exponents of disordered proteins as a function of the normalized hydropathy scale of Kyte and Doolittle and as a function of the absolute value of the net charge per residue (NCPR) for Martin et al. (magenta), Hofmann et al. (orange), Fuertes et al. (blue), Zheng et al. (red), Bowman et al. (cyan), and Ryback et al. (green).

302
6.4 Crowding effects on disordered proteins

Early experiments on disordered proteins revealed that addition of crowding agents do not lead to stabilization of structured configurations in IDPs\textsuperscript{109–111}; however, some sequences that exhibit transient secondary structure or fold upon binding with a ligand may exhibit an increase structural content in presence of crowding \textsuperscript{112,113}. Here, we want to contextualize some of the recent observations within concepts borrowed from the physics of polymers and colloids, which provides a general framework for interpreting and classifying the different mechanisms controlling conformational changes, folding, and binding of disordered proteins in a crowded milieu.

6.4.1 Investigating crowding effects using synthetic polymers.

Crowding effects are commonly studied \textit{in vitro} by mimicking the crowded cellular environment with synthetic polymers, such as polyethylene glycol (PEG), polyvinyl alcohol (PVA), Dextran, and Ficoll. These synthetic polymers are often available with different degrees of polymerization and crosslinking, have a high water solubility even at volume fractions comparable to the ones found in the cellular milieu, and are relatively inexpensive. Even when performing \textit{in cell} experiments, the comparison with artificial crowding conditions can provide a useful term of comparison to rationalize experimental results\textsuperscript{114}.

Particular attention must be used in characterizing the interaction between the molecule of interest and the crowding polymer, since the molecule composition may be more or less prone to specific interactions with the polymer, as shown in previous case for PEG \textsuperscript{8,115–118} and Dextran\textsuperscript{119}. Equally important is to consider the concentration regime in which experiments are performed. Polymers are flexible molecules and, at a sufficiently high concentration, start
overlapping and intertwining with each other. Therefore, describing polymers as hard spheres is a strong simplification that is in some measure valid only when the polymers are in a dilute regime. Polymer physics provides a precise definition for determining the threshold concentration between the dilute and semidilute regime. The threshold concentration \( \phi^* \) is equal to the one at which the volume fraction of the overall solution is identical to the volume fraction of a single chain (Fig. 6.3A). This concentration depends on the degree of polymerization of the chain and above the critical concentration, the solution properties are no longer dependent on the radius of the polymer but only on the mesh size of the semidilute solution (Fig. 6.3B). Importantly, \( \phi^* \) scales with \( P^{-4/5} \) (i.e.: \( \phi^* (P) \propto P^{-4/5} \)) where \( P \) is the degree of polymerization; hence, polymers such as PEG are often in semidilute conditions in the experimental conditions used to test crowding effects. At even higher concentrations, the polymer solution enters in the so-called “concentrated regime”, where the polymers become entangled and motions in the solution can happen only via “reptation” because of the high density packing of polymers. While often neglected, even branched and crosslinked polymers can overlap\(^{120,121} \). Finally, commercial polymers are normally polydisperse and differences in the polydispersion can impact the measured contributions.

These considerations are essential when we want to quantitatively describe the properties of a “synthetic” crowded solution. A natural question is whether such overlapping solutions are of any interest in the context of the cellular environment. In this respect, it is important to note that the concentration of disordered proteins within the nuclear pore complex can be estimated between 25% and 55% of volume fraction, about ten times the overlap concentration\(^{122} \). Similarly, a high concentration of disordered proteins are found in biomolecular condensates or reconstituted condensates. Examples from in vitro experiments suggest concentrations of the
dense phase in the 1-50 mM range\textsuperscript{85,114,123}, which are above the corresponding overlap concentrations. Overall, these observations suggest that disordered proteins may act also as crowders themselves; therefore, understanding the corresponding polymer physics may help reveal details on the impact that such protein assemblies can have on a protein of interest.

Figure 6.3. Dilute and semidilute regime of polymers. A. Schematic depiction of the dilute regime, critical overlap concentration, and semidilute regime. B. Theoretical dependence of the critical overlap concentration with the degree of polymerization reveals a rapid decrease with increasing polymer length.

6.4.2 Crowding effects on IDR conformations.

The absence of structural changes in IDRs within a crowded solution does not imply that the protein conformations are left unchanged. This problem was early recognized in the context of the effects of crowding on protein folding, where the unfolded state adopts a disordered ensemble\textsuperscript{124}.

To contextualize the effects of crowding on IDR conformations, it is useful to recapitulate the fundamental concepts at the basis of the scaled particle theory\textsuperscript{125} and its application to the
problem of crowding\textsuperscript{26}. Scaled particle theory describes a solution filled with a certain volume fraction of hard spheres and provides an estimate for the cost of inserting an empty rigid cavity within the solution (Fig. 6.4A). Here the radius of the cavity coincides with the radius of gyration of the protein (or molecule) of interest and the radius of the hard spheres recapitulates the radius of gyration of the crowding agents\textsuperscript{26}. The term “scaled” underlines how the free energy cost of inserting the cavity is completely dependent on the ratio between the radius of gyration of the cavity (protein) and the radius of gyration of the crowding agents. A typical prediction of the model is that small crowding agents have a stronger impact than large crowding agents. This can be interpreted accounting for the fact that bigger crowding agents give rise to larger interstitial cavities, where it is easier to accommodate the molecule of interest.

While this approach has been successfully applied to structured proteins, a critical element is presented when assuming a rigid cavity for describing the space occupied by an IDR, since crowding agents can enter in the volume occupied by the flexible protein (compare Fig. 6.4A-B). This problem has been explicitly discussed by Minton and a Gaussian Cloud scaled particle theory model\textsuperscript{26} has been proposed to address this limitation. In this Gaussian Cloud model, the probability of inserting the flexible protein in a solution of hard-spheres is computed by describing the disordered protein (or unfolded state) in terms of a Gaussian distribution (the cloud) and computing the probability of a clash between the protein cloud and the hard-sphere crowding agent. While there is always a probability of collision between the protein and the crowders, this probability is always different from one and decreases to zero with increasing distance from the center of mass of the protein. Furthermore, the probability depends on both the size of the protein and the size of hard spheres and, obviously, on the concentration of the crowders. At a fixed volume fraction, small crowding agents will have little impact on the chain
dimensions, because the probability of clashes between the small crowders and the protein are low. With increasing size of the hard spheres, the probability of clashes will increase resulting in a stronger protein collapse. Above a certain size of the crowding agent, the probability of clashes between the cloud and the polymers approaches the result of the classical scaled particle theory. As such, further increasing the size of the crowders will result in a lower impact of the crowding effects, because of the large interstitial spaces between hard-spheres that can accommodate the protein.

In this respect, it is interesting to compare these predictions with simulations performed on flexible molecules. Qin and Zhou\textsuperscript{126} studied the impact of crowding agents on the disordered conformational ensemble as a function of the degree of compaction of the protein for a fixed hard sphere size, finding that compact ensembles are less sensitive to crowding, whereas more expanded ensembles exhibit stronger collapse (Fig. 6.4G-H). Interestingly, the degree of compaction is not monotonic since it is a balance between the crowding effect and the intrachain interactions within the disordered ensemble. Kang et al.\textsuperscript{127} studied the effect of different sizes of crowding agents, mimicked by hard spheres, on the conformations of a disordered polymer chain: the small crowding agents cause a strong collapse of the disordered chain (Fig. 6.4E), whereas larger crowding agents have a mild impact on chain conformations (Fig. 6.4F). Miller et al.\textsuperscript{128} investigated the effects of three different sizes of crowding agents on chains with different compositions (from hydrophobic to hydrophilic sequences). They observed an increase of crowding-induced collapse for more hydrophilic (and therefore expanded) sequences as well as a systematic smaller impact of the larger crowders.

Compaction of the disordered ensemble has been reported in experiments for denatured and disordered proteins\textsuperscript{129–131}, but not all the disordered proteins have been shown to exhibit
strong collapses as expected from scaled particle predictions\textsuperscript{132,133,134}. These observations pose a question whether there are further contributions in the experiments that are not captured by scaled particle theory.

One important factor that is often neglected and that conflicts with the common assumptions of scaled particle theory is that polymeric crowding agents can overlap with each other. As such, the representation of crowding as a solution of hard spheres is no longer pertinent when the solution is in a semidilute regime (compare Fig. 6.3). In this respect, Flory-Huggins theories provide a useful framework to account for both the degree of polymerization of the protein and the one of the crowding agents. In these models, the protein collapses with increasing concentration of crowding agents to avoid trapping crowding agents within its volume, therefore resulting in a decrease of their translational entropy\textsuperscript{135}. To this end, depletion of the crowding agents from the disordered protein volume comes at the cost of chain compaction. The exact degree of compaction of a single protein depends on the volume concentration of the polymer and on whether the protein is mixed with shorter or longer polymer chains. The concept of shorter (and, vice versa, of longer) polymers is described in polymer physics by the Flory’s criterion\textsuperscript{136,137} and is summarized by the expression $P < N^{1/2}$ (or $P \geq N^{1/2}$), where $P$ is the degree of polymerization of the polymer and $N$ is the degree of polymerization of the disordered protein. Under this conditions, the chain dimensions of the protein, expressed in terms of radius of gyration $R_g$, can be described by:

$$R_g(N, P, \phi, a) = R_g^0 \left( \frac{1}{1 + a\phi/\phi^* (P)} \right)^{1/5} \text{ for } P < N^{1/2} \quad \text{Eq. 6.3a}$$
where $\alpha$ is an empirical parameter that can account for differences in the solvent quality for the different proteins and interactions between protein and polymer$^{135}$, $s_{NP}$ quantifies the interaction between the protein and the polymer chains, $\phi^*(P)$ is the overlap concentration dependent on the degree of polymerization $P$, and $f(N, P, \phi, s_{NP})$ is a function that represents the renormalization mapping in the case of a polymer in a bath of long chains$^{138-140}$.

Single-molecule experiments investigated the impact of different polymer lengths on the degree of compaction of four different disordered proteins with similar sequence lengths$^{122,135}$. Despite the similarity in sequence length, the four proteins exhibit different degrees of expansion in dilute solutions, reflecting different residue-residue and residue-solvent interactions, with scaling exponents ranging from 0.5 to 0.7 (compare Fig. 6.2H and 6.4J). When increasing the concentration of crowders, different degrees of compaction are observed (Fig. 6.4J). Flory-Huggins theories described by Eqs. 6.3a and 6.3b provide a quantitative fit of the experimental data, revealing that there are no significant changes in the strength of interaction between the polymers and the different protein sequences. Instead, the discrepancy in the degree of compaction is expected because of the different degree of expansions of the protein and associated scaling exponents. In the limit of a polymer melt, all the configurations of the proteins are expected to tend toward the conformational behavior of an ideal chain (the polymer chain in a theta solvent). Therefore, the more expanded the proteins are in good solvent, the more sensitive they will be to crowding agents; conversely, proteins that are close to the theta state will be less sensitive to changes in crowding. The same observations have been reproduced in
computational studies investigating the impact of crowding on disordered proteins of different degree of expansion in presence of polymeric crowders that mimic PEG6000.\textsuperscript{141}

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure6.4.png}
\caption{Crowding models and conformational changes of disordered proteins. A. Scaled Particle Theory B. Gaussian-cloud modification of scaled particle theory. C-D. Flory-Huggins theory in the limit of a polymer chain (the disordered protein) in a solution of shorter chains (the crowding agents) (C) or in a solution of longer chains (D). E-F. Simulations of a disordered chain in a bath or small hard-spheres (E) or in a bath of large spheres (F)\textsuperscript{127}. The $\lambda$ parameter represents the ratio between the radius of gyration of the polymer chain and the radius of the hard spheres in solution. G-H. Simulations of disordered proteins with different degrees of expansions as function of volume fraction of crowding agents\textsuperscript{128}. The parameter $\xi$ represents different intrachain interactions, from 0.5 to 1, with higher values representing stronger residue-residue interactions (G). Comparison between the predictions from scaled particle theory and gaussian cloud model with simulations (H) J. Experimental results from single-molecule FRET measurements of four distinct protein sequences with similar sequence length\textsuperscript{122}. Data represents changes of radius of gyration $R_g$ as a function of crowders concentration.}
\end{figure}

\textbf{6.4.3 Crowding effects on binding equilibrium.}

One of the important advantages of disordered proteins is their ability of binding with specificity to multiple different ligands, harnessing conformational heterogeneity as a tool for
recognizing different binding partners. Because of the interaction with the ligand, many proteins can fold upon binding and adopt more structured conformations\textsuperscript{142}, can assemble in tight complexes and remain completely disordered\textsuperscript{62}, or dynamically diffuse on their folded ligand\textsuperscript{143}. The IDR sequence encodes for the affinities of such binding reactions and, interestingly, the conformational modulation of the unbound disordered state has a pronounced contribution to the affinity\textsuperscript{144}. Despite the nature of the final complex, it is clear that such a binding process can be impacted by crowding, either favoring or disfavoring the bound state.

The impact of crowding on binding equilibrium can be described in terms of a thermodynamic cycle in which the binding of two components in a dilute solution is compared with the one in a crowded solution\textsuperscript{8} (Fig. 6.5A). Scaled particle theory offers a framework to describe association equilibria in terms of the size of the two binding components and of crowders in solution\textsuperscript{8,145}. While the theory can account for different shapes of the protein components in the bound and unbound states, e.g. by introducing spherocylindrical descriptions of the complexes\textsuperscript{146}, their application to polymeric solutions and disordered proteins remains challenging because of the flexible and overlapping nature of both the polymers used as crowding agents and the disordered proteins.

A more general framework to describe this phenomenon is offered by the theory of depletion forces introduced by Asakura and Oosawa\textsuperscript{147} and developed in the context of colloids, which offers a similar semantic to the one adopted for the compaction of the disordered ensemble in section 3.2. For a detailed description of this theory and its application, the book from Lekkerkerker and Tuinier\textsuperscript{148} provides a detailed overview of concepts and applications of depletion theory to colloids. For the case of interest here, let’s start by considering two spherical colloids that are immersed in a solution of polymers. The colloids perceive a relative attraction
because of entropic forces: similarly to the case discussed above, the two spheres prefer a closer configuration to limit the number of crowding agents that otherwise would be trapped within the interface between the two colloids. The binding free energies $\Delta G$ and $\Delta G_0$ measured in presence or absence of the crowders, respectively, are linked to the equilibrium dissociation constants $K_D$ and $K_{D,0}$ via:

$$
\Delta \Delta G = \Delta G - \Delta G_0 = -k_B T \ln \frac{K_{D,0}}{K_D}
$$

Eq. 6.4

where $k_B$ is the Boltzmann constant and $T$ the absolute temperature.

The contribution of the crowded environment can be modeled in terms of the osmotic pressure $\Pi = n k_B T$ of the solution, where $n$ is the number density of colloids (related to the concentration $c$ by $n = c/M$ with $M$ being the molar mass of the crowding agents). The distance-dependent attractive interaction potential between the colloids, $W(r)$, is the result of the product of the osmotic pressure $\Pi$ and the overlap volume $V_{\text{overlap}}(r)$ of the colloids depletion layers. It is reasonable to assume that for two proteins\(^{149}\) the net stabilization of the complex $\Delta \Delta G$ corresponds to the interaction potential when two spheres are brought into contact, $W(0)$:

$$
\Delta \Delta G = W(0) = -n k_B T V_{\text{overlap}}(0) = -\frac{c}{M} k_B T V_{\text{overlap}}(0)
$$

Eq.6.5

where the overlap volume is defined by:
\[ V_{\text{overlap}}(0) = \frac{\pi (r+R-d)^2 (d^2 - 3(r-R)^2 + 2d(r+R))}{12d} \]  

Eq. 6.6

with \( r = R_1 + \delta_s(R_1), \quad R = R_2 + \delta_s(R_2), \quad d = R_1 + R_2, \) with

\[
\frac{\delta}{R} = \left( 1 + 3\frac{\delta_0}{R} + 2.273\left(\frac{\delta_0}{R}\right)^2 - 0.0975\left(\frac{\delta_0}{R}\right)^3 \right)^3 - 1 \quad \text{[150,151]}, \quad \text{and} \quad \delta_0 = 1.07R_g.
\]

Note that the depletion layer \( \delta_0 \) is slightly larger than the radius of gyration \( R_g \) of the polymers to account for the flexible nature of dilute polymers. The overlap volume \( V_{\text{overlap}}(0) \) represents the overlap between the depletion layers of the two molecules, i.e. the space not accessible by a single polymer when the two colloids are in contact with one another.

To account for the length scale of relevance in dilute and semidilute solutions, it is convenient to define an effective depletion layer \( \delta \) that depends on the thickness of the depletion layer in a dilute solution \( \delta_0 \) and on the average mesh size \( \xi \) of the polymer solution in the semidilute regime\textsuperscript{152}:

\[
\delta^{-2} = \delta_0^{-2} + \xi^{-2} \quad \text{Eq. 6.7}
\]

In this way, when the average mesh size of the polymer \( \xi \) is large (i.e., the polymer is in the dilute regime or in the semidilute regime but close to the overlap critical concentration \( \phi^* \)), the depletion layer is equivalent to \( \delta_0 \); however, with decreasing size of \( \xi \) (i.e. increasing concentration in the semidilute regime), the mesh size \( \xi \) can become dominant over \( \delta_0 \).
Additional factors that need to be taken into account include that: 1) disordered proteins may have a smaller overlap volume when compared to hard sphere colloids, because of the different interpenetration ratio\textsuperscript{153}; 2) at a high concentration of short crowding agents, one may need to account for multiple layers of polymers being enclosed between the colloidal spheres\textsuperscript{154}. One important result of this theory is that the strongest effect happens with the largest crowding agent, whereas the smallest crowding agent has the least effect. This is in contrast with predictions from scaled particle theory, where the smallest crowding agent is expected to have the largest impact.

The predictions of this theory well describe the binding of the two disordered proteins, ACTR and NCBD, across several concentrations (from 0 to 40\%) and degrees of polymerization of PEG (from monomer to PEG35000)\textsuperscript{155} (Fig. 6.6A).

In order to describe the effects on the kinetics of binding, one needs to consider two additional factors: the impact of crowding on the viscosity of solution and on the association rate (the dissociation rate can be derived from the equilibrium constant). The depletion theory offers the advantage of providing a common language across these two other elements.
Figure 6.5. Crowding effects on binding. A. Thermodynamic cycle for the binding of two molecules A and B in a dilute solution and in a crowded environment. B. Depletion Layer in a dilute and semidilute polymer solution for a spherical particle. C. Binding equilibrium is affected by depletion of polymer molecules: this can be described in terms of the osmotic pressure of the solution and of the overlap volume of the depletion layers for the molecule A and B (see Eq. 6.5). In the case of a disordered protein, the depletion layer is reduced since the protein and the polymers can interpenetrate. D. Impact of crowding on microviscosity: the depletion layer identifies an area surrounding the molecule of interest that does not experience the same density of the remaining part of the solution. E. Binding kinetics as described in terms of the association rate, which depends on the depletion interactions favoring the binding, the increased viscosity slowing down the diffusion, and the cavity formation that accelerates diffusion when the two molecules are close to each other.

6.4.4 Crowding effects on diffusion.

Let’s now consider how crowding impacts the diffusion of disordered proteins. One important aspect to account for is that viscosity in a crowded solution is length-dependent, because the solution is not uniform across all the scales; therefore, the probed length scale dictates what type of viscosity is perceived from the protein of interest. This is not per se a specific property of disordered proteins; however, disordered proteins exhibit dynamics over multiple length scales (e.g., diffusion of the overall molecule and intrachain diffusion) and, therefore, different
viscosities can be detected depending on the phenomenon studied. While bulk viscosity can be measured with a viscometer, microviscosity can be quantified by a technique that can sense the molecule in its local environment, e.g. using Fluorescence Correlation Spectroscopy\textsuperscript{153,158}. Experimental data can be interpreted in terms of empirical relationship between the protein size and length-scale dominating protein interactions\textsuperscript{159}. Various theories account for the probe size as well as size and concentration of solutes, starting from the Einstein model for a suspension of hard spheres and its modifications\textsuperscript{160} and including applications of scaled particle theory\textsuperscript{161,162}. Here we focus on the model presented by Tuinier et al\textsuperscript{155} in the context of polymer solutions. In this picture, the probe (in our case the disordered protein) is surrounded by a depletion layer of thickness $\delta$, where the polymers crowding the solution are excluded. However, after passing this layer, the density of crowding agents increases and the diffusion is dominated by the interaction with the polymer solution (Fig. 6.5D). By using the effective depletion layer $\delta$ defined in Eq. 6.7, this model is able to account for the transition between dilute and semidilute regions in polymer solutions. If the probe is sufficiently small and the polymer solution has a concentration above the overlap concentration, the probe will experience diffusion through the mesh of the polymer network and, as such, will perceive a very different (reduced) diffusion than the one expected by the bulk viscosity. Conversely, if the crowding agent is very small in size, compared to the probe, diffusion as measured by the probe particle will be very similar to bulk viscosity. These observations can be quantified in terms of a microviscosity $\eta_{micro}$ that depends linearly on the solvent viscosity $\eta_s$ according to:

\[
\eta_{\text{micro}} = \eta_s \frac{Q(\lambda, \varepsilon)}{Z(\lambda, \varepsilon)} \tag{6.8}
\]
where $\lambda$ is equal to the ratio of solvent and bulk viscosity $\eta_s/\eta_{bulk}$, $\varepsilon$ is equal to $\delta/R$ and represents the ratio of depletion layer thickness and particle radius, and $Q(\lambda, \varepsilon)$ and $Z(\lambda, \varepsilon)$ are algebraic functions of the ratio of solvent and bulk viscosity:

$$Q(\lambda, \varepsilon) = 2(2 + 3\lambda)(1 + \varepsilon)^6 - 4(1 - \lambda)(1 + \varepsilon) \quad \text{Eq. 6.9a}$$

$$Z(\lambda, \varepsilon) = 2(2 + 3\lambda)(1 + \varepsilon)^6 - (1 - \lambda)[3(3 + 2\lambda)(1 + \varepsilon)^5 + 10(1 + \varepsilon)^3 - 9(1 + \varepsilon) + 4(1 - \lambda)] \quad \text{Eq. 6.9b}$$

This model has been extensively tested for a disordered protein in the context of multiple concentrations (across the dilute and semidilute regime) and different degrees of polymerization for PEG\textsuperscript{153} (Fig. 6.6B) with minor deviations from the expected behavior. As for the case of protein conformations, it is important to quantify the microviscosity across different polymer lengths and possibly using different polymers, since this procedure enables a more robust test of theoretical predictions. Furthermore, the characterization of synthetic crowded solutions for specific proteins of interest can provide a useful comparison for interpreting viscosity changes occurring in living cells\textsuperscript{114}. It is interesting to note that measurements in cells reveal very different viscosities as measured on the nanometer scale (where intrachain dynamics occur) and on the hundreds of nanometers scale (where the diffusion of the molecule is detected by FCS)\textsuperscript{153,163} (compare Fig. 6.6G and 6.6I).
6.4.5 Crowding effects on binding kinetics.

The same language derived for describing the effect of crowding on binding equilibrium and on the diffusion of a molecule in terms of depletion forces can be used to describe the impact of crowding on the association rate of binding between two molecules. The basic foundation of this model relies on the work of Collins and Kimball on the diffusion-controlled reaction rates\(^{164}\):

\[
\frac{1}{k_{\text{on}}} = \frac{1}{k_0} + \frac{1}{4\pi D_0} \frac{1}{R_{\text{contact}}} 
\]

Eq. 6.10

Here \(k_{\text{on}}\) is the association rate, \(D_0\) is the diffusion coefficient of the two reactants (one relative to each other) in a dilute solution, \(R_{\text{contact}}\) is the contact radius at which the reaction occurs, and \(k_0\) is the intrinsic rate of the reaction once the reactants are within the contact radius.

Recently, Berezhkowskii and Szabo extended this theory to explicitly account for the effects of crowding. In this scenario, the on-rate \(^{156}\) \(k_{\text{on}}\) is given by:

\[
\frac{1}{k_{\text{on}}} = \left( \frac{1}{k_0} + \frac{1}{4\pi D_0} \left( \frac{1}{R_{\text{contact}}} - \frac{1}{R_{\text{cavity}}} \right) \right) e^{\frac{\Delta G}{R T}} + \frac{1}{4\pi D_1 R_{\text{cavity}}} 
\]

Eq. 6.11

From this equation, it is simple to see that crowding impacts the association rate \(k_{\text{on}}\) in three different modes. First of all, in the crowded solution, the two molecules of interest diffuse one in respect to the other with a diffusion coefficient \(D_1\), which is reduced compared to the one observed in pure solvent \((D_0)\) because of the impact of viscosity (compare section 3.4). This
contribution results in a slow-down of the reaction rate. The second element is that depletion of
crowding agents in the neighborhood of the two molecules of interest give rise to a cavity with
radius $R_{cavity}$ where the two reactants experience a different diffusion coefficient than the one of
the crowded solution. In a depleted solution, this diffusion coefficient can be assumed to be the
one of dilute solutions in absence of crowding agents, $D_0$. This contribution results in a speeding
up of the reaction rate. Finally, a square-well potential arises in the cavity because of depletion
interactions, favoring the attraction between the two molecules and increasing the association
rate coefficient. The depth of the potential can be approximated using the depletion interaction
free energy, $\Delta\Delta G$, as described in Eq. 6.4 and 6.5 $^{165}$. This model provides a good quantification
of experimental data (Fig. 6.6C), comprising a non monotonic trend observed in mixtures of
large polymers where the association rate first increases at low crowder concentration and then
decreases with increasing concentration$^{153}$, which in the theory is a signature of cavity formation
effects$^{156}$. 
**Figure 6.6.** Quantification of crowding effects on disordered proteins *in vitro* and *in living cells*. A-C. Zosel et al. used single-molecule FRET to quantify the impact of crowding on the binding of two disordered proteins, ACTR and NCBD, in presence of PEG molecules at various crowder concentrations and for different degree of polymerization of the PEG molecule. Small crowders have a smaller impact than larger crowders on the binding equilibrium (A), microviscosity measured as $D_\eta/D_1$ (B), and on the association rate (C). Image adapted from [153].

Investigation of crowding effects by modulating the cell volume via changes in the osmolarity of the cell medium (D) and temperature (E). Ensemble FRET provides a readout that can be attributed to shift in folding stability of the protein (left column) or conformational changes of the unfolded state (right column). The comparison between the theoretical predictions (solid lines) supports the case for compaction of the disordered state [166]. F-I. In cell single-molecule FRET [114]. Microinjection enables delivering controlled concentrations of labeled molecules in the cell for single-molecule detection. Changes in osmolarity of the extracellular medium enable modulating the impact of cellular crowding on diffusion (G), conformations (H), and reconfiguration time (I) of the disordered protein Prothymosin alpha.

### 6.5 Summary and Outlook

The last ten years have provided fundamental insights on the sequence properties encoded in disordered proteins that control their conformations and dynamics as well as their propensity to assemble in condensates. The advancements made in describing the conformations and interactions of disordered proteins in crowded solutions support the importance of a systematic investigation of these effects across multiple concentrations for different degrees of polymerization of the same polymer and as a function of different polymer chemistry. Similarly, the investigation of crowding effects as a function of temperature (Fig. 6.6D-E) and
pressure\textsuperscript{167–170} provides additional parameters that can be contrasted with theory and simulations. This is essential to enable a robust test of theoretical expectations, whether they are from scaled particle theory, Flory-Huggins models, or depletion theories. The example in Fig. 6.6E well describes the inherent challenge of interpreting cellular results due to the limited manipulation that can be performed in living cells and the importance of accurate models. Modulating both cellular osmolarity and temperature enables to observe changes in the signal, but only the comparison with model predictions enable discriminating changes in the collapse of the unfolding state from a shift of the protein melting temperature.

Studies of the interaction between disordered proteins and synthetic polymers are also of high relevance for all those experiments that use synthetic polymers as mimicker of the cellular crowded environment to favor phase separation of proteins \textit{in vitro}. Interestingly, recent experiments found that crowding alters the material properties of \textit{in vitro} reconstituted biomolecular condensates\textsuperscript{171}.

While we described here results that support a better understanding on how the disordered ensemble responds to crowding, new models are needed to integrate these new observations in the context of folding (for what concerns the unfolded state), folding upon binding (for what concerns disordered proteins), and phase separation. In this respect, atomistic and coarse-grained simulations offer a powerful tool to test theoretical expectations and interpret experimental results.

The current advancements in the understanding of crowding effects on IDRs have been made possible by the development and refinement of both experimental and computational methods aimed to address the complexity of the conformational ensemble of disordered proteins. The same approaches have started to provide quantitative readouts on protein conformations,
stability, and binding not only in synthetic crowded solutions, but also directly in living cells, paving the way to use the cell as the definitive “test tube” for exploring biochemistry where it happens.

Acknowledgements. A.S. would like to thank Priya Banerjee, Simon Ebbinghaus, Jeetain Mittal, Gary Pielak, Rohit Pappu, Tom Record, and Ben Schuler for the insightful discussions on the effects of crowding on proteins over the years.

Keywords: proteins, polymers, colloids, intrinsically disordered proteins, crowding,

Conflict of Interests.
All authors declare that they have no conflicts of interest.
6.6 References


106. Zheng, W. et al. Inferring properties of disordered chains from FRET transfer


120. Ioan, C. E., Aberle, T. & Burchard, W. Light Scattering and Viscosity Behavior of


Chapter 7

Excluded volume and weak interactions in crowded solutions modulate conformations and RNA binding of an intrinsically disordered tail.

This chapter is adapted from: Madison A. Stringer+, Jasmine Cubuk+, J. Jeremías Incicco, Debjit Roy, Kathleen B. Hall, Melissa D. Stuchell-Brereton, Andrea Soranno (in submission to Journal of Physical Chemistry B)

+:Equal contribution

Author’s contribution. M.A.S. and J.C. equally contributed to the work. M.A.S collected the single-molecule data for the conformations of the NTD-RBD_L and collected the nsFCS data. J.C collected the RNA-binding data and the temperature dependence data. M.A.S. and J.C. performed the experiments. D.R. and J.J.I. provided research tools. M.A.S., J.C., J.J.I., and A.S. analyzed the data. M.D.S.B., K.B.H, and A.S. supervised the experiments. A.S. designed research. M.A.S., J.C, J.J.I, K.B.H, M.D.S-B. and A.S wrote the manuscript.
7.1 Abstract

The cellular milieu is a solution crowded with a significant concentration of different components (proteins, nucleic acids, metabolites, etc.). Such a crowded environment affects protein conformations, dynamics, and interactions. Intrinsically disordered proteins and regions are particularly sensitive to these effects. Here, we investigate the impact on an intrinsically disordered tail that flanks a folded domain, the N-terminal domain (NTD) and RNA Binding Domain (RBD) of the SARS-CoV-2 Nucleocapsid protein. We mimic the crowded environment of the cell using polyethylene glycol (PEG) and study its impact on protein conformations using single-molecule Förster Resonance Energy Transfer (FRET). We found that high molecular weight PEG induces a collapse of the disordered N-terminal tail, whereas low molecular weight PEG induces a chain expansion. Our data can be explained by accounting for two opposing contributions: favorable interactions between the protein and crowder molecules, and screening of excluded volume interactions. We further characterized the interaction between protein and RNA in the presence of crowding agents. While for all PEG molecules tested we observed an increase in the binding affinity, the trend is not monotonic as a function of the degree of PEG polymerization. This points to a role of non-specific protein-PEG interactions on binding in addition to the entropic effects due to crowding. To separate the enthalpic and entropic components of the effects, we investigated the temperature dependence of the association constants in the absence and presence of crowders. Finally, we compared the effects of crowding across mutations in the disordered region and found that the three-fold difference in association constants for two naturally occurring variants of the SARS-CoV-2 Nucleocapsid protein is reduced to almost identical affinities in the presence of crowders.
Overall, our data provide new insights into understanding and modeling the contribution of crowding effects on disordered regions, including the impact of interactions between proteins and crowders and their interplay when binding a ligand.
7.2 Introduction

At variance with the dilute conditions in which the majority of biochemical experiments and simulations are conducted, proteins perform their function within crowded cells in the midst of a myriad of other molecules. In eukaryotic cells, the total concentration of molecules can account for \( \sim 10\% \) of the overall volume \(^1^4\), significantly reducing the available space that each molecule can explore. This phenomenon is commonly referred to as crowding. In addition to pure excluded volume effects, the high concentration of components within the cell gives rise to non-negligible “weak” interaction effects, often referred to as quinary interactions \(^5^8\). As a result of these effects, the crowded cellular environment impacts protein folding stability, conformations, dynamics, diffusion, and interactions with ligands \(^5^9^\text{–13}\). Moreover, both crowding effects and quinary interactions are likely crucial within the context of biomolecular condensates, which select and concentrate a specific subset of components to very high concentrations (if not for the scaffolding elements, at least for the host components) \(^5^\text{,14–16}\). Quantifying and modeling such contributions is essential to better understand how protein function and interactions are modulated by the cellular environment.

Lacking a stable 3D-structure, intrinsically disordered proteins (IDPs) and regions (IDRs) are particularly susceptible to changes in the surrounding milieu \(^1^4\). Studies with synthetic crowders have revealed that IDP conformations tend to adopt more compact ensembles in denser environments \(^1^4,1^7^\text{–19}\). The precise degree of compaction depends on the specific polymer properties of the protein and of the synthetic crowder \(^1^4\). Similar observations were made also in the context of simulations \(^1^7,2^0,2^1\). Crowding effects apply not only to the conformations of disordered proteins, but also to their interactions with ligands, where entropic forces modulate both equilibrium and kinetics of binding \(^2^2\). The same effects can be observed in vitro, using
synthetic crowders \(^22\) or cell extract \(^{23,24}\), and in living cells \(^{18,19}\). As for conformations, these effects have been rationalized using polymer theories that account for screening of excluded volume and depletion effects \(^{25–28}\).

Previous efforts have contributed to building a general theoretical framework for understanding the impact of crowding on disordered proteins, linking the protein response to fundamental properties of the sequence, such as the scaling exponent. While the scaling exponent reflects the monomer-solvent and monomer-monomer interactions, and therefore depends on the primary structure of the protein, different disordered proteins may attain the same scaling exponent through completely different sequence compositions \(^{29–31}\). Sequence specific effects are relatively unexplored, but they can encode for nonspecific affinities for crowding agents, modulate conformations across different length scales, and even introduce transient secondary structure motifs. Furthermore, the contribution of macromolecular crowding has been previously tested only for full-length IDPs. Little is known about the impact of crowding on disordered regions flanking folded domains: different effects may occur when a disordered region is attached to one or more folded domains, including stabilization of interactions between the disordered region and the folded interface.

Here, we studied the effects of crowding on the disordered N-terminal domain (NTD) of the SARS-CoV-2 Nucleocapsid protein (Wuhan-Hu-1 variant) when attached to the contiguous RNA Binding Domain (RBD) (Fig. 7.1, Supplementary Table 7.1 and 7.2). This is a well characterized system where we previously found that the N-terminal tail is disordered, flexible, and retains its dynamics even when complexed with nucleic acids \(^{32,33}\). To investigate the effects of crowding we opted for single-molecule Förster Resonance Energy Transfer (FRET), which
provides a powerful toolbox to quantify the impact of crowding on protein conformations, dynamics, and diffusion, both in presence and in absence of ligands \(^\text{14,18,22,34,35}\). To monitor the conformational changes in the disordered tail, we used a previously designed construct \(^\text{32,33}\) that comprises both NTD and RBD and carries cysteine mutations for introducing fluorescent dyes at position 1 and 68 (Fig. 7.1, Supplementary Table 7.2). Because the labeling positions probe the NTD, we will refer to the construct as NTD\(_L\)-RBD. The construct has been labeled with Alexa Fluor 488 and 594, as previously described \(^\text{32,33}\). Using polyethylene glycol (PEG) as a model synthetic crowder, we set out to investigate how the presence of crowder molecules impacts protein conformations and its binding to RNA.

![Figure 7.1. N-terminal domain of SARS-CoV-2 and RNA-Binding-Domain (NTD-RBD). SARS-CoV-2 Nucleocapsid protein comprises an N-terminal domain (NTD, in magenta), an RNA Binding Domain (RBD, in lilac), a linker region (LINK, in gray), a dimerization domain (DIMER, in gray), and a C-terminal domain (CTD, in gray). In the present study we focus on an NTD-RBD truncation variant (magenta and lilac domains). Positions of fluorophores are reported in yellow over an all-atom van der Waals surface representation of the NTD-RBD construct.](image)

### 7.3 RESULTS.

We started by investigating the conformations of the protein under dilute conditions (no crowding). Single-molecule FRET measurements of the NTD\(_L\)-RBD protein in HEPES buffer (see Materials and Methods) provided a distribution of transfer efficiencies characterized by a
single shot-noise-limited population with a mean value of $0.698 \pm 0.006$ at $23^\circ$C, consistent with
the previously reported value of $0.709 \pm 0.009$ obtained in Tris buffer under analogous solution
conditions $^{32,33}$. As previously discussed, this single population represents a flexible and dynamic
tail. Having determined a reference state of the protein structural ensemble under dilute solution
conditions, we set to quantify the impact of crowding on protein conformations. To this end, we
studied the effects as a function of polymer concentration and length, which previously have
been shown to provide a robust tool to discriminate across theoretical frameworks $^{14,15}$. In
particular, polymer “length” determines the type of response that a test chain (e.g., our protein)
will perceive in a bath of other chains (e.g., PEG). When the degree of polymerization (number
of monomers) $P$ of the crowders is larger than the degree of polymerization $N$ of the test chain,
the test chain compacts and adopts conformations close to the one of the ideal state. Flory’s
criterion $^{36}$ provides a quantitative reference for this crossover regime, which is: $P > N^{1/2}$. For
the system in this study, this threshold approximately coincides with PEG 1000 (see
Supplementary Information). With this criterion in mind, we started quantifying the impact of
large and small crowders on protein conformations.

7.3.1 Large crowders compact protein conformations. First, we investigated the
contribution of high molecular weight PEG polymers. We identified a series of PEG molecules
with average molecular weight equal to or larger than 1000 Da and with sufficient purity (lack of
fluorescent contaminants) to allow for single-molecule detection: PEG 1000, PEG 2050, PEG
3350, PEG 4600, PEG 6000, PEG 8000, PEG 35000 (see Supplementary Table 7.3). For each
molecular weight, we explored the impact of increasing concentrations of PEG up to 20% of
weight per volume (w/v), a similar concentration to the volume fraction occupied in a crowded
cellular environment $^{15}$. 343
Figure 7.2. Impact of crowders on IDR conformations. A. Transfer efficiency distributions as function of crowder concentrations (weight per volume, w/v, in g/ml) for large crowders (average molecular weight larger than 600 Da). n indicates the number of bursts per histogram. B. Transfer efficiency distributions as a function of crowder concentrations (weight per volume, w/v, in g/ml) for small crowders (average molecular weight smaller than 1000 Da). PEG1000 is shown as a reference for comparison with the case of larger crowders. n indicates the number of bursts per histogram. C. Plot of measured transfer efficiencies for the NTD_{L}-RBD in the range between 0 and 20 % (w/v). Lines are guides for the eyes.

Inspection of the transfer efficiency distributions, after selection of the donor-acceptor population (see Supplementary Information), reveals under all conditions one single population (Fig. 7.2A-B). Changes in mean transfer efficiency with increasing crowding
concentrations reflect conformational changes occurring in the protein. Interestingly, for PEG 1000 we detected no substantial shift in the mean transfer efficiency at any of the concentrations tested. In contrast, for all the other large molecular weights of PEG (PEG 2050, PEG 3350, PEG 4600, PEG 6000, PEG 8000, PEG 35000), we observed a clear shift of the mean transfer efficiency toward higher values (Fig. 7.2A,C), indicating a compaction of the IDR. Note that the change in refractive index of the solution can contribute to a shift toward lower transfer efficiencies, de facto counteracting the collapsing effect of the crowding agents.

We further compared the degree of compaction and its dependence with the molecular weight of PEG. This is best captured by comparing the variation in transfer efficiency at a constant concentration of crowders, such that the number of PEG monomers in solution is approximately constant, and all the differences in conformations are due to the changes in the PEG chain length. Analysis of the results at 15% w/v (Supplementary Figure 7.1 and Fig. 7.3C) shows that increasing the molecular weight of PEG above 1000 Da results in compaction of the disordered tail, approaching a saturation limit of the degree of compaction at molecular weights higher than 8000 Da (Supplementary Figure 1 and Fig. 3C). This is consistent with previous observations for other disordered proteins 14.

7.3.2 Small crowding agents expand protein conformations. One notable difference from previous results 14, where a monotonic increase on the strength of compaction was found with increasing molecular weight, is that we do not observe any change in transfer efficiency for PEG 1000. This observation suggests that small crowding agents (≤ 1000 Da) may have no impact on the compaction of the NTD region. Therefore, we investigated this case scenario by measuring the impact of Ethylene Glycol (EG), PEG 400, and PEG 600. Here, we observed that PEG 600 behaves similarly to PEG 1000, whereas for all crowding agents smaller than PEG 600
the mean transfer efficiency does not shift toward higher values, but instead moves toward lower ones (Fig. 7.2B). The lower the molecular weight, the larger the shift toward lower transfer efficiencies. A shift toward lower values of this extent exceeds the changes due to the refractive index of the solution and implies an expansion of the tail. This rules out that small crowding agents have no effect on protein dimensions and suggests that these smaller PEG molecules are interacting with the disordered tail. We further compared the change in mean transfer efficiency at 15% w/v concentration (Fig. 7.2C, 7.3C and Supplementary Fig. 7.1), and we noticed that the smallest molecular weight molecule, EG (the constitutive monomer of PEG), has the largest impact and expansion on the chain, whereas increasing the molecular weight leads to a decrease of chain expansion. This further supports the idea that two different forces are at balance: one attractive that leads to chain expansion for low molecular weights and one repulsive that leads to chain compaction for high molecular weights of PEG. Length of the polymeric crowder seems to determine one scenario or the other.
**Figure 7.3. Degree of compaction as a function of crowding concentration.**

**A.** Conversion of the mean transfer efficiencies in root-mean-square interdye distance for all the molecules. The dashed line represents the value in absence of crowders. Solid lines are the global fit to the model described in the Supplementary Information. **B.** Dependence of the interdye root mean square distance at a fixed concentration (15% w/v) as a function of the PEG molecular weight. The dashed line represents the value in absence of crowders. The solid cyan line is the expectation of conformational changes at this concentration of PEG for the effect of pure excluded volume screening. The measured values depart from the predicted trend suggesting the occurrence of additional interactions. The blue line is a fit to the model described in the Supplementary Information and reported in panel C and D. **C.** Estimate of the excess interaction not accounted for by the pure excluded volume screening as a function of the PEG concentration. The trend is fitted to the Schellman weak binding model (as previously proposed for small solvent molecules). **D.** The number of binding sites associated to the Schellman weak binding model (γ) decreases with increased polymer molecular weight, with a characteristic molecular weight of approximately 1200 ± 100 Da. This observation is qualitatively consistent with previous observations from the Record lab, where the PEG ends are found to have different preferential interaction than the backbone of PEG.  

### 7.3.3 The disordered tail remains dynamic under crowded conditions.

To rule out that the observed compaction or expansion is due to folding of the IDR or specific interactions between the IDR and the folded RBD, we performed nanosecond Fluorescence Correlation Spectroscopy (nsFCS) measurements, which report on chain dynamics. Based on previous observations, we expect anticorrelated behavior in the cross-correlation of donor-acceptor photons over time due to fast chain dynamics in the 100 ns time scale. This
anticorrelated behavior reflects the intrinsic anticorrelated nature of FRET, where an increase in acceptor emission requires a decrease in donor emission and vice versa. Measurements for PEG 400 and PEG 3350 at 15% w/v are reported in Supplementary Fig. 7.2. For all measured samples, we found that the donor-acceptor cross-correlation is anti-correlated, confirming fast conformational dynamics in the disordered region between the fluorophores. The donor-donor and acceptor-acceptor correlations and the donor-acceptor cross-correlation can be described by a global fit where the correlation time is shared (Supplementary Table 7.4), supporting that the same dynamics are captured across the three correlations. The correlation time is approximately 120 ns, in agreement with previously reported values in the absence of crowders. Overall, our data confirm that in both scenarios, whether the crowding agents cause a shift toward higher or lower transfer efficiencies, the NTD remains disordered and flexible.

### 7.3.4 Balancing excluded volume screening and local interactions.

To try to make sense of the experimental observations, we conceptualized the observed phenomenon in terms of a polymer model.

The theory of coil-to-globule transition is a useful model to compute the impact of solution conditions on the conformation of a disordered protein. We used this model to quantify the solvation free energy change upon mixing of the IDR with each crowded solution.

Previous experiments have also supported the use of a renormalized-group polymer theory of ternary solutions to capture the compaction of IDPs in the presence of crowding polymers. The model accounts for screening of a single chain excluded volume (the one associated with the protein) by the excluded volume of other chains (the PEG molecules). It is called renormalized because it accounts for the different length-scales at play, including the
characteristic size of the protein and the one of the crowding polymers. We used this model to estimate the excluded volume screening contribution to the free energy of solvation.

Finally, we accounted for the polymeric nature of PEG molecules and for the fact that above a critical concentration, PEG molecules start to overlap (overlap concentration \(42\), see \textbf{Supplementary Fig. 7.3}). The characteristic length-scale of the polymer solution is no longer represented by the radius of gyration of PEG, but by the mesh size of the solution, due to overlapping PEG molecules. This regime is commonly referred to as the semi-dilute regime, and PEG molecules longer than 50 monomers (~2500 Da) can easily enter such a regime at volume fractions lower than 5\% \(^{15}\) (\textbf{Supplementary Fig. 7.3}).

A detailed mathematical description of the model is provided in the \textbf{Supplementary Information}, but can be summarized as:

\[
R_g(N, P, \phi, \epsilon) = R_{g0} f(N, P, \phi, \epsilon)
\]

\textbf{Eq. 7.1}

where \(R_{g0}\) is the radius of gyration of the protein in absence of PEG, \(N\) and \(P\) are the degree of polymerization of protein and PEG, \(\phi\) is the volume fraction occupied by the polymer, and \(\epsilon\) is the effective interaction energy between monomers after subtracting the contribution of excluded volume screening effects. The term effective is used to highlight that \(\epsilon\) accounts for several factors, including the contribution of solvent and co-solutes.

Since \(R_{g0}, N, P, \phi\) can be derived from known parameters (the length of the protein sequence, the conformations adopted by the protein in absence of PEG, and PEG solution concentrations), we can use the model to directly extract the \(\epsilon\) parameter as a function of PEG concentration and molecular weight (\textbf{Fig. 7.3B}). The change in the interaction parameter
between any crowded condition and the uncrowded reference state, \( \Delta \varepsilon = \varepsilon - \varepsilon_0 \), is directly related to the change in the solvation free energy \(^{41}\). \( \Delta \varepsilon \) can be fitted to the Schellman weak binding model \(^{43,44}\) according to \( \Delta \varepsilon = \gamma \ln(1 + Kc) \), where \( \gamma \) is the average number of binding sites, \( K \) is a binding constant, and \( c \) is the molar concentration of PEG monomers in solution. Fitting results are reported in Supplementary Table 7.5. A plot of \( \gamma \) as a function of PEG molecular weight (Fig. 7.3D) reveals two interesting features. First, the number of binding sites associated with the larger PEG molecules is negligible (close to zero), which suggests that the renormalization group theory well captures this regime. Second, \( \Delta \varepsilon \) is the largest for EG and decreases with increasing molecular weight (Fig. 7.3B). Data can be fitted to an empirical exponential decay, \( \gamma = \Delta \gamma \text{Exp}( - P/P'_c ) + \gamma_\infty \), with \( P'_c \) representing a “correlation molecular weight” (900 ± 200 Da) over which the effect of chemical interaction between the protein and the PEG end-groups becomes negligible.

Alternatively, we can assume that the terminal ends and the backbone of PEG interact differently with the protein and account for the fact that at constant volume fraction, the concentration of PEG terminal groups decreases with increasing the degree of polymerization (Eq. S7.15 in Supplementary Information). The resulting fit captures the trend of experimental data across all concentrations (Fig. 3A) and provides a global quantification of the corresponding interactions (Supplementary Table 7.6). Our observations further support that the end-groups of PEG interact with different residues compared to the backbone of PEG and result in a stronger conformational change.
Figure 7.4. Crowding impact on the NTD-RBD and (rU)$_{40}$ association constant. **A.** Examples of titrations of (rU)$_{40}$ on NTD$_L$-RBD as a function of a fixed concentration of PEG (15% weight/volume) for different PEG molecular weights. Color scheme is provided in the legend. **B.** Schematic of depletion theory. Two colloidal particles (purple and cyan circles) in a polymer solution are surrounded by a depletion layer (dashed blue line), where polymers (purple coils) do not enter. Bringing the two colloidal particles together gives rise to an overlap between the two depletion layers, which is maximized when colloidal particles are in contact. The larger the overlapping volume of the depletion layers, the stronger the attractive entropic force between the two particles. For the case treated in this work, the NTD-RBD is described as a colloidal particle (accounting for the presence of a folded domain), whereas RNA is represented by a smaller effective radius since PEG and RNA can interpenetrate, resulting in a significantly smaller depletion layer than the colloidal particle. **C.** Upper panel. Association constant $K_{A0}$ (in absence of crowding) divided for the measured value $K_A$ in presence of 15% w/v PEG. The blue band is the estimate based on depletion theory when assuming a radius of gyration of the RNA equal to 0.8 ± 0.3 nm. Dashed line is the reference state. Central panel. Conversion of the association constant ratio to the corresponding $\Delta \Delta G$ as a function of the PEG molecular weight. Colored points represent measurements. Error bars are standard deviations of two independent sample repeats (see Table S7.6). The blue band is the estimate based on depletion theory when assuming a radius of gyration of the RNA equal to 0.8 ± 0.3 nm. Dashed line is the reference state. Lower panel. Expected values for the theory when assuming a fixed radius of gyration (as estimated for the protein in absence of crowding from simulations $^{33}$, $R_g = 2.1$ nm) and increasing molecular weight of PEG molecules that adopt a radius of gyration between 0.7 to 3.0 nm.
7.3.5 Crowder effects on binding affinity equilibrium. With the observation that the NTD-RBD sequence encodes for interactions with the crowder molecule and that this effect can be counterbalanced by excluded volume screening, we turn to investigate how these effects will be reflected on binding equilibria. We have previously shown that single-molecule FRET can be used to quantify binding of the NTD₇-RBD to RNA because it is accompanied by a shift in transfer efficiency, which reflects conformational changes of the tail. Here, we monitored the effects of crowding on the binding of NTD₇-RBD to (rU)₄₀. The specific length of the oligonucleotide has been chosen based on previous experiments to maximize the sensitivity of the transfer efficiency change for determining both protein affinity and conformations. For each molecular weight of PEG, we studied the change in binding affinity at a fixed concentration of 15% w/v by titrating NTD₇-RBD with increasing concentrations of (rU)₄₀ in presence and absence of the crowder (Fig. 7.4A). The change in binding affinity is expressed as $K_A / K_{A,0}$ and the corresponding change in binding free energy $\Delta \Delta G$ (Fig. 7.4C):

$$\Delta \Delta G = -k_B T \ln \frac{K_A}{K_{A,0}}$$

Eq. 7.2

In particular, we are interested in understanding what happens to binding affinity with respect to PEG1000, which acts as the divider between the regimes where attractive or screening interactions dominate.

We found that for crowding agents smaller than PEG 1000, there is an increase of the association constant (tighter binding) with increasing molecular weight of PEG, until reaching a maximum at approximately the molecular weight of PEG 600 (Fig. 7.4C and Supplementary Table 7.7). For molecules larger than PEG 600, the trend reverts, and the association constant decreases (weaker binding) with increasing polymer molecular weight. However, it does not decrease below the
value measured in absence of crowding agents. Our experimental observations suggest a non-trivial relation between protein conformations and binding. To better understand the source of this behavior, we turned to established models that have previously captured analogous behavior for disordered proteins and nucleic acids.22,35

7.3.6 Depletion effects can describe binding affinities. From the equilibrium constant, we estimated the $\Delta \Delta G$ associated with the crowding process and compared it to predictions from the depletion theory. We used a model where depletion forces are estimated for two colloidal particles in a solution crowded by one type of polymer (Fig. 4B). In this framework, the polymers in the solution are depleted from the surface of the colloidal particle within a given layer that is commonly referred to as the depletion layer. This layer reflects the fact that polymers cannot interpenetrate with the colloidal particle. When the two colloidal particles are brought in proximity, the depletion layers start to overlap, leading to an increase of the volume that the polymers can explore or, in other words, an increase in the entropy of the solution. As a result, the two colloidal particles in a polymer solution experience an attractive entropic force that depends on the radius of the colloidal particles, radius of gyration of the polymers, and on the concentration of polymers:

$$\Delta \Delta G = - nk_B T V_{\text{overlap}} (R_{g, \text{protein}}, R_{g, \text{RNA}}, R_{g, \text{PEG}})$$ Eq. 7.3

where $n$ is the concentration of polymer expressed in number density, $V_{\text{overlap}}$ represents the overlapping volume of depletion layers, and $R_{g, \text{protein}}, R_{g, \text{RNA}},$ and $R_{g, \text{PEG}}$ are the radius of gyration of the protein, RNA, and PEG, respectively. As for the case of protein conformations, we also accounted for overlapping of polymers under semi-dilute conditions. Mathematical details are presented in Supplementary Information (Eq. S7.19-22).
We assigned the radius of gyration of one particle to be equal to the value obtained from the simulations of the NTD-RBD. We further rescaled the radius to account for the small change in size observed with increasing molecular weight of PEG. Finally, we used the radius of gyration of the RNA as a model parameter and computed the case of a purely entropic contribution of PEG. We plot our predictions as a function of the molecular weight at a constant concentration of crowding agents of 15% w/v (Fig. 7.4C). For a large $R_g^{RNA}$ (3.0 nm), we found that $\Delta\Delta G$ is expected to decrease monotonically with increasing molecular weight; however, when decreasing the $R_g^{RNA}$ we recover a non-monotonic trend that is qualitatively compatible with our experimental observations. The experimental trend is better captured assuming an effective radius of the RNA in the range of 0.8 ± 0.3 nm. The reduced size of RNA is consistent with previous treatment of binding between two disordered proteins with the same theory, where only a segment of the disordered protein was proposed to sense depletion interactions. This can be explained by accounting for the fact that polymer crowders and disordered proteins (or RNA, in the current case) can interpenetrate in the solution. Small effective radii of the RNA and PEG molecules (due to the mesh size) reduce the depletion layers and corresponding overlap volume, consequently reducing the entropic forces and resulting in the non-monotonic trend for long PEG molecules.

### 7.3.7 Temperature effects on protein conformations in absence of crowders.

The emerging picture from our experiments is of a balance of enthalpic and entropic forces that determine protein conformations. However, depletion theory suggests that entropic contributions alone can be sufficient to qualitatively describe the trend of association constants for binding to RNA.
To better quantify the contribution of the entropic and enthalpic components on the binding of NTD-RBD to RNA, we determined changes in the association constant as a function of temperature. To this end, we used a temperature-controlled cuvette holder, analogous to previous studies \(^{45-47}\), which allowed us to monitor the conformations of the NTD, in the context of NTD\(_L\)-RBD, at temperatures from 10°C to 56°C. Similar to other studies of the temperature dependence of disordered proteins \(^{45-47}\), we found that the FRET efficiency of NTD\(_L\)-RDB shifts toward higher values with increasing temperature from 10°C to 37°C, indicating a compaction of the disordered tail. Above 37°C we additionally observe a broadening of the distribution of transfer efficiencies, revealing another conformational change in the NTD\(_L\)-RBD. Measurements performed on the RBD\(_L\) construct (Supplementary Fig. 7.4) reveal a small deviation in transfer efficiencies that is not compatible with a complete unfolding of the RBD. Together with the conformational changes observed for NTD\(_L\)-RBD above 37°C, which indicate a compaction of the probed segment, we reason that this transfer efficiency change reflects the formation of an intermediate configuration of the RBD.

### 7.3.8 Enthalpic and entropic contributions to RNA binding in the absence of crowders

Having established a range of temperatures where the NTD\(_L\)-RBD shows a single population (10°C-37°C), we then proceed to study the temperature dependence of RNA binding in the absence of PEG. At saturating concentrations of (rU)\(_{40}\), we observe a single population up to 44°C and the appearance of a second population, shifted to higher transfer efficiencies, only above that temperature. This implies that RNA stabilizes the “low temperature” conformation of the NTD-RBD. (Fig. 75A-C, Supplementary Fig. 7.5). In addition, the mean transfer efficiency of the new population does not change upon addition of RNA, independently from the concentration, suggesting that the new conformational state is non competent for RNA binding.
and dissociation of RNA is required to adopt that conformation. To avoid the complications associated with this conformational change, we restricted our analysis of the association constant to temperatures equal and below 37°C (Supplementary Table 7.8).

According to Van’t Hoff equation, if in a given range of temperatures the enthalpy of a reaction does not change, the logarithm of the association constant $K_A$ is a linear function of the reciprocal of absolute temperature $T$, where the slope reports about the enthalpy $\Delta H$ and the intercept reports about the entropy $\Delta S$ of binding:

$$\ln K_A = -\frac{\Delta H}{R} \frac{1}{T} + \frac{\Delta S}{R}$$  \hspace{1cm} \text{Eq. 7.4}

where $R$ is the gas constant.

We observed a linear increase of $\ln K_A$ as a function of $1/T$, in the range of temperatures from 10°C to 37°C. Fitting of Eq. 7.4 provides $\Delta H$ and $\Delta S$ of $-7 \pm 2 \text{ kcal/mol}$ and $0.014 \pm 0.005 \text{ kcal/mol/K}$, respectively, for the binding of NTD$_L$-RBD to (rU)$_{40}$ in the absence of crowding agents (see Supplementary Table 7.9), indicating that the binding reaction is mildly exothermic and mostly enthalpically driven under this experimental conditions (Fig. 7.5J). These values provide the reference for understanding the impact of crowding agents on the entropy and enthalpy of binding.
Figure 7.5. NTD-RBD conformations and binding of (rU)$_{40}$ as a function of temperature. A-C. Upper panels. Representative transfer efficiency distributions of the NTD$_L$-RBD in 50 mM HEPES buffer (A), with addition of 20 nM (rU)$_{40}$ (B), and with addition of 500 nM (rU)$_{40}$ (C) as a function of temperature from 10 °C to 56 °C. Lower panels. Contour plots of the transfer efficiency distributions of NTD$_L$-RBD as a function of temperature. D-F. Upper panels. Representative transfer efficiency distributions of the NTD$_L$-RBD in 50 mM HEPES buffer supplemented with 15% w/v PEG 600 in absence of RNA (D) and in presence of of 2.5 nM (rU)$_{40}$ (E) and 12.5 nM (rU)$_{40}$ (F), as a function of temperature from 10 °C to 37 °C. Lower panels. Contour plots of the transfer efficiency distributions of NTD$_L$-RBD as a function of temperature for the corresponding experimental conditions in the upper panels. G. Temperature dependence of the interdye distance in absence of PEG (black) and with EG (red), PEG 600 (orange), and PEG 8000 (gray). H. Dependence of Δε(T)=ε(T)-ε$_0$(T) on the inverse of the temperature. J. Plot of the ln(K$_A$/M) as a function of 1/T (K$^{-1}$) in absence of PEG (black) and in presence of 15% w/v EG (red), PEG 600 (orange), and PEG 8000 (gray). Error bars represent errors of the fit. The linear fit reports about the enthalpy (slope) and entropy (intercept) of binding according to Eq. 7.4. K. ΔH, ΔS and ΔΔS obtained from the fit of the data in panel J.
7.3.9 Temperature dependence of protein conformations in crowded conditions. As a next step, we characterized the temperature response of the protein in presence of three archetypal crowders (EG, PEG 600, and PEG 8000) at a fixed concentration (15% w/v) (Fig. 7.5E-G, Supplementary Fig. 7.6-8). These specific molecular weights have been chosen because they reflect three distinct regimes: low molecular weight and no significant effect on $K_A$ (EG), low molecular weight and significant effect on $K_A$ (PEG 600), high molecular weight and no significant effect on $K_A$ (PEG 8000).

We found that, at a fixed volume fraction of crowding agents of 15% w/v, the protein conformations exhibit different temperature responses depending on the type of crowding agent (Fig. 7.5H). For the case of EG and PEG 600, the presence of the crowder appears to limit the extent of conformational collapse of the protein, resulting in an almost unchanged protein configuration. For PEG 8000, we observed an opposite trend, with the protein conformations expanding with increasing temperature instead of collapsing. These results are apparently discordant, with the overall protein conformation expanding or collapsing as a function of the crowding degree of polymerization.

To better understand the contribution of the forces at play, we estimated $\Delta \varepsilon(T) = \varepsilon(T) - \varepsilon_0(T)$ from Eq. 7.1, which separates the contribution of solvent and crowders interactions from excluded volume screening effects. We found that $\Delta \varepsilon$ increases with increasing temperature for all the three crowding agents (Fig. 7.5I), indicating that the specific temperature response of chain dimensions for the different degree of polymerization of PEG is due to excluded volume screening effects. We note that the temperature dependence of $\Delta \varepsilon$ is not surprising and does reflect change in the interaction of PEG and water molecules with the protein (number of sites on
the protein and binding constant), as implied by the Schellman weak binding model. An exact quantification of these effects would require comparison of different sequence compositions for the disordered tail, which is beyond the scope of this work.

Having characterized the effect of temperature on the conformations of the protein in presence of crowders, we proceeded to study how the association constant for RNA binding depends on temperature.

### 7.3.10 Enthalpic and entropic components to RNA binding in presence of crowders.

For all three crowding agents (EG, PEG 600, and PEG 8000), we found a linear dependence of the logarithm of the association constant as a function of the inverse of the temperature. As mentioned before, the slope of this linear dependence reports about the enthalpic contribution to binding. In this respect, while binding remained exothermic in the presence of all three crowders at 15% w/v, there were clear differences in the magnitude of their effects (Fig. 7.5J-K). Interestingly, although PEG 600 favors binding more than EG and PEG 8000, it exerts a negligible effect on the enthalpy ($\Delta H_{\text{exp}} = -5 \pm 5$ kcal/mol, Supplementary Table 7.10), indicating that the contribution introduced by this crowding agent to binding is purely of entropic nature. The experimental change in the entropy of binding induced by addition of PEG 600, $\Delta S_{\text{exp}} = -0.02 \pm 0.02$ kcal/mol/K, (Supplementary Table 7.10) is compatible within the large error with the value from depletion theory (assuming an effective radius of gyration of 0.8 ± 0.3 nm for RNA, see Depletion effects can describe binding affinities), $\Delta S_{\text{dt}} = 0.002 (\pm 0.006)$ kcal/mol/K. PEG 8000, on the other hand, the crowder of the three that impacts binding the least, presents a more significant change in enthalpy ($\Delta H_{\text{exp}} = -6 \pm 4$ kcal/mol, Supplementary Table 10), but this favorable effect on binding is completely compensated by a negative change in the entropy of binding $\Delta S_{\text{exp}}, -0.02 \pm 0.01$ kcal/mol/K. In this case, the measured change in entropy
is of opposite sign and greater in absolute value than what is predicted by depletion theory, $\Delta S_{dt} = 0.0029 \pm 0.001\text{ kcal/mol/K}$. Lastly, at variance with PEG 600 and PEG 8000, EG shows a significant change in both enthalpy and entropy, with $\Delta H_{exp} = -14 \pm 3\text{ kcal/mol}$ and $\Delta S_{exp} = -0.05 \pm 0.01\text{ kcal/mol/K}$ (Supplementary Table 7.10), also in opposite direction to the estimated value from depletion theory of $\Delta S_{dt} = 0.0014 \pm 0.0003\text{ kcal/mol/K}$, and displaying a strong compensation between enthalpic and entropic components of their effects on the binding of NTD-RBD to (rU)$_{40}$ (Supplementary Fig. 7.10). The result is robust with respect to the fitting procedure used to estimate the association constants as a function of temperature (Supplementary Fig. 7.11 and Supplementary Table 7.10). Overall, our data suggest that whereas depletion theory can capture the overall trend of the equilibrium constants for this system, the origin of the observed phenomenon is more complicated than the entropic explanation proposed by the model.

### 7.3.11 Sequence specificity.

We have previously shown that mutations characteristic to the Omicron variant of the SARS-CoV-2 do not alter the conformations of the disordered tail, compared to the Wuhan-Hu-1 variant, but lead to a 4-fold change in binding affinity for poly(rU)$_{33}$. Here, we asked whether crowding effects can modulate the binding affinity of these two different variants. In other words, we asked whether the “functional” differences encoded in the sequence are maintained under crowded conditions. While this in first approximation seems a reasonable assumption, there is the possibility that conformational changes and depletion effects induced by crowders may normalize the contribution of mutations.
Figure 7.6. Effects of crowding on the Omicron variant. A. Distribution of transfer efficiencies for wildtype (Wuhan-Hu-1) and Omicron variants in absence of crowding. Vertical dashed line provides a reference for comparing the mean transfer efficiency. B. Corresponding fraction bound and unbound based on titrations of (rU)₄₀ in absence of crowding and fit to Eq. S7.17. Errors are from multiple independent repeats of the same condition. C. Distribution of transfer efficiencies for wildtype (Wuhan-Hu-1) and Omicron variants in presence of crowding. Vertical dashed line provides a reference for comparing the mean transfer efficiency. D. Corresponding fraction bound and unbound based on titrations of (rU)₄₀ in presence of 15% w/v PEG 600 and fit to Eq. S7.17. Errors are from multiple independent repeats of the same condition.

To this end, we compared the binding of the two variants in absence and presence of 15% w/v PEG 600 (Fig. 7.6, Supplementary Fig. 7.12). We selected PEG 600 because in our previous binding experiments this crowding agent showed the strongest effect on $K_A$. In the HEPES buffer and no-crowders condition, we found that the association constants for (rU)₄₀ of the Wuhan variant was $2.7 \pm 0.3$ times the value for the Omicron variant, with $K_A^{WUHAN}$ and $K_A^{Omicron}$ equal to $0.097 \pm 0.006$ nM⁻¹ and $0.036 \pm 0.004$ nM⁻¹, respectively. Interestingly, when
measured in presence of 15% w/v PEG 600, $K_A^{\text{Wuhan}}$ is equal to $0.32 \pm 0.04 \text{ nM}^{-1}$ and $K_A^{\text{Omicron}}$ is $0.46 \pm 0.06 \text{ nM}^{-1}$. Therefore, addition of crowding reduces the change of affinity between the two variants to a factor of $0.7 \pm 0.1$. This suggests that crowding effects can effectively “normalize” some of the interactions encoded in disordered regions and that the chain response will be effectively different depending on the environmental context.

### 7.4 DISCUSSION.

#### 7.4.1 Crowding and IDR conformations. Our results expand on previous investigation of full-length intrinsically disordered proteins by studying the impact of crowding on a disordered tail attached to a folded domain. Within the range of occupied volume fractions that we explored, we did not observe stabilization of contacts with the folded domain, as confirmed by nanosecond FCS measurements. While a stable conformation is not adopted, the disordered region definitely was found to be sensitive to the surrounding environment. The emerging picture is of a competition between weak-binding interactions that alter the free energy of solvation of the protein and the excluded volume screening effects due to the spatial hindrance of other crowders. The key parameters controlling protein conformations are the degree of expansion of the protein, the “size” (degree of polymerization and scaling exponent) of the crowding agents, and the extent of interactions between crowding agents and protein. The degree of expansion (or compaction) of the protein is measured with reference to the dimensions of the “ideal state” of the protein, which can be estimated based on scaling exponent measurements or simulations. The Flory criterion provides the reference to define large ($P > N^{1/2}$) and small ($P < N^{1/2}$) crowders when compared to the length of the disordered protein. This is consistent with approximately the size of PEG 1000 (Supplementary Information). In this framework, the protein conformations
in the presence of small polymeric crowders sense a small to negligible contribution of the screening excluded volume effect. If there are weak attractive interactions of the protein to the crowders, they can modulate the conformations of the protein. This modulation can occur in one of two ways where either the interaction with the crowders can increase the solvation free energy of the chain, therefore leading to an expansion (as in the case studied here), or decrease the solvation free energy and therefore lead to a collapse. Another scenario can occur if there are weak repulsive interactions between the protein and the crowders. Repulsive interactions between the protein and the crowder will result in an amplification of the excluded volume screening effects, as predicted by polymer theories of ternary solutions \(25-27\) and shown in previous experiments \(^{14}\). In the case of large polymeric crowders, the contribution of excluded volume screening becomes significant and can drive compaction of the chain if the chain is more expanded than its ideal state conformation \(25-27\). Weak attractive interactions can counterbalance in some measure the effect of screening if the protein-crowder local interactions favor a better solvation of the polymer chain, which is the case for the system in this study.

**7.4.2 Crowding and IDR interactions.** In our experiment we tested the impact of a fixed concentration of crowding on binding when modulating the degree of polymerization of PEG. We found a non-monotonic behavior of the association constant of NTD-RBD for RNA (Fig. 7.4), with the strongest effect (favoring association) detected for PEG 400 and PEG 600. We compared experimental estimates of the change in free energy with expectations from depletion theory and found that a purely entropic contribution can explain the non-monotonic trend. We took a step forward and verified whether this interpretation holds true by quantifying the enthalpic and entropic contributions to binding from the Van’t Hoff analysis of the temperature dependence of the association constant. We found that the predicted entropic changes, based on
the depletion theory, are smaller than the one observed in the experiments. This is not completely surprising, given that the version of the model adopted here operates various simplifications in the description of the protein and, in particular, of the RNA. The enthalpic contribution is truly negligible only for PEG 600 and, within errors, for PEG 8000, whereas it becomes dominant for EG. We speculate that this is due to the interactions of EG with both the protein and RNA, bridging the two molecules together. The difference in binding affinity of Wuhan-Hu-1 and Omicron variants in 15% w/v PEG 600 solution further suggests that few mutations in the sequence are sufficient to alter the entropic and enthalpic contributions to binding, pointing to sequence specific encoded interactions.

7.4.3 PEG interactions. Our results reveal weak attractive interactions between the disordered tail and low molecular weight PEG molecules. The existence of interactions between protein (or nucleic acids) and PEG has been reported before in literature\(^9,13\), leading even to recommendations against the use of PEG as a crowding agent because it is not inert\(^9\). Indeed, weak interactions complicate the interpretation of ensemble experiments, where different protein (or nucleic acid) conformations cannot be resolved. However, single-molecule measurements\(^{14,22,34,35}\) and carefully designed ensemble experiments\(^ {5,48}\) can leverage such interactions to explore the interplay between purely entropic effects and enthalpic contributions. In this respect, the use of cytomimetic solutions based on polymers and other solutes can provide physical insights on the properties of the cellular medium\(^ {49}\). Different from previous single-molecule experiments focused on the folding of nucleic acids\(^ {34,35}\), we observed enthalpic contributions on both conformations of the proteins and interactions of the protein with RNA that could not be simply described in terms of depletion effects. The dependence of the interaction on the molecular weight of PEG molecules was similar to recent determinations of enthalpic and
entropic crowding effects on a foldable protein \(^{48}\), with enthalpic effects dominating low molecular weights of PEG. Interestingly, we found compensating effects of enthalpic and entropic effects on RNA binding, which counterbalance each other. We speculate that the dynamic nature of the complex facilitates a tradeoff between increased enthalpic interactions, for example mediated by short PEG, and entropic changes. Finally, it is worth mentioning that the use of the Schellman weak binding model implicitly accounts for the dynamic exchange of water molecules with crowding agents in the proximity of the protein. Indeed, addition of osmolytes are known to alter the water activity and contribute entropically to protein conformations, interactions, and folding\(^{50,51}\).

7.4.4 Implications for SARS-CoV-2. Finally, it is interesting to consider the implications of our experiments in the context of SARS-CoV-2. We found that the Nucleocapsid NTD\(_L\)-RBD and RBD\(_L\) undergo a cooperative conformational change in absence of RNA above 37 °C, with NTD\(_L\)-RBD adopting a more compact conformation and RBD\(_L\) slightly expanding. By comparison with previous states identified in denaturation experiments\(^{33}\), we reason this change is linked to interactions of the tail with an intermediate (partially unfolded) state of RBD. Interestingly, binding of single-stranded nucleic acid favors the stability of the open configuration (folded RBD) and the transition toward the intermediate state is shifted toward higher temperatures. This supports a picture where the NTD-RBD region of the SARS-CoV-2 Nucleocapsid protein not only has a preference for the binding of single-stranded RNA, but the binding of RNA also stabilizes specific conformations. The comparison between variants provides further insights on the impact of evolutionary mutations in the protein sequence. The small (yet measurable) difference of the association constants between the Wuhan-Hu-1 and the Omicron variants in presence of PEG suggests that the specific mutations are largely tolerated by
the virus because, under crowded conditions, they result in a similar affinity. However, our observations open the door to a more complex interpretation of the role of disordered regions in the cellular environment, where the function (here represented by RNA binding affinity) can become context dependent: in an uncrowded environment, the two variants may bind to RNA differently than in a crowded environment.

Figure 7.7 Weak interactions and Excluded Volume (EV) screening effects on disordered tails. Schematic summary of the competition between weak interactions and excluded volume screening as a function of the degree of polymerization (or molecular weight) of PEG molecules.

7.5 CONCLUSIONS. Our experiments and theoretical modeling provide new insights on the balance between weak-interactions and excluded volume screening effects on disordered regions, in particular in the context of disordered tails. We found weak attractive interactions, previously not observed, that reveal a sequence specificity in the interaction with PEG. While interpretations of ensemble studies are complicated by attractive interactions, single-molecule FRET enables quantification of these impacts on both conformations and binding. These
quantifications are essential elements for rigorously testing current polymer models. In this respect, the theory presented here helps quantify the role of excluded volume screening and weak interactions. However, more work is needed to establish quantitative models capable of capturing the role of entropic and enthalpic interactions of crowders on IDR binding. Future work will focus on studying effects from the perspective of RNA, testing the contribution of cellular crowding, as well as exploring sequence compositional effects within the conformation-affinity-crowding space.

7.6 METHODS. All single-molecule FRET experiments have been performed in a 50 mM HEPES buffer pH 7.4 at a temperature of 23 °C, unless differently specified, with addition of 200 mM β-Mercaptoethanol and 0.001% v/v Tween20. Single-molecule measurements were performed on a modified Picoquant MT200, as previously described 32,52. Volume fractions of PEG solutions have been computed by converting from the weight/volume (w/v) fraction using the density of PEG molecules, as previously reported 14,22. Further experimental and theoretical details are reported in Supplementary Information.

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

Acknowledgements. The research reported in this work was funded by the NIH National Institute on Allergic and Infectious Diseases (R01AI163142).
7.7 Supporting Information

7.7.1 REAGENTS.

**Crowding agents.** Ethylene Glycol (EG), PEG 400, PEG 600, PEG 1000, PEG 2050, PEG 4600, PEG 8000 and PEG 35000 were purchased from Sigma Aldrich (USA). PEG 6000 was purchased from Millipore Sigma (USA). PEG 3350 was purchased from INTEGRA Chemical Company (USA).

**Protein and RNA.** RNA used in this study was synthesized and purified with HPLC by IDT technologies (USA). NTD\textsubscript{L}-RBD and RBD\textsubscript{L} were expressed, purified, and labeled as previously described\textsuperscript{32,33}.

**Cuvette for single-molecule experiments.** Glass cuvettes were assembled using glass cylinders from Hilgenberg (Germany), and coverslips from Deckglaser (Germany), glued together using optical adhesive 61 from Norland (USA), and PEGylated using amino silane from UCT Specialties LLC (USA), sodium bicarbonate from Santa Cruz Biotechnology (USA), and mPEG-SVA from Laysan Bio (USA) as previously described\textsuperscript{32}.

7.7.2 Single-molecule fluorescence setup. Single-molecule confocal fluorescence measurements were performed using a Picoquant MT200 instrument (Picoquant, Germany). Pulsed Interleaved Excitation (PIE) was obtained by synchronizing 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) and pulsed at 20 MHz. Lasers were focused in the sample through a 60x1.2 UPlanSApo Superapochromat water immersion objective (Olympus, Japan) and emitted photons were collected through the same objective. Separation of excitation and emission photons was operated by a dichroic mirror (ZT568rpc, Chroma, USA),
and emitted photons were further filtered by a long pass filter (HQ500LP, Chroma Technology) to suppress scattering light at wavelength of excitation and by the confocal pinhole (100 μm diameter). Finally, the emitted photons were separated into four channels by a polarizing beam splitter (which differentiate between perpendicular and parallel polarization), followed by a dichroic mirror (585DCXR, Chroma) for each polarization that further select 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 70-100 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 300 pM, estimated from dilutions of concentrated samples with known concentration, where concentration was previously determined via absorbance measurements.

7.7.3 Measurement conditions. All measurements, unless differently specified, were performed in 50 mM HEPES, pH 7.4 at 23°C (NaOH), 200 mM mercaptoethanol, 0.001% v/v Tween20 (for surface passivation) and PEG at the reported concentrations. All measurements were performed in PEGylated glass cuvettes. Each sample was measured for at least 10 min at 23 ± 0.5 °C unless otherwise indicated. Most measurements were performed at least in duplicate (independent replicates from a new sample preparation) to confirm reproducibility of the results.
7.7.4 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 15 photons in each bin and contiguous bins were merged if the total number of photons was larger than at least 20 photons. Transfer efficiencies for each burst were calculated according to

\[ E = \frac{n_A}{n_A + n_D} \]  

Eq. S7.1

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:

\[ S = \frac{I_D}{I_D + \gamma_{PIE} I_A} \]  

Eq. S7.2

where \( I_D \) and \( I_A \) represent the total intensities observed after donor and acceptor excitation and \( \gamma_{PIE} \) 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).
7.7.5 **Fit of transfer efficiency distributions.** To estimate the mean transfer efficiency and extract multiple populations from the transfer efficiency distributions, each population was approximated with a Gaussian distribution function. When fitting more than one peak, the histogram is analyzed with a sum of two Gaussian functions. In the case of binding of RNA, the two Gaussian distributions associated with the bound and unbound state are constrained to the values of mean transfer efficiency and width of the sample in absence of ligand and at saturation of ligand.

7.7.6 **Estimate of fraction bound.** To estimate the fraction bound and the corresponding $K_A$, we adopted three independent methods. In the first method, we globally fitted the distribution of transfer efficiencies with two Gaussian distributions, one with a fixed conformation determined from the sample in absence of RNA. In the second method, we fitted the mean transfer efficiency of the distribution as it changes with increasing ligand concentration. In the third method, we normalized the histogram and subtract the histogram of the ligand free population from all the subsequent histograms. We then use the area of the population below a threshold given by the “isosbestic” point across the RNA titration to evaluate the fraction bound at each concentration. All three methods provide analogous results on $K_A$ as shown in Fig. S7.11.

7.7.7 **Nanosecond FCS analysis.** Autocorrelation of acceptor, donor, and cross-correlation curves between acceptor and donor channels were calculated as described previously $^{53,54}$. All measurements were performed at single-molecule concentrations ($\sim 300$ pM), and donor-only
bursts were discarded. Finally, the correlation was computed over a time window of 5 μs, and characteristics timescales were extracted according to:

\[
g_{ij}(\tau) = 1 + \frac{1}{M} \left( 1 - c_{AB} \exp\left[-\frac{(\tau - \tau_0)}{\tau_{AB}}\right] \right) \times \\
\times \left( 1 + c_{CD} \exp\left[-\frac{(\tau - \tau_0)}{\tau_{CD}}\right] \right) \left( 1 + c_T \exp\left[-\frac{(\tau - \tau_0)}{\tau_T}\right] \right)
\]

Eq. S7.3

where \( M \) 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).

### 7.7.8 Determination of root mean square interdye distances from mean 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 \(^{55}\). In the Gaussian model, the conversion rely on one single fitting parameter, the root mean square interdye distance \( r = \sqrt{\langle R^2 \rangle} \).

Estimates of this parameter is obtained by numerically solving:

\[
\langle E \rangle = \int_0^\infty E(R) P(R) \, dR
\]

Eq. S7.4
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 \( R_0 \):

\[
E(R) = \frac{R_0^6}{R_0^6 + R^6}
\]

Eq. S7.5

where \( R_0 \) denotes the Förster distance for the pair of fluorophores employed (5.4 nm for Alexa Fluor 488/Alexa Fluor 594 in water at 23°C 56). The Gaussian chain distribution is given by:

\[
P(R) = 4\pi R^2 \left( \frac{3}{2\pi R^2} \right)^{3/2} \exp \left( -\frac{3R^2}{2R} \right)
\]

Eq. S7.6

Eqs. S7.5 and S7.6 are substituted into Eq. S7.4 and \( r \) is numerically optimized such that it equals the experimentally determined value for mean transfer efficiency.

\( R_0 \) in Eq. S7.5 is proportional to the refractive index \( r^{-2/3} \) 57 and therefore, upon addition of EG or PEG of different sizes, it is corrected for changes in refractive index according to \( R_{0,\text{PEG}} = R_0 (r_{\text{PEG}}/r_0)^{55} \), where \( r_0 \) is the refractive index in absence of crowders.

7.7.9 Protein conformations. To capture the competitive effects of the weak interactions of PEG molecules with the protein as well as the excluded volume screening effects due to crowding, we approximated the problem by separating the two contributions. Although weak interactions are favored by the high concentration of polymers, only a small fraction of polymers interact with the protein. These weak interactions are treated analogously to the ones observed for denaturants and osmolytes. By occupying a large fraction of the available volume, the
remaining PEG molecules cause screening of the excluded volume of the protein and lead to its compaction.

To describe this type of scheme, we used the approach developed in Hofmann et al.\textsuperscript{41}, where the conformations of the protein are described by the coil-to-globule model of Sanchez\textsuperscript{38}, following the work of Sherman and Haran\textsuperscript{39} and subsequent modifications\textsuperscript{40}. In this framework the degree of compaction of the polymer chain can be described in relation to the theta state configuration as:

\[
\alpha^5 - \alpha^3 = \frac{3}{14} N \frac{N b^3}{R_g^3} (1 - \epsilon) + \frac{N b^3}{R_g^3} \frac{N}{7\alpha^3} \tag{Eq. S7.7}
\]

with $\alpha$ being equal to the ratio of the radius of gyration of the protein $R_g$ compared to the radius of gyration of an ideal chain $R_{\theta0}$ and $\epsilon$ is the interaction energy, $N$ is the number of protein monomers, $b$ is the length associated with each monomer.

To be consistent with the definitions used in the renormalization group theory results\textsuperscript{25–27} that we will apply later, we define the radius of gyration of the protein in buffer conditions as:

\[
R_{\theta0} = B_N 0.8^{0.5} N^{0.588} \tag{Eq. S7.8a}
\]

under the assumption that the protein chain is in the good solvent regime. This is the case for our protein chain, which has more expanded configurations than an ideal chain because of repulsive electrostatic interactions. The terms $B_N$ summarizes the specific excluded volume information of the protein chain. We define the ideal reference state as:

374
Based on the definitions in Eqs. S7.7 and S7.8 we can define an interaction energy for the reference state that is given by:

$$\varepsilon_0 = 1 + b^{-3} B_N 3 \left( \frac{3.34}{N^{0.236}} - \frac{3.34}{N^{0.06}} \right) + 0.93 b^{-3} B_N N^{-0.764} \quad \text{Eq. S7.9}$$

The interaction term can be related to the change in the free energy of solvation through $\Delta \varepsilon = \varepsilon - \varepsilon_s = \Delta G_{sol}$.

Minimization of Eq. S7.7 provides the effect of the contribution of weak protein-PEG interactions to the dimension of the chain, which alters the characteristic excluded volume parameter $B_N$ as

$$B_{N,w} = \alpha B_N N^{-0.088} \quad \text{Eq. S7.10}$$

where the subscript $N,w$ is to remember that this quantity is derived in presence of weak interactions.

We can define an analogous quantity for PEG, which is given by

$$B_p = R_{g}^{PEG} 0.8^{-0.5} P^{-0.588} \quad \text{Eq. S7.11}$$

where $P$ is the number of Kuhn segments in PEG. $R_{g}^{PEG}$ is given by

$$R_{g}^{PEG} = 0.21 \left( \frac{M_P}{M_{P,mon}} \right)^{0.583}$$

as previously reported in literature 58.

As discussed previously 14,15,18, excluded volume screening can be described using renormalized estimates of the excluded volume parameters for both protein and PEG as well as their relative
interaction. To simplify the case scenario, we will use the renormalized theory also for short PEG polymers (according to Flory’s criterion): while this is not necessarily accurate, it does not introduce large deviations in the expected trend and restricts the number of variables at play.

With these definitions, we can now define the fundamentals parameters for the renormalized group theory including renormalized values for the number of protein and PEG Kuhn segments (\(N_R\) and \(P_R\), respectively) concentration of PEG (\(c^P_R\)), and characteristic renormalized length of the protein Kuhn segment (\(l^P_R\)):

\[
y = \frac{p}{N} \left( \frac{B_p}{B_{s,w}} \right)^{-0.588} \quad \text{Eq. S7.12a}
\]

\[
s_p = u B^3_p \phi b^{-3} P^{0.76}_0 y^{-0.76} \quad \text{Eq. S7.12b}
\]

\[
N^{-1}_R + 2 s_p N^{-0.76}_R = 1 \quad \text{Eq. S7.12c}
\]

\[
P_R = y N_R \quad \text{Eq. S7.12d}
\]

\[
c^P_R = 1/2 \left( 1 - N^{-1}_R \right) \quad \text{Eq. S7.12e}
\]

\[
l_R = B_{s,w} (N/N_R)^{0.588} \quad \text{Eq. S7.12f}
\]

\[
f_{N,p} = 1 \pm \left( \frac{l_R}{s_{N,p}} \right)^{-0.4} \quad \text{Eq. S7.12g}
\]

The interaction between the monomers of PEG and proteins is provided by the \(s_{N,p}\) parameter and the associated sign represents compatible vs. incompatible solutions \(^{25,26}\). In this case we set
\( s_{N,P} \) equal to zero to restrict to the case of pure excluded volume interactions. The parameter \( u \) is a constant set to be equal to 5.756.

Finally we can compute the root mean square radius of gyration in presence of the excluded volume screening effects of PEG as:

\[
R_N^g = \left( t_R^2 N_R \left( 0.636 + 0.165 N_R^{0.5} - 0.292 N_R^{0.5} f_{N,P} G\left( c_R^P P_R, N_R / P_R \right) \right) \right)^{0.5} \tag{Eq. S7.13}
\]

where:

\[
G(W, y) = W \int_0^\infty z^{-1/2} \frac{B(z) D(z) y}{1 + W D(z) y} dz \tag{Eq. S7.14a}
\]

\[
B(z) = \frac{1}{6} e^{-z} - 2 D(z) - \frac{8}{z} (D(z) - 1) - \frac{10}{z^2} (D(z) - 1 + \frac{z}{3}) \tag{Eq. S7.14b}
\]

\[
D(z) = \frac{2}{z^2} \left( e^{-z} - 1 + z \right) \tag{Eq. S7.14c}
\]

The set of equations Eqs. S7.7-S14 is then deployed to fit the experimental dataset across all molecular weights of PEG, from EG to PEG 35000, and all volume fractions. Results are displayed in Fig. 7.3B.

Assuming a relation analogous to the one proposed by Schellman \(^{41,43,44}\) and assuming the end-groups of PEG and the backbone of PEG interact with independent sites on the protein, we can write:

\[
\Delta G_{ee} = \gamma_{ee} \ln \left[ 1 + 2 K_{ee} \left( \Phi_{w/v} / M_{P,mon} \right) / \left( M_p / M_{P,mon} + 1 \right) \right] \tag{Eq. S7.15a}
\]
\[ \Delta G_{\text{int}} = \gamma_{\text{int}} \ln \left[ 1 + K_{\text{int}} \left( \phi_{w/v}/M_{p,\text{mon}} \right) \left( M_p/M_{p,\text{mon}} - 1 \right)/\left( M_p/M_{p,\text{mon}} + 1 \right) \right] \]

Eq. S7.15b

where here \( \gamma_{ee} \) and \( \gamma_{\text{int}} \) represent the (fractional) number of binding sites associated to each group, \( \phi_{w/v} \) is weight per volume concentration, \( K_{ee} \) and \( K_{\text{int}} \) provide the association constant in mol\(^{-1} \) units, whereas \( M_p \) is the molecular weight of PEG and \( M_{p,\text{mon}} \) is the molecular weight of the monomer of PEG. Here we approximate \( M_{p,\text{mon}} \) to 50 Da. For simplicity in the fit we also assume that \( K_{ee} \) and \( K_{\text{int}} \) are similar, \( K_{ee} \sim K_{\text{int}} \), and the overall difference is encoded in the relative concentration of ends and backbone segments (as rescaled by the length of the polymer) and a change in the number of binding sites. In Eq. S7.15, the volume fraction of the polymer is weighted by fractional occurrence of the ends and backbone elements. For a given polymer there are \( \left( M_p/M_{p,\text{mon}} - 1 \right) \) backbone elements and 2 ends contribution over a total of \( (M_p/M_{p,\text{mon}} + 1) \) interacting element per polymer in the solution.

These two different contributions to the free energy are included to account on whether the interaction arises from the PEG end-groups or from the PEG backbone, as previously proposed by the Record lab\(^{37,59} \), such that

\[ \varepsilon = \varepsilon_0 - \Delta G_{ee} - \Delta G_{\text{int}} \]

Eq. S7.16

The linear additivity is a strong simplification of the more complex solvent exchange model for four components in solutions.

### 7.7.10 A note regarding the renormalized group theory model.

It is important to note some limitations in the treatment proposed here and corresponding nuances in the
renormalized group theory for the ternary solutions. The theory is developed and is strictly valid for a polymer chain in a solution of longer chains. However, we found that deviations occurring for shorter chains are reasonably small and the convenience of applying a single theory is of greater benefit when including and comparing further elements. Importantly, this theory allows for accounting attractive and repulsive interactions with the crowder. However, we found that the impact of the interaction with shorter PEG molecules (EG and PEG 400), which are attractive and expand the chain, could not be reproduced by the theory, if not in part, using reasonable estimates of the interaction parameters. We think this is possibly an intrinsic limitation of the theory that may arise because experimentally there is a difference between favorable/unfavorable mixing interactions and the physical binding of a ligand to the chain. For this reason, in this work, we have opted for using the theory only to establish the pure excluded volume limit and assign to other processes the remaining changes that are observed in the experiments. Finally, we want to mention that the theory is solved in the limit of an excluded volume chain and in a mixture of excluded volume chains. While this applies well for PEG, the disordered region is not exactly in the excluded volume limit. In future works, we aim to adapt the theory to capture intermediate regimes between the excluded volume regime and the theta state conditions.

7.7.11 Flory’s criterion. Application of Flory’s criterion for identifying the regime of long vs short polymers is partially complicated in this context by the presence of the folded RBD. We assumed an identical segment length for each PEG element and Ca-Ca distance. The interdye distance measured within NTD$_L$-RBD is $N=68$ residues, which results in $P=N^{1/2}=8$, equivalent to PEG 400 (assuming a molecular of ~ 50 Da for each monomer of PEG). If we consider the whole NTD$_L$-RBD region, this encompasses $N = 173$ residues, which results in $P=N^{1/2}=13$,
equivalent to ~ PEG 650. Therefore, we considered PEG 400 and PEG 600 as polymers in the crossover regime, and PEG 1000 (and higher molecular volumes) as “larger” than the protein.

### 7.7.12 Protein and nucleic acid binding.

Binding of RNA ligands to labeled N protein constructs in the presence and absence of PEG, was monitored by following the fraction of bursts associated with the bound and unbound population.

When the fraction of bound protein $f_b$ is directly estimated from the distribution of transfer efficiencies, titration curves were analyzed according to:

$$f_b = \frac{K_A [RNA]_{tot}}{1 + K_A [RNA]_{tot}}$$

Eq. S7.17

where $K_A$ is the association constant and $[RNA]_{tot}$ is the total concentration of RNA. Since the protein and (rU)$_{40}$ form a complex with a 1:1 stoichiometry and we measure directly the fractions of free and bound populations not just a signal variation, we can determine $K_A$ with good accuracy from a few points of the titration.

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 of fluorescently labeled protein used in the experiments (~200pM).

### 7.7.13 Depletion theory.

The impact of crowding on the binding of IDRs to ligands has been previously demonstrated by Zosel et al. In brief, the variation in the binding free energies in presence and absence of
crowders, $\Delta G$ and $\Delta G_0$ respectively, can be computed by comparing the the equilibrium constants $K_A$ and $K_{A,0}$ according to:

$$\Delta \Delta G = \Delta G - \Delta G_0 = -k_B T \ln \frac{K_A}{K_{A,0}} \quad \text{Eq. S7.18}$$

where $k_B$ is the Boltzmann constant and $T$ the temperature in Kelvin.

The osmotic pressure of the solution is given by $\Pi = n k_B T$, with $n$ being the number density of colloids. The number density $n$ is linked to the concentration $c$ of colloids the concentration by $n = c/M$, with $M$ being the molar mass of the crowding agents. The overlap volume $V_{\text{overlap}}(r)$ of the colloids depletion layers gives rise to a distance-dependent attractive interaction potential between the colloids, $W(r)$. For two spherical colloidal particles, the binding $\Delta \Delta G$ is equivalent to the interaction potential when the two spheres are brought into contact $^{60}$. This can be computed setting $r$ equal to zero in the interaction potential, i.e. $W(0)$, which leads to:

$$\Delta \Delta G = W(0) = -nk_B TV_{\text{overlap}}(0) = -\frac{c}{M} k_B TV_{\text{overlap}}(0) \quad \text{Eq. S7.19}$$

where the overlap volume is defined by:

$$V_{\text{overlap}}(0) = \frac{\pi (r+R-d)^2(d^2-3(r-R)^2+2d(r+R))}{12d} \quad \text{Eq. S7.20}$$

with $^{61,62}$
\[ r = R_1 + \delta_s (R_1) \quad \text{Eq. S7.21a} \]
\[ R = R_2 + \delta_s (R_2) \quad \text{Eq. S7.21b} \]
\[ d = R_1 + R_2 \quad \text{Eq. S7.21c} \]
\[ \frac{\delta}{R} = \left( 1 + 3 \frac{\delta_0}{R} + 2.273 \left( \frac{\delta_0}{R} \right)^2 - 0.0975 \left( \frac{\delta_0}{R} \right)^3 \right)^{\frac{3}{2}} - 1 \quad \text{Eq. S7.21d} \]
\[ \delta_0 = 1.07 R_g \quad \text{Eq. S7.21e} \]

The depletion layer \( \delta_0 \) accounts for the “soft” nature of the polymer compared to a rigid hard sphere of equivalent radius of gyration \( R_g \).

To account for the crossover of length scales when passing from a dilute to a semidilute solution, an effective depletion layer \( \delta \) is defined as:
\[ \delta^{-2} = \delta_0^{-2} + \xi^{-2} \quad \text{Eq. S7.22} \]

where \( \delta_0 \) is the depletion layer in a dilute solutions and \( \xi \) is the average mesh size of the polymer solution in the semidilute regime."
7.8 Supporting Figures.
Supplementary Figure S7.1. Mean transfer efficiency at 15% w/v as a function of PEG Molecular Weight. Line is a guide for the eyes.
**Supplementary Figure S7.2. NTD\textsubscript{1}-RBD dynamics under crowding conditions.** Example of nanosecond-FCS (nsFCS) traces of NTD\textsubscript{1}-RBD in the presence of 15% w/v PEG 400 (A) and PEG 3350 (B) solutions. Donor-donor, acceptor-acceptor, and donor-acceptor correlations are shown in green, red, and orange (respectively). Solid lines report about a global fit to Eq. S7.3 and corresponding residuals.
Supplementary Figure S7.3. Dilute and semidilute regimes. Schematic of the overlap concentration $\phi^*$ as a function of the degree of polymerization (number of monomers) of the PEG molecule.
Supplementary Figure S7.4. Temperature dependence of RBD$_L$. *Upper panel.* Representative transfer efficiency distributions of RBD$_L$ in 50 mM HEPES buffer. *Lower panel.* Corresponding contour plot as a function of temperature and transfer efficiency. Solid lines represent measured temperatures. A transition occurs between 37 and 43 °C with a clear shift of transfer efficiencies to lower values. The trend is opposite to what is observed for NTD$_L$-RBD and suggests a conformational change in the RBD. The specific shift in transfer efficiency is reminiscent of the intermediate state populated at low denaturant concentrations $^{32,33}$. 
Supplementary Figure S7.5. NTD-RBD conformations and binding of (rU)$_{40}$ as a function of temperature. A-C. Upper panels: Representative transfer efficiency distributions of the NTD$_{L}$-RBD in 50 mM HEPES buffer (A), with addition of 20 nM (rU)$_{40}$ (B), and with addition of 500 nM (rU)$_{40}$ (C) as a function of temperature from 10 °C to 56 °C. Lower panels. Contour plots of the transfer efficiency distributions of NTD$_{L}$-RBD as a function of temperature for the corresponding experimental conditions in the upper panels. D. Plot of the log($K_A$/nM) as a function of $1/T$ (K$^{-1}$). Error bars represent error of the fit when calculating $K_A$ as a function in changes in transfer efficiency. The linear fit reports about the enthalpy (slope) and entropy (intercept) of binding according to Eq. 7.4. Two different fits are provided for the regime where one single population is observed (up to 37 °C) and a second linear fit is performed for higher temperatures. Fitting results are reported in Supplementary Table 7.7 and 7.8.
Supplementary Figure S7.6. NTD-RBD conformations and binding of \((rU)_{40}\) as a function of temperature in the presence of 15% w/v EG. A-C. Upper panels. Representative transfer efficiency distributions of the NTD-RBD in 50 mM HEPES, 15% w/v EG buffer (A), with addition of 30 nM \((rU)_{40}\) (B), and with addition of 500 nM \((rU)_{40}\) (C) as a function of temperature from 10 °C to 37°C. Lower panels. Contour plots of the transfer efficiency distributions of NTD-RBD as a function of temperature for the corresponding experimental conditions in the upper panels. D. Plot of the \(\ln(K_a/\text{mM})\) as a function of \(1/T\) (K\(^{-1}\)). Error bars represent errors of the fit. The linear fit reports about the enthalpy (slope) and entropy (intercept) of binding according to Eq. 7.4. Fitting results are reported in Supplementary Table 7.8 and 7.9.
Supplementary Figure S7.7. NTD-RBD conformations and binding of (rU)_{40} as a function of temperature in the presence of 15% w/v PEG 600. A-C. Upper panels. Representative transfer efficiency distributions of the NTD_{L}-RBD in 50 mM HEPES buffer (A), with addition of 2.5 nM (rU)_{40} (B), and with addition of 12.5 nM (rU)_{40} (C) as a function of temperature from 10 °C to 37 °C. Lower panels. Contour plots of the transfer efficiency distributions of NTD_{L}-RBD as a function of temperature for the corresponding experimental conditions in the upper panels. D. Plot of the log(K_{A}/nM) as a function of 1/T (K^{-1}). Error bars represent errors of the fit. The linear fit reports about the enthalpy (slope) and entropy (intercept) of binding according to Eq. 7.4. Fitting results are reported in Supplementary Table 7.8 and 7.9.
Supplementary Figure S7.8. NTD-RBD conformations and binding of (rU)₄₀ as a function of temperature in the presence of 15% w/v PEG 8000. A-C. Upper panels. Representative transfer efficiency distributions of the NTD₁-RBD in 50 mM HEPES buffer (A), with addition of 20 nM (rU)₄₀ (B), and with addition of 500 nM (rU)₄₀ (C) as a function of temperature from 10 °C to 37 °C. Lower panels. Contour plots of the transfer efficiency distributions of NTD₁-RBD as a function of temperature for the corresponding experimental conditions in the upper panels. D. Plot of the log(Kᵦ/nM) as a function of 1/T (K⁻¹). Error bars represent errors of the fit. The linear fit reports about the enthalpy (slope) and entropy (intercept) of binding according to Eq. 7.4. Fitting results are reported in Supplementary Tables 7.8 and 7.9.
Supplementary Figure S7.9. NTD1-RBD conformations as a function of temperature in absence and presence of 15\% w/v PEG solutions. A. Representative histograms of conformational changes in absence of crowders, 15\% w/v EG, PEG 600, and PEG 8000. B. Corresponding mean transfer efficiencies as function of the temperature. The gray area represents the region where a second conformation is absence of crowding. Addition of large crowders shifts the transition toward lower temperatures as indicated by the sudden increase in transfer efficiency at 37 C.
Supplementary Figure S7.10. ΔΔH and ΔΔS at 15% w/v for EG (red), PEG 600 (yellow), PEG 8000 (gray). Cyan area indicates the region where binding is favored, the white one, the area where binding is unfavored. The measured values for EG, PEG 600, and PEG 8000 lie on a line (black line) parallel to the bisectrix (red line). Error bars are propagated from the errors on ΔH and ΔS.
Supplementary Figure S7.11. Temperature dependence of $\ln(K_A)$ as estimated from different fitting models. Here we compared the temperature dependence of $\ln(K_A)$ based on different methods of estimation: i) fitting the change in the area of normalized histograms based on the; ii) fitting the change in the mean transfer efficiency; iii) analyzing the change in the area fractions obtained by fitting two Gaussian distributions on the transfer efficiency distribution. A. Comparison of $\ln(K_A)$ across different PEG conditions (as indicated in the legend) for each of the methods. B. Comparison of $\ln(K_A)$ for each individual set across the three different determination methods.
Supplementary Figure S7.12. Binding curves from single-molecule FRET experiments for wild type Wuhan-Hu-1 and Omicron variants in absence (empty and cyan dots) and presence (orange and blue dots) of 15% w/v PEG 600.
7.9 Supporting Tables

Supplementary Table 7.1. Sequence of WT NTD-RBD. Labeling positions are reported in purple.

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
<th>Start Position (WT)</th>
<th>End Position (WT)</th>
<th>Labeling Positions (WT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTD&lt;sub&gt;1&lt;/sub&gt;-RBD</td>
<td>GP&lt;sub&gt;CSDNGPO&lt;/sub&gt;QQR&lt;sub&gt;NAPRITFGGSP&lt;/sub&gt;ST&lt;sub&gt;DS&lt;/sub&gt;T&lt;sub&gt;GS&lt;/sub&gt;NQN&lt;sub&gt;Q&lt;/sub&gt;ER&lt;sub&gt;SG&lt;/sub&gt;ARS&lt;sub&gt;KQ&lt;/sub&gt;R&lt;sub&gt;RP&lt;/sub&gt;QGL&lt;sub&gt;LPNNT&lt;/sub&gt;A&lt;sub&gt;SA&lt;/sub&gt;W&lt;sub&gt;F&lt;/sub&gt;TALTQH&lt;sub&gt;H&lt;/sub&gt;GC&lt;sub&gt;EDLKFP&lt;/sub&gt;C&lt;sub&gt;G&lt;/sub&gt;G&lt;sub&gt;Q&lt;/sub&gt;GP&lt;sub&gt;INTNSSP&lt;/sub&gt;D&lt;sub&gt;DQ&lt;/sub&gt;IG&lt;sub&gt;GY&lt;/sub&gt;YRA&lt;sub&gt;RR&lt;/sub&gt;T&lt;sub&gt;RR&lt;/sub&gt;IR&lt;sub&gt;RR&lt;/sub&gt;GD&lt;sub&gt;G&lt;/sub&gt;D&lt;sub&gt;KK&lt;/sub&gt;MD&lt;sub&gt;L&lt;/sub&gt;LSP&lt;sub&gt;RY&lt;/sub&gt;W&lt;sub&gt;F&lt;/sub&gt;Y&lt;sub&gt;Y&lt;/sub&gt;LGT&lt;sub&gt;G&lt;/sub&gt;PEAG&lt;sub&gt;LPY&lt;/sub&gt;G&lt;sub&gt;ANK&lt;/sub&gt;KD&lt;sub&gt;G&lt;/sub&gt;I&lt;sub&gt;H&lt;/sub&gt;W&lt;sub&gt;V&lt;/sub&gt;TE&lt;sub&gt;G&lt;/sub&gt;ALN&lt;sub&gt;TP&lt;/sub&gt;D&lt;sub&gt;H&lt;/sub&gt;IG&lt;sub&gt;TR&lt;/sub&gt;N&lt;sub&gt;P&lt;/sub&gt;ANN&lt;sub&gt;AA&lt;/sub&gt;AIL&lt;sub&gt;L&lt;/sub&gt;Q&lt;sub&gt;LP&lt;/sub&gt;Q&lt;sub&gt;GTL&lt;/sub&gt;LPK&lt;sub&gt;G&lt;/sub&gt;FY&lt;sub&gt;A&lt;/sub&gt;</td>
<td>1</td>
<td>173</td>
<td>1, 68</td>
</tr>
<tr>
<td>NTD&lt;sub&gt;1&lt;/sub&gt;-RBD Omicron (P13L, Δ31-33)</td>
<td>GP&lt;sub&gt;CSDNGPO&lt;/sub&gt;QR&lt;sub&gt;NAL&lt;/sub&gt;RR&lt;sub&gt;ITFGG&lt;/sub&gt;PS&lt;sub&gt;DT&lt;/sub&gt;T&lt;sub&gt;GS&lt;/sub&gt;NQN&lt;sub&gt;Q&lt;/sub&gt;ER&lt;sub&gt;G&lt;/sub&gt;ARS&lt;sub&gt;KQ&lt;/sub&gt;R&lt;sub&gt;RP&lt;/sub&gt;QGL&lt;sub&gt;LPNNT&lt;/sub&gt;A&lt;sub&gt;SA&lt;/sub&gt;W&lt;sub&gt;F&lt;/sub&gt;TALTQH&lt;sub&gt;H&lt;/sub&gt;G&lt;sub&gt;K&lt;/sub&gt;E&lt;sub&gt;DLKFP&lt;/sub&gt;C&lt;sub&gt;G&lt;/sub&gt;G&lt;sub&gt;Q&lt;/sub&gt;GP&lt;sub&gt;O&lt;/sub&gt;N&lt;sub&gt;SSPDQ&lt;/sub&gt;IG&lt;sub&gt;G&lt;/sub&gt;YRR&lt;sub&gt;ATR&lt;/sub&gt;R&lt;sub&gt;IR&lt;/sub&gt;GG&lt;sub&gt;D&lt;/sub&gt;G&lt;sub&gt;K&lt;/sub&gt;MD&lt;sub&gt;L&lt;/sub&gt;LSP&lt;sub&gt;RY&lt;/sub&gt;W&lt;sub&gt;F&lt;/sub&gt;Y&lt;sub&gt;Y&lt;/sub&gt;LGT&lt;sub&gt;G&lt;/sub&gt;PEAG&lt;sub&gt;LPY&lt;/sub&gt;G&lt;sub&gt;ANK&lt;/sub&gt;D&lt;sub&gt;G&lt;/sub&gt;I&lt;sub&gt;H&lt;/sub&gt;W&lt;sub&gt;V&lt;/sub&gt;TE&lt;sub&gt;G&lt;/sub&gt;ALN&lt;sub&gt;TP&lt;/sub&gt;D&lt;sub&gt;H&lt;/sub&gt;IG&lt;sub&gt;TR&lt;/sub&gt;N&lt;sub&gt;P&lt;/sub&gt;ANN&lt;sub&gt;AA&lt;/sub&gt;AIL&lt;sub&gt;L&lt;/sub&gt;Q&lt;sub&gt;LP&lt;/sub&gt;Q&lt;sub&gt;GTL&lt;/sub&gt;LPK&lt;sub&gt;G&lt;/sub&gt;FY&lt;sub&gt;A&lt;/sub&gt;</td>
<td>1</td>
<td>173</td>
<td>1, 68</td>
</tr>
<tr>
<td>RBD&lt;sub&gt;L&lt;/sub&gt;</td>
<td>HHHHHHHHHLFQGPSWFTALTQHGKE DLKFCGQVPINTNSSPDDQIGYRRATRR IRGGDGKMDLSPRWYFYYLGTGPEAGLPY GANKDGIHWVATEGALNTPKH&lt;sub&gt;T&lt;/sub&gt;IGTRNPANN AAIVQLPQGTTLPKGFA</td>
<td>51</td>
<td>173</td>
<td>68,172</td>
</tr>
</tbody>
</table>
Supplementary Table 7.3. PEGs used in this study

<table>
<thead>
<tr>
<th>Company</th>
<th>Catalog Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>EG</td>
<td>Sigma-Aldrich</td>
</tr>
<tr>
<td>PEG 400</td>
<td>Sigma-Aldrich</td>
</tr>
<tr>
<td>PEG 600</td>
<td>Sigma-Aldrich</td>
</tr>
<tr>
<td>PEG 1000</td>
<td>Sigma-Aldrich</td>
</tr>
<tr>
<td>PEG 2050</td>
<td>Sigma-Aldrich</td>
</tr>
<tr>
<td>PEG 3350</td>
<td>INTEGRA Chemical Company</td>
</tr>
<tr>
<td>PEG 4600</td>
<td>Sigma-Aldrich</td>
</tr>
<tr>
<td>PEG 6000</td>
<td>EMD Millipore Corp.</td>
</tr>
<tr>
<td>PEG 8000</td>
<td>Sigma-Aldrich</td>
</tr>
<tr>
<td>PEG 35000</td>
<td>Sigma-Aldrich</td>
</tr>
</tbody>
</table>

Supplementary Table 7.4. Dynamics of the NTD<sub>L</sub>-RBD in the presence of 15% Crowder.

<table>
<thead>
<tr>
<th>PEG</th>
<th>(\tau_{CD}) (ns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PEG 400</td>
<td>123 (\pm) 4</td>
</tr>
<tr>
<td>PEG 3350</td>
<td>119 (\pm) 3</td>
</tr>
</tbody>
</table>
**Supplementary Table 7.5. Schellman weak binding model individual fit results.** Values of \( K \) are not well defined because of the small amplitude or variation in data points, whereas value of \( \gamma \) can be determined more robustly from the amplitude change in \( \Delta \varepsilon \). Corresponding values of \( \gamma \) are reported in Fig. 3D.

<table>
<thead>
<tr>
<th>individual fits</th>
<th>( K ) (M(^{-1}))</th>
<th>( \gamma )</th>
</tr>
</thead>
<tbody>
<tr>
<td>EG</td>
<td>1.2 ± 0.4</td>
<td>0.28 ± 0.05</td>
</tr>
<tr>
<td>PEG 400</td>
<td>1.9 ± 0.9</td>
<td>0.21 ± 0.05</td>
</tr>
<tr>
<td>PEG 600</td>
<td>0.4 ± 0.4</td>
<td>0.3 ± 0.3</td>
</tr>
<tr>
<td>PEG 1000</td>
<td>1.6 ± 0.8</td>
<td>0.13 ± 0.03</td>
</tr>
<tr>
<td>PEG 2050</td>
<td>50 ± 300</td>
<td>0.02 ± 0.02</td>
</tr>
<tr>
<td>PEG 3350</td>
<td>100 ± 300</td>
<td>0.02 ± 0.01</td>
</tr>
<tr>
<td>PEG 4600</td>
<td>100 ± 2000</td>
<td>0.008 ± 0.03</td>
</tr>
<tr>
<td>PEG 6000</td>
<td>0.6 ± 0.6</td>
<td>0.08 ± 0.05</td>
</tr>
<tr>
<td>PEG 8000</td>
<td>100 ± 800</td>
<td>0.01 ± 0.02</td>
</tr>
<tr>
<td>PEG 35000</td>
<td>100 ± 5000</td>
<td>0.005 ± 0.05</td>
</tr>
</tbody>
</table>
**Supplementary Table 7.6. Schellman weak binding model global fit results.** Values of \( K, \gamma_{ee}, \) and \( \gamma_i \) are obtained based on Eq. S12 and are reported in Fig. 3A. Note that Eq. S12 introduces a different definition of the concentrations and therefore the values in Supplementary Tables 4 and 5 are not expected to be identical, though trends can be compared.

<table>
<thead>
<tr>
<th>global fit</th>
<th>( K (\text{M}^{-1}) )</th>
<th>( \gamma_{ee} )</th>
<th>( \gamma_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.3 ± 0.8</td>
<td>0.16 ± 0.03</td>
<td>0.037 ± 0.004</td>
</tr>
</tbody>
</table>

**Supplementary Table 7.7.** Association constants for the binding of NTD\(_L\)-RBD to (rU)\(_{40}\) in presence and absence of crowding agents at 23°C in 50 mM HEPES buffer.

<table>
<thead>
<tr>
<th>15% w/v</th>
<th>( K_A ) for (rU)(_{40}) (nM(^{-1}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>no crowders</td>
<td>0.22 ± 0.03</td>
</tr>
<tr>
<td>EG</td>
<td>0.08 ± 0.02</td>
</tr>
<tr>
<td>PEG 400</td>
<td>0.45 ± 0.07</td>
</tr>
<tr>
<td>PEG 600</td>
<td>0.32 ± 0.04</td>
</tr>
<tr>
<td>PEG 1000</td>
<td>0.21 ± 0.03</td>
</tr>
<tr>
<td>PEG 2050</td>
<td>0.16 ± 0.02</td>
</tr>
<tr>
<td>PEG 3350</td>
<td>0.43 ± 0.06</td>
</tr>
<tr>
<td>PEG 8000</td>
<td>0.168 ± 0.001</td>
</tr>
<tr>
<td>PEG 35000</td>
<td>0.098 ± 0.006</td>
</tr>
</tbody>
</table>
**Supplementary Table 7.8.** Association constants for the binding of NTD₄-RBD to (rU)₄₀ in presence and absence of 15% w/v crowding agents at temperatures from 10 °C to 37 °C in 50 mM HEPES buffer. Errors represent the error of the fit when analyzing the areas from the fraction bound.

<table>
<thead>
<tr>
<th>(rU)₄₀</th>
<th>10 °C</th>
<th>16 °C</th>
<th>23 °C</th>
<th>30 °C</th>
<th>37 °C</th>
</tr>
</thead>
<tbody>
<tr>
<td>15% w/v</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no crowders</td>
<td>0.24 ± 0.01</td>
<td>0.24 ± 0.01</td>
<td>0.21 ± 0.07</td>
<td>0.11 ± 0.01</td>
<td>0.09 ± 0.01</td>
</tr>
<tr>
<td>EG</td>
<td>0.27 ± 0.01</td>
<td>0.15 ± 0.01</td>
<td>0.08 ± 0.01</td>
<td>0.02 ± 0.01</td>
<td>0.01 ± 0.01</td>
</tr>
<tr>
<td>PEG 600</td>
<td>0.54 ± 0.01</td>
<td>0.36 ± 0.01</td>
<td>0.43 ± 0.05</td>
<td>0.19 ± 0.01</td>
<td>0.07 ± 0.01</td>
</tr>
<tr>
<td>PEG 8000</td>
<td>0.20 ± 0.01</td>
<td>0.13 ± 0.01</td>
<td>0.12 ± 0.03</td>
<td>0.05 ± 0.01</td>
<td>0.028 ± 0.004</td>
</tr>
</tbody>
</table>
Supplementary Table 7.9. Enthalpic and entropic contribution to the binding free energy in presence and absence of 15% w/v crowding agents at temperatures from 10 °C to 37 °C in 50 mM HEPES buffer. Errors associated with ΔΔH_exp and ΔΔS_exp are obtained through propagations of the errors.

<table>
<thead>
<tr>
<th></th>
<th>ΔH</th>
<th>ΔS</th>
<th>ΔΔH</th>
<th>ΔΔS</th>
<th>-TΔΔS at 23°C</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>noPEG</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean E</td>
<td>-13 ± 1</td>
<td>-0.004 ± 0.004</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normalized Histo</td>
<td>-14 ± 2</td>
<td>-0.008 ± 0.006</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Area Fracs</td>
<td>-7 ± 2</td>
<td>0.014 ± 0.005</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>EG</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*No 37C Mean E</td>
<td>-29 ± 6</td>
<td>-0.062 ± 0.019</td>
<td>-16 ± 7</td>
<td>-0.06 ± 0.02</td>
<td>17 ± 7</td>
</tr>
<tr>
<td>Normalized Histo</td>
<td>-24 ± 4</td>
<td>-0.047 ± 0.012</td>
<td>-10 ± 6</td>
<td>-0.04 ± 0.02</td>
<td>11 ± 5</td>
</tr>
<tr>
<td>Area Fracs</td>
<td>-21 ± 2</td>
<td>-0.035 ± 0.007</td>
<td>-14 ± 3</td>
<td>-0.049 ± 0.012</td>
<td>15 ± 4</td>
</tr>
<tr>
<td><strong>PEG 600</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean E</td>
<td>-13 ± 3</td>
<td>-0.005 ± 0.011</td>
<td>0 ± 5</td>
<td>-0.00 ± 0.02</td>
<td>0 ± 4</td>
</tr>
<tr>
<td>Normalized Histo</td>
<td>-13 ± 4</td>
<td>-0.005 ± 0.013</td>
<td>1 ± 6</td>
<td>0.00 ± 0.02</td>
<td>-1 ± 6</td>
</tr>
<tr>
<td>Area Fracs</td>
<td>-12 ± 3</td>
<td>-0.002 ± 0.011</td>
<td>-5 ± 5</td>
<td>-0.016 ± 0.016</td>
<td>4 ± 5</td>
</tr>
<tr>
<td><strong>PEG 8000</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean E</td>
<td>-11 ± 4</td>
<td>-0.002 ± 0.014</td>
<td>2 ± 5</td>
<td>0.00 ± 0.02</td>
<td>-1 ± 5</td>
</tr>
<tr>
<td>Normalized Histo</td>
<td>-13 ± 5</td>
<td>-0.009 ± 0.016</td>
<td>1 ± 7</td>
<td>-0.00 ± 0.02</td>
<td>0.0 ± 7</td>
</tr>
<tr>
<td>Area Fracs</td>
<td>-12 ± 2</td>
<td>-0.006 ± 0.008</td>
<td>-6 ± 4</td>
<td>-0.02 ± 0.013</td>
<td>6 ± 4</td>
</tr>
</tbody>
</table>
Supplementary Table 7.10. Theoretical expectations from depletion theory ($\Delta \Delta S_{dt}$). $\Delta \Delta S_{dt}$ is computed assuming an effective radius of gyration for the RNA of $0.8 \pm 0.3$ nm. Errors associated with $\Delta \Delta S_{dt}$ are computed from the variation associated with the effective radius of gyration of RNA.

<table>
<thead>
<tr>
<th>15% w/v</th>
<th>$\Delta \Delta S_{exp}$ (kcal/mol/K)</th>
<th>$\Delta \Delta S_{dt}$ (kcal/mol/K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EG</td>
<td>-0.02 ± 0.02</td>
<td>0.0014 (± 0.0003)</td>
</tr>
<tr>
<td>PEG 600</td>
<td>0.01 ± 0.01</td>
<td>0.0021 (± 0.0006)</td>
</tr>
<tr>
<td>PEG 8000</td>
<td>0.01 ± 0.01</td>
<td>0.0029 (± 0.001, -0.0014)</td>
</tr>
</tbody>
</table>

Supplementary Table 7.11. Association constants for the binding of NTD$_L$-RBD to (rU)$_{40}$ calculated using the change in mean transfer efficiencies in absence of crowding agents from 10 °C to 56 °C in 50 mM HEPES buffer.

<table>
<thead>
<tr>
<th>(rU)$_{40}$</th>
<th>$K_A$ (nM$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10 °C</td>
</tr>
<tr>
<td>50 mM HEPES</td>
<td>0.7 ± 0.2</td>
</tr>
</tbody>
</table>
7.10 REFERENCES.


26. Schäfer, L. & Kappeler, C. Interaction effects on the size of a polymer chain in ternary


53. Nettels, D., Hoffmann, A. & Schuler, B. Unfolded protein and peptide dynamics investigated with single-molecule FRET and correlation spectroscopy from picoseconds to


Chapter 8

Conclusions and Future Directions
8.1 Main Findings

My thesis work has focused on dissecting the contribution of disordered regions within the SARS-CoV-2 N protein in modulating the conformational properties of the protein and its interaction with nucleic acids. To this end, my thesis work has centered on the application of single-molecule confocal fluorescence spectroscopy and, in particular, single-molecule FRET to quantify structural and dynamical changes in the N protein, in absence and presence of RNA.

In Chapter one, I have introduced material necessary for understanding the importance of the N protein in the context of COVID-19 and summarizing what was known at the time I started investigating the N protein. Previous investigations of SARS-CoV-1 provided important insights. Back in 2003, when the first SARS outbreak occurred, many studies pointed to the fact that the SARS-CoV-1 N protein was not fully structured, with three regions (NTD, LINK, CTD) adopting disordered configuration. At the time, very little was known about disordered regions and the majority of the studies disregarded their role, focusing their efforts solely on the structured domains\textsuperscript{1-5}.

Twenty years later, when the COVID-19 pandemic began, like for the N protein of SARS-CoV-1, studies rushed to solve structures of the folded domains\textsuperscript{6-10}. However, the main focus did not stay isolated to the folded domains. Other work had a strong emphasis on the ability of the N protein to undergo phase separation \textit{in vitro}\textsuperscript{11-15} and \textit{in cell}\textsuperscript{13,14,16-18}. From the start of the pandemic (early 2020) to present day, countless studies probing N protein’s role in phase separation have been published. Overall, the main picture that arises is that the N protein has a strong affinity for nucleic acids, and under the right conditions, will phase separate into liquid-like droplets \textit{in vitro} and \textit{in cell}. While the phase separation propensity of the protein is an
interesting and important topic of study, my contributions to the study of the N protein aimed to understand the molecular details that drive these interactions.

In chapter two, I provided an introduction to intrinsically disordered proteins, and the use of single-molecule spectroscopy to study their conformations and functions. This review exemplifies the value of single-molecule methods such as fluorescence correlation spectroscopy (FCS) and Förster resonance energy transfer FRET, using both confocal and TIRF microscopy. We highlighted that studying intrinsically disordered proteins or regions using single-molecule spectroscopy provides important biophysical parameters such as conformations and dynamics of the system.

The work I discussed in chapter three describes the first biochemical and biophysical characterization of the SARS-CoV-2 N protein that focused on the intrinsically disordered regions. Using single-molecule FRET, I showed that the three predicted disordered regions were in fact disordered and dynamic in monomeric N protein. Furthermore, the work done in chapter three demonstrated the phase separation capabilities of the N protein, and provided a model where single-genome condensates form through N protein-RNA interactions, which are driven by a small number of high-affinity binding sites.

**Chapter four** has a strong focus on the dimerization ability of the SARS-CoV-2 N protein. My work provides evidence of a partially unstructured dimerization domain in monomeric N protein, undergoing significant conformational changes upon dimerization. I also demonstrate the effects of dimerization on other domains of the protein, showing that the NTD remains unchanged while the LINK and CTD have a partial expansion upon dimerizing. Finally, I also quantified the effects of thermal denaturation of the dimerization domain, providing insights into the
conformations of the protein under more physiological conditions. Overall, this chapter provided new insights into the conformations and energetics of the dimerization domain, which are likely applicable to dimerization domains in other coronavirus N proteins.

Given that the role of the N protein is to bind to and package RNA, chapter five of my thesis is focused on studying the role of the disordered NTD in RNA binding. By studying the RBD, and NTD-RBD in isolation, I found that the presence of the NTD enhances the affinity to RNA about 40-80 fold in comparison to the RBD in isolation. Using nanosecond fluorescence correlation spectroscopy (nsFCS), I found that, even under saturating conditions, the NTD remains dynamic when bound to RNA, indicating that it forms a fuzzy complex\(^9\) with the RNA. This is further confirmed by coarse grained simulations. To study the effects of sequence specificity, I also characterized a small single-stranded region of the 5’UTR of the SARS-CoV-2 genome and found that the NTD adopts different conformations and stoichiometries when binding to specific viral RNA compared to homo-oligonucleotides (rU\(_\text{x}\)). Finally, sequence changes in the intrinsically disordered region found in the Omicron variant (a residue substitution and a three-residue deletion) have negligible effects on conformations, but significant ones in terms of affinity. Specifically, I note that the Omicron variant has a weaker affinity compared to WT. Overall, this chapter helped to define the importance of the NTD in binding RNA. We hypothesized that the dynamic complex formed between the NTD-RBD and RNA may help searching along the genome for specific high affinity motifs.

The sixth chapter discussed the role of macromolecular crowding and intrinsically disordered proteins. This review summarizes the current interpretation of crowding effects on intrinsically disordered proteins and regions, in comparison to their folded counterparts. This work posed the basis for investigating crowding effects on the Nucleocapsid protein.
The seventh chapter focused on investigating how excluded volume effects and weak interactions with crowders can modulate the conformational ensemble of the NTD-RBD and its ability to bind nucleic acids. We find that short PEG molecules are dominated by interactions with the disordered region, whereas long polymer molecules largely contribute in terms of excluded volume effects. The comparison with theoretical expectations from polymer and colloidal particle theories provided new insights on the role of crowding on disordered proteins. For what concerns the role in the SARS-CoV-2, we found that crowding normalizes the effect of the Omicron mutation compared to WT.

In summary, the monomeric full-length protein is flexible and dynamic, undergoes phase separation, and can dimerize at low nanomolar concentrations ($K_D = \sim 11$ nM). The inherent flexibility of the protein is maintained also in the dimer form and when bound to RNA.

In the following, I have discussed the current literature on the N protein in light of the observations from my work.

8.2 SARS-CoV-2 Nucleocapsid interactions with nucleic acids

The ability of N protein to interact and package the viral genome is a critical step in the SARS-CoV-2 life cycle. The knowledge of N protein's essential role led to many studies, over the last three years, that focused on investigating the mechanism by which N protein interacts with RNA.

Electrophoretic mobility shift assays (EMSAs) show that full-length N protein binds to both 7mer ssRNA and dsRNA \(^{20}\). Results suggest N protein has a slight preference for ssRNA\(^{20}\). Wu
et al. further compared binding of full-length N protein to various truncations using fluorescent polarization assays and a FITC-ss-20mer and FITC-stemloop, they demonstrate the capability of multiple domains of the N protein to bind to both single-stranded and structured RNA. Furthermore, they conclude that, similar to our findings, the RBD in isolation binds significantly weaker to dsRNA in comparison to ssRNA. In vitro NMR measurements by Dinesh et al. identified residues in the RNA binding domain, specifically an arginine-rich arm of the domain, that interacts with a 7mer and 10mer oligo from the 5’UTR of the SARS-CoV-2 genome, which contains the CoV-1 transcriptional regulatory sequence. When comparing the regions of interaction from NMR with our work, we observed similar regions of interaction from coarse-grained simulations (See chapter 5, Figure 5.5). Interestingly, in the presence of the full disordered NTD, we observed a modulation of the interactions in both NTD and RBD, suggesting that presence of the tail alters the exact mode of binding of RNA.

In cell work by Iserman et al. identified structural elements of the SARS-CoV-2 genome that interact with the N protein via cross-linking studies. The 5’end and frameshift regions of the genome were predicted using SHAPE-MaP. RNP-MaP was utilized to crosslink the lysine residues of N protein to proximal nucleotides. Specific regions of the 5’end were identified as sites of interaction and promoted phase separation, whereas the frameshifted regions significantly decreased the phase separation and promoted solubility. From the 5’end, the V21 sequence measured in chapter five was observed to interact with the N protein. From these regions, interactions were shown to be different depending on the state of the N protein (in a diffuse state or in a droplet state from their phase separation assays). This may resemble some of the differences observed in dilute and crowded environments in our experiments in Chapter 7.

To gain an understanding of how N protein compacts the viral genome, studies using optical
tweezers and long DNA strands have highlighted the capability of N protein to bind to, and compact nucleic acids. Specifically, the work of Morse et al. focused on the full-length protein and demonstrated that the protein compacts DNA in a multi-step process. Furthermore, they show that the protein binds to and compacts ssDNA even at protein concentrations as low as 1 nM. Truncations of the protein including the RBD alone, RBD-LINK, dimerization domain alone, or dimerization domain and CTD, show that the presence of the linker enhances the compaction of the DNA compared to RBD alone, with minimal effects compared to the dimer with and without the linker. Overall, their work suggested that the RBD is responsible for initial binding to the nucleic acids, with strong preferences to the disordered linker. The dimerization domain constructs were required for the second event consisting of the fast compacting. Furthermore, it was noted that at higher concentrations, compaction occurred quicker, indicating a role of self assembly in comparing the nucleic acids. Though this study provides a model by which compaction of DNA is driven at different steps by N protein, work using viral RNA will provide a more precise mechanistic understanding of the role of each region of N protein.

**Figure 8.1 N protein-nucleic acid interactions** A) Representative EMSA gel with full-length N protein and RNA. Modified from Ye et al. B) Representative HSQC for the RBD of N protein (top) and corresponding CSPs per residue (bottom). Modified from Dinesh et al. C) Representative extension traces for the full length protein (left) and two truncations (left). Modified from Morse et al.
8.3 SARS-CoV-2 Nucleocapsid with other proteins.

During the viral life cycle, it is essential that the N protein maintains interactions with other components of the virus, aside from the nucleic acid. Of these interactions, the one with the SARS-CoV membrane (M) protein interaction is crucial for assembly of the virion\textsuperscript{26,27}. N-M interactions have recently been probed using GFP-tagged M protein CTD (aa 104–222 of 222) and Cy3-N protein\textsuperscript{28}. It was found that in isolation, and with RNA, the M-CTD did not form any phase separated condensates. However, in the presence of Cy3-N, even in the absence of RNA, N and M interactions drove phase separation. While N has been proposed to phase separate on its own\textsuperscript{14,23}, the enrichment in both M and N protein in the dense phase when mixed in the same solution suggests that the interaction is crucial. Using truncations of N protein, it was found that deletion of the C-terminal IDR (CTD) of N protein reduced the capability of N-M phase separation, and deletion of the dimerization domain and CTD completely eliminated phase separation at the concentrations tested, suggesting that the dimerization domain and CTD of N protein interact with the CTD of the membrane protein.

Aside from the M protein, N protein has also been shown to interact with the Spike protein\textsuperscript{29}. Immunoprecipitation assays using anti-spike mAb showed indirect potential interactions with the SARS-CoV-2 N protein. Furthermore, synthetic peptides of the CTD of the spike protein competed for interaction with N protein.

N protein does not only interact with the structural components of the virus, but recent works in vitro and in silico have shown that the N protein directly binds to non-structural protein 3 (nsp3)\textsuperscript{30–32}. In silico studies suggested that residues at N protein’s CTD might be involved in interacting with nsp3, a member of the replication transcription complex for SARS-CoVs RNA processing.
*In vitro* studies using nuclear magnetic resonance (NMR) spectroscopy and isothermal titration calorimetry (ITC) demonstrate a direct interaction of the N protein with nsp3a. Specifically, these studies suggest that the disordered linker region of N protein is required to bind the ubiquitin-like folded region of nsp3. Furthermore, NMR results indicate that the interaction is dynamic.

![Figure 8.2 N protein interactions](image)

**Figure 8.2 N protein interactions** A) Phase separation of FL and and the CTD of M protein (top). No phase separation observed when the C-terminal half of N protein is deleted (bottom). Modified from Wang et al. B) Representative HSQC of nsp3 when bound to the linker of the N protein. Modified from Bessa et al.

Overall, the N protein is a complex, multivalent protein that interacts with many components of the virus, including other structural and non-structural proteins. Along with components of the virus, recent works have demonstrated the ability of the N protein to interact with G3BP1 (see section 8.4), a protein not typically involved in the coronavirus viral cycle. Liu *et al.* proposed that interaction of N with G3BPs facilitates viral replication. *In vitro* work using immunoprecipitation and surface plasmon resonance (SPR) showed the interaction of the N protein with both G3BP1 and G3BP2. The same study analyzed lung tissue samples from
infected mice and humans with SARS-CoV-2 and found that the two were co-localized. Truncation analysis confirmed that residues 1-43 contributed to the interaction, but unlike the previous study mentioned, this group identified residues 255-364 (dimerization domain) to be crucial for the interaction. Understanding the multivalent nature of N protein and its capabilities in enhancing viral potency is considered to be of great importance for the development of potential inhibitors and therapies for future coronaviruses.

8.4 SARS-CoV-2 Nucleocapsid phase separation

Phase separation is a phenomenon where two or more components in a solution, under specific conditions, can demix into two or more phases\textsuperscript{38-40}. Phase separation has been proposed as a key mechanism in controlling the assembly and functioning of biomolecular condensates\textsuperscript{41-43}. Biomolecular condensates are not extraneous to viral replication and they have emerged as an intriguing class of intracellular organelles where viral protein components hijack existing cellular condensates or lead to the formation of new ones\textsuperscript{44-48}. For this reason, a certain attention has been dedicated to investigating the propensity of phase separation for the SARS-CoV-2 nucleocapsid protein, \textit{in vitro} and \textit{in living cells}. Like other well known phase-separating proteins, such as TRF2\textsuperscript{49} and G3BP1\textsuperscript{50,51}, the N protein contains an RNA binding domain, a dimerization domain, and intrinsically disordered regions enriched in positively charged residues. All these factors pointed to a strong phase separation propensity of the system. Contextually to our observation of N protein phase separation with RNA, many other groups have observed phase separation of the protein both \textit{in vitro} and \textit{in living cells}.\textit{In vitro} phase separation assays done by Iserman \textit{et al.}\textsuperscript{23} showed that N protein phase separates in absence of ligands in a temperature dependent manner at a concentration of 1-4 μM. Given that
N protein’s propensity to phase separate increases with increasing temperature, it has been proposed as an example of a system exhibiting a lower critical solution temperature (LCST) phase separation. In other words, there is a critical temperature above which the protein exhibits demixing. Based on Iserman et al., this temperature is \( \sim 45 \, ^\circ C \). The molecular properties controlling N protein’s ability to phase separate are not well understood. Interestingly, my work on the dimerization domain discussed in chapter four demonstrates that dissociation of dimer protein occurs above 37 \(^\circ C\) at similar conditions to the one used for phase separation (1\(\mu\)M). Destabilization of the dimer complex may favor an increased multivalence of the protein. Even with the addition of nucleic acid (SL5B or U10), we do not observe either a stabilization or de-stabilization of dimer at 37 \(^\circ C\).

To study the importance of the different regions of the N protein, Chen et al.\(^{12}\), tested a variety of truncations spanning deletions of the folded domains to the domains in isolation. They found that deletion of the RBD did not significantly alter the phase separation, while the RBD in isolation did not phase separate at all under the conditions tested. However, deletion of the dimer domain resulted in decreased phase separation, but the dimerization domain in isolation had a very weak capability of phase separating under the conditions tested. This observation indicates that a dimer is not completely necessary for the protein to undergo phase separation.

In vitro turbidity measurements and in cell work done by Savastano\(^{18}\) et al. suggest that alone, N protein does not have a strong propensity to phase separate on its own. However, in the presence of nucleic acids (polyU in vitro), they notice a strong increase in the turbidity of the solution, indicating the presence of phase separation. In cell, using HeLa cells, they observe Alexa-488 labeled N protein condensates co-localizing with G3BP1 under stress conditions.
Additional work observing N protein-G3BP1\textsuperscript{36} in cell phase separation, found that the intrinsically disordered NTD was crucial for droplet formation, as a deletion of residues 1-39 impaired aggregation and droplet formation in HeLa Cells. Furthermore, when cells were given arsenite, a known poison which induces stress in cells, the GFP-N protein localized to puncta (stress granules, which are usually enriched in G3BP1/2). Co-localization of endogenous G3BP1 and overexpressed N protein was observed in the HeLa cells, indicating an interaction of the two. Aside from the interactions of N protein with G3BP1, N protein has been shown to phase separate with the M protein (see above, 8.3), but also with a variety of nucleic acids, ranging from synthetic oligos to portions of the SARS-CoV-2 genome. Studies using synthetic oligos\textsuperscript{12,37,52} show that the N protein can phase separate under a variety of conditions in vitro including fairly high salt concentrations\textsuperscript{52} >150mM. Furthermore, these phase separated droplets are dynamic, showing full recovery after photobleaching within tens of seconds\textsuperscript{12,37,52}. When comparing single-stranded (ss) oligos of varying lengths of uracil, the authors note that there was a length-dependent behavior on the properties of the phase-separated droplets. It was found that the longer ss-RNAs formed larger droplets with the N protein\textsuperscript{12}.

Studies using viral RNA\textsuperscript{16,53,54} have been performed both in vitro and in cell, and similar to work done with synthetic oligos, highlight the capability of the N protein to undergo phase separation at fairly low concentrations of protein and RNA. Studying phase separation of GFP-N with viral RNA show that the protein (concentration dependant) phase separates even at high salt concentrations >300 mM, within the pH range of 4.5-7.5, and at temperatures ranging from 4 °C to 40 °C.\textsuperscript{53} Furthermore, the same study identified the dimerization domain to be an essential driver of phase separation under the conditions tested.
Knowing that N protein interacts so strongly with nucleic acids in a phase separated state, phase separation has been used as an amplifier of protein-nucleic acid interactions. In particular, this is useful when trying to test small molecules that could inhibit such interactions. Jack et al.\textsuperscript{14} demonstrated the use of small molecules to inhibit phase separation. In particular, they noted that 10\% 1,6 hexanediol was sufficient to obliterate phase separation under the conditions tested.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure8.3.png}
\caption{N protein phase separation disrupted by small molecules. Modified from Jack et al\textsuperscript{14}.}
\end{figure}

Given the nature of the molecule, it was suggested that phase separation of N protein was driven by hydrophobic interactions. Furthermore, nilotinib did not suppress phase separation of N protein, but instead changed the morphology of the condensates.

Overall, the strong propensity of the SARS-CoV-2 N protein for phase separation further stressed the role of multivalence encoded in the protein, as discussed already in the context of protein-protein interactions.
8.5 Future directions.

Overall, my thesis work shed light on the role of intrinsically disordered regions in the context of the SARS-CoV-2 N protein, providing quantification of protein conformations, dynamics, and interactions with nucleic acids. Though much effort was focused on characterizing biochemical and biophysical aspects of the N protein, there are many aspects that remain to be explored.

While I started to address the mechanism of interaction of the protein with nucleic acids using a “divide and conquer” approach and focusing on the contribution of the disordered NTD with respect to the RBD, more work is needed to contextualize the role of other disordered regions with respect to the RBD and the dimerization domain. Currently, there is a strong interest in better understanding the role of the linker region, as it is highly enriched in arginine residues, as well as serine residues, which have been proposed to undergo phosphorylation. Work done by Bessa et al. (see section 8.3) highlighted the importance of the linker region and its functional role in the viral environment. Work by Wu et al. and Morse et al. further demonstrated the importance of the linker region for RNA binding, both double-stranded and single-stranded RNA. Preliminary studies (Appendix A) using smFRET suggest that the addition of the linker has an even greater enhancement on the protein affinity for RNA. Comparison of different labeling positions can enable characterizing the role of the linker in binding to RNA and how this may alter binding site size, stoichiometry, and local conformations of the protein.

Ultimately, the aim is to understand how the full-length protein binds to and interacts with RNA. Preliminary measurements using smFRET suggest that the protein oligomerization is favored in the presence of nucleic acids, lowering the dimerization binding constant to picomolar affinities (Appendix B). These observations are in contrast with the proposed idea that nucleic acids may
disfavor oligomerization\textsuperscript{17}. Under optimized solution conditions (temperature, length of the nucleic acid, and salt), single-molecule studies of stable dimers can provide insights on the conformational changes associated to the dimer form when binding to RNA.

Genome packaging and condensation is one of the essential roles of the nucleocapsid protein in all coronaviruses, including SARS-CoV-2. The actual mechanism by which the N protein is able to condense and package a 30kb genome remains undetermined. Recent work has begun to investigate the mechanism by which N protein can condense long strands of nucleic acid, using DNA\textsuperscript{25}. Future studies using RNA and eventually fragments of the genome of SARS-CoV-2 will provide a more accurate representation of how N protein directly condenses the long genome to fit into a 100 nm virion\textsuperscript{57}. Using optical tweezers, viral RNA fragments can be inserted between tethers and attached to beads. Similar to Morse \textit{et al.}\textsuperscript{25}, adding N protein (or truncations variants) and monitoring changes in force vs extension of the RNA will provide insight into the N protein’s ability of compacting the genome\textsuperscript{25}. Extending these experiments to sequences of other coronaviruses where a different compaction of the genome is observed (larger helical compaction) may provide insights on the mechanisms controlling the conformation of nucleic acid inside the virion.

When considering the role of phase separation in the function or dysfunction of the N protein, one poses the question as to what can be used to inhibit the process from occurring? Work done by the Yildiz lab\textsuperscript{14} demonstrated that certain small molecules interfered with N proteins ability to phase separate either alone or with polyC. To investigate the role of small molecules, a drug screen using hundreds/thousands of small molecules would be necessary to evaluate potential candidates. Condensates used would be composed of N protein and a synthetic oligonucleotide, such as polyU, and eventually use of a genomic RNA for better accuracy of the
Once candidates are chosen, further studies such as binding using ITC or smFRET can be integrated to elucidate the mechanism by which the drug interferes.

In conclusion, single-molecule fluorescence spectroscopy has proven to be a powerful toolbox for investigating protein conformations and interactions of disordered regions. The work has provided new insights on the molecular interactions encoded in the SARS CoV 2 N protein and has paved the way to quantitative studies of interactions with other binding partners, viral genome RNA, and small molecules.
8.6 References


separation stimulated by RNA and partitions into phases of human ribonucleoproteins. doi:10.1101/2020.06.09.141101.


39. Gelb, L. D., Gubbins, K. E., Radhakrishnan, R. & Sliwinska-Bartkowiak, M. Phase


Appendix A

The LINK region significantly enhances the binding affinity of N protein to RNA
In chapter five, I described the role of the intrinsically disordered NTD in aiding the RBD in binding to RNA using single-molecule FRET. In Appendix A, I have summarized preliminary single-molecule FRET measurements highlighting the role of the serine arginine-rich linker in further aiding these interactions.

Here, I used a truncation of the full-length protein containing the NTD, the RBD, and the linker, (NTD-RBD-LINK) labeled at positions 172 and 245 to monitor the conformation of the linker, as described in chapter three in 50 mM Tris, pH 7.4, room temperature (295 ± 0.5 K).

In the absence of RNA, the NTD-RBD-LINK adopts a narrow population with a mean transfer efficiency of ~ 0.80 ± 0.03 (Fig. A1 and A2). This is consistent with previous observations of the linker domain. Titrating a single-stranded synthetic oligonucleotide of poly(rU) (Fig. A1) uracil (rU)20 (Fig. A2) or results in the rise of a second population. In the case of poly(rU), the bound population is characterized by a mean transfer efficiency of 0.57 ± 0.03. We fit the data to a 1:1 binding model (eq. AA1), and obtain an association constant $K_A$ of 0.04 ± 0.01 nM, which is about 10-fold stronger than the one quantified under identical conditions for the NTD-RBD (Figure A1 and Table A1).
Figure A1: Binding of Poly(rU) to the NTD-RBD-LINK

A) Representative histogram of transfer efficiencies of the NTD-RBD-LINK in absence of and with increasing concentrations of poly(rU). B) Fraction bound vs concentration of poly(rU) for the NTD-RBD-LINK (green) and NTD-RBD (red).

In the presence of saturating (rU)$_{20}$, we observe an extended state with a mean transfer efficiency of 0.65 ± 0.03. While it is extended, it is not as extended as it is in the presence of poly(rU), which indicates that the expansion of the disordered linker is dependent on the length of the nucleic acid polymer, similar to that observed for the NTD in chapter five. When comparing the affinities, we obtain an association constant $K_A$ of 0.024 ± 0.002nM, which is about 5-fold tighter than the NTD-RBD (Figure A2 and Table A2). For (rU)$_{20}$, which is significantly shorter than poly-(rU), we cannot exclude that multiple copies can bind to a single protein. To confirm the exact stoichiometry of the protein:RNA complex, future work will characterize the binding of RNA in presence of different concentrations of protein by mixing labeled and unlabeled NTD-RBD-LINK constructs.
Figure A2: Binding of \((rU)_{20}\) to the NTD-RBD-LINK  

A) Representative histogram of transfer efficiencies of the NTD-RBD-LINK in absence of and with increasing concentrations of \((rU)_{20}\). B) Fraction bound vs concentration of \((rU)_{20}\) for the NTD-RBD-LINK (light green) and NTD-RBD (teal).

To tease out the contribution of other domains, such as the RBD and NTD, to the interaction of nucleocapsid protein with RNA, future studies will use constructs of the NTD-RBD-LINK with the NTD labeled, RBD labeled, and LINK labeled. By comparing the affinities and modes of interaction of the three domains, we can more precisely determine the affinities of the interactions.

Table A2: Sequence of WT NTD-RBD-LINK. Labeling positions are reported in green.

<table>
<thead>
<tr>
<th>Position</th>
<th>Amino Acid Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MSDNGPQQR NAPRITFGGP SDSTGSNQNG ERSGARSKQR RPQGLPNTA</td>
</tr>
<tr>
<td>51</td>
<td>SWFTALTQHG KEDLKFPGRG QGPINTNSSP DQIGYYRRA TRRIRGGDGK</td>
</tr>
<tr>
<td>101</td>
<td>MKDLSRPRWF YYLGTGPEAG LPYGANKDGI IHWATEGALN TPKDHIGTRN</td>
</tr>
<tr>
<td>151</td>
<td>PANNAAILQ LPQTTLPKG PÆAEGSRGGS QAASSRSSRS RNASSRNSTPG</td>
</tr>
<tr>
<td>201</td>
<td>SSRTSPARM AGNNGDAALA LLLLRLNQL ESKMSGKGOQ QGQTVF</td>
</tr>
</tbody>
</table>
Table A2: Association constants for the NTD-RBD-LINK in comparison to the NTD-RBD (chapter four).

<table>
<thead>
<tr>
<th></th>
<th>Ka (nM⁻¹)</th>
<th>(rU)₂₀ (molecules)</th>
<th>Poly(rU) (nt)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTD-RBD</td>
<td>0.0054 ± 0.0007</td>
<td>0.0040 ± 0.0003</td>
<td></td>
</tr>
<tr>
<td>NTD-RBD-LINK</td>
<td>0.024 ± 0.002</td>
<td>0.036 ± 0.01</td>
<td></td>
</tr>
</tbody>
</table>

Binding of Nucleic Acid: When the fraction of bound protein \( f_b \) is directly estimated from the distribution of transfer efficiencies, titration curves were analyzed according to:

\[
f_b = \frac{K_A[RNA]_{tot}}{1 + K_A[RNA]_{tot}} \quad \text{Eq. A1}
\]

where \( K_A \) is the association constant and \( [RNA]_{tot} \) is the total concentration of RNA. Since the protein and (rU)₂₀ and poly(rU) form a complex that can be observed with a 1:1 stoichiometry and we measure directly the fractions of free and bound populations not just a signal variation, we can determine \( K_A \).

References

Appendix B

Nucleic acid binding induces oligomerization of the SARS-CoV-2 Nucleocapsid Protein
In chapter four, I characterized the conformational changes of the dimerization domain upon dimerization. In a supporting figure from chapter four (S4.4) using single-molecule FRET, I show that binding to nucleic acids does not dissociate the dimer, as suggested previously 1. In Appendix B, I present preliminary single-molecule FRET measurements suggesting that binding of RNA lowers the dissociation constant of dimerization of the full-length protein, with dimer formation becoming prominent at low concentration as 100 pM.

Here, I use the full-length protein labeled at positions 1 and 68 to monitor the conformation of the NTD, NTD FL, as described in chapter three (in 50 mM Tris, pH 7.4, room temperature (295 ± 0.5 K)). An advantage of single-molecule FRET is the capability of determining the stoichiometry of the labeled sample given by Eq. B1:

\[ S = \frac{I_D}{(I_D + \gamma_{PIE} I_A)} \]  

where \( I_D \) and \( I_A \) represent the total intensities observed after donor and acceptor excitation, \( \gamma_{PIE} \) is a correction factor accounting for differences in donor and acceptor detection efficiency and laser intensities. Donor-only and acceptor-only labeled molecules have a stoichiometry ratio of 1 and 0, respectively, while a (monomeric) 1:1 (acceptor:donor) labeled protein exhibits a stoichiometry of 0.5. In the event that multiple labeled proteins associate with one another, multiple stoichiometric cases arise (Fig. B1). In the case of a dimer, there are three possibilities. The first being two proteins labeled with 1 acceptor and one donor 1:1 yielding a 0.5 stoichiometric ratio. This can be distinguished from the 1:1 labeled monomer by studying the dependence of this population in presence of denaturant. The second case involves the binding of a 1:1 donor:acceptor protein with a donor-only protein, yielding a stoichiometry ratio of 0.75. The third case involves the binding of a 1:1 donor:acceptor protein with an acceptor-only protein,
yielding a stoichiometry ratio of 0.25. For this calculation we assumed the donor-only and acceptor only species to be both double-labeled with only donor or acceptor dyes. This is the most likely scenario based on sequential labeling of the protein, where proteins are first sub-stoichiometrically labeled with donor, then purified, and labeled with an excess of acceptor. If only a single acceptor or a single donor is present in the donor- and acceptor- only populations, the stoichiometry ratio will result in a value of ~0.67 and ~0.33. (Fig. B1). For the case of the NTD\textsubscript{FL}, in the absence of nucleic acids, a narrow histogram centered at 0.5 stoichiometry ratio is observed. With the addition of poly(rU), a broadening of the 0.5 stoichiometry population is observed. For the complex containing 120 nM poly(rU) and 500 nM poly(rU), there is a broadening of the stoichiometry ratio seen in the histograms, as well as a diagonal stretching of the population on the 2D plot. In both cases with the RNA, populations can be extracted at 0.75 and 0.25 stoichiometry ratio, indicating self-assembly of at least two or more labeled molecules.

Future studies will aim to optimize conditions where the protein can be studied in the context of the full-length protein when binding to nucleic acids.
Figure B1: N protein is monomeric at 100 pM concentrations in the absence of RNA. Transfer efficiency vs stoichiometry plot with expected values based on the number of labels in solution. The orange distribution represents the measured value for the NTD_{FL} in 50mM Tris, pH 7.4.
**Figure B2:** N protein oligomerizes at 100 pM concentrations in the presence of RNA. Representative stoichiometry ratio histograms (left) and stoichiometry vs transfer efficiency 2D plots (right) for the NTD$_{FL}$ construct, in the absence (top) and presence of increasing concentrations of poly(rU). The stoichiometry ratio is estimated from the bursts, as detected after donor excitation (cyan distribution) or after both donor and acceptor excitation (red distribution). A broadening of the peak centered at 0.5 stoichiometric ratio is observed in the presence of poly(rU), indicating more than just a 1:1 ratio of labeled protein to RNA (left). Similarly, the 2D plots exhibit a broadening of the population, which if 1:1, would be centered at 0.5. For samples containing 120 nM and 500 nM poly(rU), there is a clear extension of the population reaching both a stoichiometric ratio of 0.75 and 0.3.

**References**

Appendix C

Apolipoprotein E4 has extensive conformational heterogeneity in lipid-free and lipid-bound forms


My contributions to this manuscript include protein purification and data collection
Apolipoprotein E4 has extensive conformational heterogeneity in lipid-free and lipid-bound forms

Melaissa D. Stuchell-Brereton\textsuperscript{a,b}, Maxwell I. Zimmerman\textsuperscript{a,b}, Justin J. Miller\textsuperscript{a,b}, Upasana L. Mallimadugula\textsuperscript{a,b}, J. Jeremías Incicco\textsuperscript{a,b,1}, Debjit Roy\textsuperscript{a,b,1}, Louis G. Smith\textsuperscript{a,b}, Jasmine Cubuk\textsuperscript{a,b}, Berevan Baban\textsuperscript{a}, Gregory T. DeKoster\textsuperscript{c}, Carl Frieden\textsuperscript{d}, Gregory R. Bowman\textsuperscript{a,b}, and Andrea Soranno\textsuperscript{a,b,1}

Edited by William Eaton, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD; received September 10, 2022; accepted January 4, 2023

The ε4-allele variant of apolipoprotein E (ApoE4) is the strongest genetic risk factor for Alzheimer’s disease, although it only differs from its neutral counterpart ApoE3 by a single amino acid substitution. While ApoE4 influences the formation of plaques and neurofibrillary tangles, the structural determinants of pathogenicity remain undetermined due to limited structural information. Previous studies have led to conflicting models of the C-terminal region positioning with respect to the N-terminal domain across isoforms largely because the data are potentially confounded by the presence of heterogeneous oligomers. Here, we apply a combination of single-molecule spectroscopy and molecular dynamics simulations to construct an atomically detailed model of monomeric ApoE4 and probe the effect of lipid association. Importantly, our approach overcomes previous limitations by allowing us to work at picomolar concentrations where only the monomer is present. Our data reveal that ApoE4 is far more disordered and extended than previously thought and retains significant conformational heterogeneity after binding lipids. Comparing the proximity of the N- and C-terminal domains across the three major isoforms (ApoE4, ApoE3, and ApoE2) suggests that all maintain heterogeneous conformations in their monomeric form, with ApoE4 adopting a slightly more compact ensemble. Overall, these data provide a foundation for understanding how ApoE4 differs from nonpathogenic and protective variants of the protein.

Significance

Despite being identified as the strongest genetic risk factor for Alzheimer’s disease more than 20 years ago, a connection between the biochemical properties of apolipoprotein E (ApoE) and its role in the disease remains elusive. This is largely due to the limited structural information available for the different forms adopted by the protein (monomer, dimer, tetramer, and lipid bound) across pathogenic and nonpathogenic variants. Here, we provide the characterization of the full-length pathogenic ApoE4 in its monomeric form both in the presence and absence of lipids. We demonstrate that the protein does not adopt a single structure, but a multiplicity of different conformations, which impacts the interpretation of the structure-function mechanism of ApoE.


The authors declare no competing interest.

This article is a PNAS Direct Submission.

Copyright © 2023 the Author(s). Published by PNAS. This open access article is distributed under Creative Commons Attribution License 4.0 (CC BY).
to the four-helix bundle. Ensemble Förster resonance energy transfer (FRET) and Electron Paramagnetic Resonance (EPR) studies (24) suggest that ApoE4 forms a close contact between the four-helix bundle and the C-terminal domain, whereas ApoE3 explores more open conformations. This is at odds with the compact set of structures determined by Nuclear Magnetic Resonance (NMR) on a monomeric ApoE3-like variant (22). Recent Hydrogen Deuterium Exchange Mass Spectrometry (HDX-MS) experiments identified isoform-dependent differences in solvent accessibility of the four-helix bundle, hinting that single amino acid substitutions affect the ability of the C-terminal domain to shield specific regions of the four-helix bundle (19). However, the interpretation of ensemble FRET, EPR (24), and HDX-MS experiments (19) is complicated by the fact that measurements were performed under conditions in which the protein is a stable tetramer (16, 19) and, therefore, are not representative of the conformations of the protein in its monomeric form. The same limitation applies to previous investigations of the folding stability of the protein domains (16, 25–27) (SI Appendix, Table S3).

Here, we circumvent these experimental difficulties by harnessing single-molecule fluorescence spectroscopy, an approach that enables working at sufficiently low protein concentrations to avoid oligomerization and directly access the protein in its monomeric form (SI Appendix, Fig. S1). Single-molecule FRET provides a direct readout on the conformations and stability of specific domains within full-length ApoE4 in both the lipid-free and lipid-bound states. We further complement single-molecule observations with molecular dynamics (MD) simulations to obtain an atomically detailed representation of protein conformations that is consistent with our experimental data.

Results
To study the conformations of ApoE4 via single-molecule FRET, we designed, expressed, and purified five distinct full-length double-cysteine mutants of the protein (Fig. 1A and SI Appendix). We used the ApoE3-like structure determined by NMR (22) (Fig. 1B) as a blueprint to guide our choice of labeling positions, such that each dye pair combination probes one of the four regions of the protein.

Folding and Stability of the Four-Helix Bundle. We first focus on the ApoE46,165 construct, where labeling positions are located in the random coil between helices H2 and H3 as previously defined (20, 30) and serves as a common reference point to investigate the folded N-terminal domain from two different perspectives. When paired with position ASC (ApoE46,165), which is situated upstream of the start of the H1 helix, A86C monitors the conformational properties and folding stability of the N-terminal tail. When paired with position G165C (ApoE46,165), which is located at the end of the H4 helix, A86C provides a readout for the folding of the four-helix bundle (22, 30). Positions G182C and A241C (ApoE4182,241) allow monitoring the behavior of the hinge domain with respect to the C terminus, while positions S223C and A291C (ApoE4223,291) provide information on the structural properties of the C-terminal domain. Finally, probe positions located at A5C and A241C (ApoE45,241,C) allow us to monitor long-range interactions between the N- and C-terminal domains. (8) One hundred and eighty-degree rotated views of the monomeric ApoE3-like variant NMR structure (PDB: 2L7B) highlighting labeling positions shown in orange. Structure color differentiates the major protein domains described in A.

Fig. 1. Protein structural regions and single-molecule constructs of full-length ApoE4. (A) Schematic representation of the secondary structure content in ApoE4 based on the NMR structure (Protein Data Bank (PDB): 2L7B) of the ApoE3-like variant with corresponding designations and identification of the major protein domains: N-terminal tail (gray), four-helix bundle (teal), hinge region (green), and C-terminal domain (light purple). Helical notations are reported for each helix. Labeling positions are identified on the linear sequence by green and red dots (the color scheme is only indicative of FRET labels and not of residue labeling for a specific fluorophore). Yellow dots identify the mutations associated with ApoE3 and ApoE2 variants. Position A86C is located in the random coil between helices H2 and H3 as previously defined (20, 30) and serves as a common reference point to investigate the folded N-terminal domain from two different perspectives. When paired with position ASC (ApoE46,165), which is situated upstream of the start of the H1 helix, A86C monitors the conformational properties and folding stability of the N-terminal tail. When paired with position G165C (ApoE46,165), which is located at the end of the H4 helix, A86C provides a readout for the folding of the four-helix bundle (22, 30). Positions G182C and A241C (ApoE4182,241) allow monitoring the behavior of the hinge domain with respect to the C terminus, while positions S223C and A291C (ApoE4223,291) provide information on the structural properties of the C-terminal domain. Finally, probe positions located at A5C and A241C (ApoE45,241,C) allow us to monitor long-range interactions between the N- and C-terminal domains. (8) One hundred and eighty-degree rotated views of the monomeric ApoE3-like variant NMR structure (PDB: 2L7B) highlighting labeling positions shown in orange. Structure color differentiates the major protein domains described in A.
A5C is situated upstream of the start of helix H1 and when paired with A86C monitors the conformational properties of the N-terminal tail (Fig. 1). Single-molecule FRET measurements of ApoE45,86 reveal two distinct populations in equilibrium under aqueous buffer conditions. The more abundant population has a mean transfer efficiency of 0.61 ± 0.02, while the less abundant population sits at 0.21 ± 0.05 (Fig. 2). Comparing the donor lifetime vs. transfer efficiency indicates that the population at low transfer efficiency is compatible with a rigid distance where positions 5 and 86 are located ~7 nm apart (SI Appendix, Fig. S2). Conversely, the population at higher transfer efficiency follows the expected trend of a dynamic conformational ensemble, that is, an ensemble of interdye distances that are sampled in a timescale much shorter than the residence time of the protein in the confocal volume. Interestingly, the results are better described using a wormlike chain distribution with persistence length $l_p$ (an estimate of the minimal flexible segment) equal to 2.5 nm and contour length $l_c$ (the maximum extension of the probed region) equal to 7.7 nm (SI Appendix, Fig. S2). Note that this contour length is just ~25% of the contour length expected for an equivalent fully disordered region, suggesting that secondary structure formation occurs within this population. To further test for the presence of secondary structure, we investigated the effect of denaturant. We observe that the population at low transfer efficiency is completely destabilized at 0.5 M GdmCl and that the population at higher transfer efficiency tends to shift toward lower values with increasing denaturant (Fig. 3). This result is consistent with a population that is not completely structured and contains a certain degree of flexibility (31). Interestingly, a noticeable shift in the mean transfer efficiency of this population occurs between 1 and 2 M GdmCl accompanied by a change in the width of the distribution (SI Appendix, Figs. S5 and S6). We interpret this behavior as the result of the coexistence of two populations with similar transfer efficiencies within the same observed peak. By fitting two independent populations within the mean transfer efficiency distribution (Fig. 2), we obtain a midpoint of the transition ($c_{1/2}$) equal to 2.06 ± 0.01 M and a $\Delta G_0$ equal to 5.2 ± 0.2 RT (Fig. 3C, compare alternative analysis in SI Appendix). This observation can be understood considering that positions 5 and 86 sample not only the N-terminal tail but also helices H1 and H2 of the four-helix bundle.

**Hinge Region.** Positions G182C and A241C (ApoE4182,241) allow monitoring of the behavior of the hinge domain with respect to the C terminus. Analysis of the corresponding transfer efficiency histograms reveals an asymmetric distribution of transfer efficiencies under aqueous buffer conditions. We analyze the asymmetric distribution in terms of two distinct populations (Fig. 2). The population associated with lower mean transfer efficiency ($E = 0.62 ± 0.02$) accounts for 60% of the observed molecules, whereas the high transfer efficiency population ($E = 0.83 ± 0.02$) accounts for the remaining 40%, corresponding to a free energy difference between these states of 1.0 ± 0.2 RT (SI Appendix, Tables S1 and S2). The asymmetry of the distribution...

---

**Fig. 2.** Single-molecule fluorescence experiments of lipid-free full-length ApoE4. Transfer efficiency histograms for selected bursts with fluorescence stoichiometry ratio between 0.3 and 0.7 across the five full-length constructs ApoE45,86 (gray), ApoE486,165 (teal), ApoE4182,241 (green), ApoE223,291 (light purple), and ApoE486,241 (magenta) at increasing concentrations of GdmCl. Under aqueous conditions, all histograms reveal coexistence of multiple states. Lines are visual guides for contrasting the native and completely unfolded configurations in each construct.
shown only for population fractions larger than 10% or when analyzed to simply provide a visual guide. Mean transfer efficiencies are in transfer efficiency histograms of Fig. 2. Solid lines connect mean transfer, Fig. S4 for free energy diagrams of each single construct. Different folded states identified in the N-terminal tail, hinge region, C-terminal and from long-range measurements. Solid lines represent the equilibrium of identified states in the four-helix bundle, hinge, and N and C termini. Dashed lines are used to underline that these trend expected for a dynamic ensemble, excluding the formation of stable contacts that would give rise to rigid configurations of the protein. Interestingly, a small percentage of the collapsed state, represented by the high transfer efficiency population, persists up to concentrations of denaturant that are compatible with the unfolding of the N-terminal domain. This implies the formation of a small fraction of more compact configurations of the protein that, nevertheless, retain a dynamic nature.

**Proximity of the N- and C-Terminal Domains.** To better understand whether the four-helix bundle and the C-terminal region form stable contacts and to which extent they are brought in close proximity, we investigate the transfer efficiency distribution between A86C and A241C (ApoE486,241). Under aqueous buffer conditions, we observe the occurrence of at least three populations with corresponding mean transfer efficiencies of 0.24 ± 0.01, 0.59 ± 0.02, and 0.87 ± 0.02 (Fig. 2). This is consistent with observation of multiple configurations in both the hinge and C-terminal regions. When comparing donor lifetime and transfer efficiency (SI Appendix, Fig. S2), all these populations lie on the trend for a dynamic ensemble and whose relative abundance is stabilized by increasing concentrations of denaturant. Both elements point toward a population that is more flexible and, at least, partially disordered, as further supported by the continuous shift of the peak from high to low transfer efficiencies when tuning the solvent quality from a poorer solvent (aqueous buffer) to a better solvent (GdmCl). The increased broadening of the width of this population below 1 M GdmCl (SI Appendix, Fig. S5), which exceeds the width measured for other constructs, points to an increased heterogeneity due to structure formation. This is consistent with previous characterizations of the C-terminal region, where destabilization of the secondary structure was observed above 1 M GdmCl (25, 26). The third population at ~0.85 mean transfer efficiency represents more compact configurations of the C-terminal domain, where positions 223 and 291 are brought in close proximity. Interestingly, the small relative abundance of this population decreases above 1 M GdmCl and disappears at 2.75 M GdmCl (Fig. 3). This regime of concentrations coincides with the folding of the four-helix bundle and mirrors that observed for the hinge region, suggesting that folding of the four-helix bundle induces conformational changes in the C-terminal region.

**MD Simulations Confirm Structural Heterogeneity.** To gain insights into the structural details of the conformational ensemble of ApoE4, we performed all-atom MD simulations of the full-length protein on the distributed computing platform Folding@home for a total aggregated time of 3.45 ns. We then constructed a Markov state model to bin the conformational ensemble into
unique states. For each observed state, we modeled fluorophores onto the labeling positions post hoc and reconstructed a set of transfer efficiency histograms that accounts for shot noise and the kinetic averaging of conformations in the observation timescale (SI Appendix). The comparison between simulated and measured transfer efficiency histograms is shown in Fig. 4A. We find good agreement between both datasets, including the occurrence of a multimodal transfer efficiency distribution for ApoE223,291, which

![Fig. 4. Comparison between transfer efficiency histograms in single-molecule measurements and MD simulations for lipid-free ApoE4. (A) Single-molecule FRET histograms of the five investigated constructs ApoE4_5,86 (gray), ApoE4_86,165 (teal), ApoE4_182,241 (green), ApoE4_223,291 (light purple), and ApoE4_86,241 (purple) are compared with equivalent distribution of transfer efficiencies computed from MD simulations (white). (B) Distance pair correlations from MD simulations contrasting the distance r_86, 165 with the distances r_5,86, r_182,241 and r_223,291, r_86,241. Colored boxes (yellow, red, and brown) identify three major configuration regimes of the four-helix bundle and corresponding changes in the other protein regions. (C) The 15 most probable configurations for each of the three states closed, open, and extended, as identified from the data in panel B. Position of 86 and 165 fluorophores is highlighted in orange, whereas the N-terminal tail is displayed in gray, the four-helix bundle in teal, the hinge region in green, and the C-terminal domain in light purple (compare with Fig. 1).]
was not captured in microsecond-long simulations, stressing the importance of an extensive sampling of the energy landscape with long simulation times (SI Appendix, Fig. S7). Deviations in the mean transfer efficiencies and relative abundance of populations are within experimental errors and known limits of comparing these approaches (SI Appendix and SI Appendix, Fig. S8). To better disentangle the conformations underlying the simulated transfer efficiency histograms, we analyzed the simulation data for the occurrence of correlations across all distance pairs (Fig. 4B and SI Appendix, Figs. S9–S12). This analysis reveals three subpopulations associated with the distance between positions 86 and 165 whose mean transfer efficiencies fall within the observed distribution for ApoE(_86,165)_C. The conformational changes in these subpopulations are not restricted to these specific labeling positions but propagate across the entire protein, highlighting correlated changes in the hinge region and antecorrelated ones in the C-terminal domain. In particular, the identified subpopulations in each distance pair correlation parallel the distance and relative abundance trends observed in the experiments. All three identified subpopulations differ from the ApoE3-like NMR structure, where numerous contacts previously identified between the four-helix bundle and the C-terminal domain are not observed even in the more compact conformations (SI Appendix, Fig. S10). Alignment of subpopulation structures reveals how these conformational differences do not stem from varying degrees of secondary structure but propagate across the entire protein, as previously proposed (20, 21, 32), or even across the four-helix bundle, open, where the C-terminal domain is undocked, and extended, where the undocked C-terminal domain adopts more extended conformations. Interestingly, these conformational differences do not stem from varying degrees of secondary structure in the C-terminal domain (SI Appendix, Fig. S11). We further analyzed the simulations to verify whether specific residue contacts are maintained despite the extensive conformational heterogeneity. We identified a set of persistent contacts within the four-helix bundle and the HC1 helix of the C-terminal domain, which suggests that the relative position of HC1 with respect to the four-helix bundle is maintained across all the subpopulations (Fig. 5). At variance with the closed subpopulation, the open and extended ensembles show an increase in the number of contacts of the N-terminal tail with the four-helix bundle and the HC1 helix, which may dictate whether the C-terminal domain docks onto the four-helix bundle. Interestingly, there are no shared contacts across the three subpopulations within the N-terminal tail or the hinge region (Fig. 5, highlighted in yellow), which suggests that these regions are adopting different conformational states. Indeed, the position of the hinge region differs across the three subpopulations and is directed by interactions between the helix H1 and either the N-terminal tail or the four-helix bundle (Fig. 5). Specifically, in the closed configuration, the hinge region mainly interacts with helices H1 and H2, whereas in the open and extended configurations, the hinge explores the surface of helices H2 and H3 with differing extent of specificity. Altogether, MD simulations confirm the experimental observation that lipid-free ApoE4 adopts a dynamic structural ensemble with at least three distinct states.

Conformational Heterogeneity Is Maintained across Isoforms.

We then turn to investigate whether mutations at residue 112 (as in ApoE3) and at residues 112 and 158 (as in ApoE2) alter the proximity of the four-helix bundle and C-terminal region of the protein, as previously proposed (20, 21, 32), or even suppress conformational heterogeneity, as observed in the ApoE3-like NMR structure (22). To this end, we create two constructs ApoE3*(_86,241) and ApoE2*(_86,241) where we insert serine residues in either position 112 or both positions 112 and 158. Serine residues are chosen because they do not interfere with the maleimide chemistry labeling. Importantly, the substitution Arg112Ser is known to replicate the effects of ApoE3 (Arg112Cys), with similar reduced domain interaction (33), lipid- (34) and lipopolysaccharide-binding properties (35), and formation of SDS-resistant complex with ApoA1 that is unique to apoE3 (36).

In our single-molecule FRET experiments, we found that all three constructs exhibit three distinct populations, indicating that conformational heterogeneity is maintained across the isoforms (Fig. 6A). Comparison of lifetime vs. transfer efficiency confirms the dynamic nature of these states. While ApoE4(_86,241) and ApoE3*(_86,241) exhibit similar mean transfer efficiencies, ApoE2*(_86,241) shows a minor shift toward higher mean transfer efficiencies for the population with ~0.6 transfer efficiency (Fig. 6B). All three isoforms are measured from the perspective of the same interdye distance between the N- and C-terminal regions, and therefore, we conclude that ApoE2* adopts slightly more compact conformations than ApoE3* and ApoE4.

Testing Salt Bridge Formation.

Previous experiments proposed a close proximity of residues 76 and 241 (24) in ApoE4 that is helped by a salt bridge formation between residues 61 and 255 (15, 30). Such a close configuration is expected to be altered in ApoE3, leading to more extended configurations. While we did not find the salt bridge in the simulations (SI Appendix, Fig. S10) and we did not observe significant changes in the distribution of transfer efficiencies for ApoE3* and ApoE4, we further tested this hypothesis by introducing a R61T mutation in ApoE4, ApoE4<sub>R61T</sub>_(_86,241) which suppresses the putative salt bridge formation between 61 and 255. As shown in Fig. 6A and B, we do observe a minimal shift toward more expanded conformations. The expansion pertains to the population with a transfer efficiency of ~0.6 and is associated with an average distance of about 5 nm. Given our labeling positions are near residues 76 and 241, our observations suggest that 76 and 241 are in close proximity for only a small fraction of sampled configurations (represented by the high transfer efficiency shoulder). If a stable close configuration of the protein was formed upon salt bridge formation, we would expect to observe a clear change in the relative fractions of each population when suppressing salt bridge formation. Interestingly, the R61T mutation only minimally decreases the fraction of molecules associated with the high transfer efficiency population, suggesting that this population does not represent a salt bridge–dependent conformation. Altogether, our results suggest that, in the monomeric form, the R61T mutation does not introduce significant changes when compared to ApoE3* and ApoE4.

Contribution of Electrostatic Screening.

Given the large proportion of surface-exposed charged residues within the N- and C-terminal regions, we further tested the effect of salt on modulating electrostatic contribution to the conformational ensemble of the protein. Titration of increasing concentrations of NaCl on ApoE4(_86,241) does not significantly alter the proportion of the relative fractions (Fig. 6C), implying that the interactions between the four-helix bundle and the C-terminal domain are not exclusively of electrostatic nature. However, the mean transfer efficiency associated with the major population shifts toward lower values, indicating an expansion of the conformational ensemble upon ion screening of the electrostatic interactions. This suggests that salt concentration can modulate the distal organization of ApoE domains but does not alter the equilibrium between the three major identified states.
Lipid Association of ApoE4. Finally, we turn to investigating how the structural heterogeneity of ApoE4 is impacted by binding to lipids, which reflects the most likely populated configuration under physiological conditions. To this end, we focus on the interaction between the ApoE4 constructs and dipalmitoylphosphatidylcholine (DMPC) liposomes with an average radius of 40 ± 20 nm (Fig. 7A and SI Appendix, Fig. S13). We chose DMPC because it is a good mimic of the lipids found in lipoproteins both in terms of hydrophilic head group and average length of the fatty acid chain (37, 38). Using single-molecule FRET and a high concentration of liposomes (100 μg/mL), we tested whether the labeled constructs could bind to lipids. ApoE4Δ4,8,6(N-terminal tail), ApoE4Δ223,291(C-terminal domain), and ApoE4Δ6,241(long-range contacts) all exhibit a single narrow distribution of transfer efficiencies with a clear shift of the mean toward values lower than 0.2, representing very extended states of the protein (Fig. 7B, E, and F and SI Appendix, Fig. S14). The complete disappearance of the populations observed for lipid-free ApoE4 confirms that these three constructs are fully associated with lipids. Interestingly, the construct ApoE4Δ8,241(four-helix bundle) exhibits two coexisting populations in equilibrium, one at high transfer efficiency (0.894 ± 0.004) and one at low transfer efficiency (0.037 ± 0.006) (Fig. 7C and SI Appendix, Table S4). Neither transfer efficiency is compatible with the population measured in aqueous conditions in the absence of lipids. This suggests that the four-helix bundle can undergo unpacking and restructuring when associated with lipids and that a certain degree of heterogeneity, represented by these two distributions of transfer efficiencies, is conserved even in the lipid-bound state (Fig. 7D).
Discussion

Conformational Heterogeneity in Lipid-Free ApoE4. Our single-molecule experiments and MD simulations clearly reveal that ApoE4 does not adopt a single structure but, instead, explore a complex and dynamic conformational ensemble. Using the ApoE3-like structure as a reference (22), we observe large deviations in the conformations of the hinge and C-terminal domains of the protein and dynamic fluctuations in the four-helix bundle (Fig. 4C). Interestingly, we do not find evidence in the experiments and simulations of previously proposed contacts between residues 76 and 241 (24) or residues 61 and 255 (15, 30) (Fig. 5 and SI Appendix, Figs. S19 and S20), and our experimental and computational data agree with the orientation of the N- and C-terminal domains observed in the ApoE3-like structure (22). This discrepancy with previous data can be rationalized by noting that experiments that identified these close contacts (24) were performed under conditions where the protein exists as a dimer or tetramer and therefore may be specific only to these forms of the protein. Similarly, salt bridges (15, 30) have been tested via mutational analysis in the context of lipoproteins or nonmonomeric forms of the protein and may reflect other interactions at play in those specific forms, which either do not occur or rarely occur in the monomeric case. While capturing a similar orientation of the domains, our data are at variance also with the “closed” NMR structure (22) (SI Appendix, Figs. S10 and S19). This observation supports that mutations along the sequence, as the ones used to monomerize ApoE3 in the NMR experiments, may alter the delicate balance between specific conformers in the structural ensemble. Indeed, the simulations suggest that the hinge region competes with the C-terminal domain for interactions with the four-helix bundle, where specific contacts involving the N-terminal tail and the four-helix bundle can sway the preference of interaction for one region or the other (Fig. 6). Study of mutations in positions 112 and 158 reveals that long-range conformations in ApoE3* resemble the one observed in ApoE4, whereas ApoE2* adopts slightly more compact...
configurations. This result differs from what may be expected based on the functional differences previously described for each isoform. However, our experiments only probed one long-range distance within the protein; therefore, we cannot exclude that local regions of the protein or even other long-range distances are not affected by the same mutations. In addition, although the cysteine-to-serine mutations were previously shown to not influence the function of ApoE isoforms (33–36), these amino acid substitutions may also introduce local and global conformational changes. Finally, the comparison between our work and previous observations points to a key role of oligomerization in modulating the protein conformational ensemble. These three aspects will be investigated in future works.

**Folding Equilibrium of Lipid-Free ApoE4.** Our single-molecule experiments also enable a direct quantification of the stability associated with each conformer of the monomeric protein and provide insights on the overall folding reaction. The denaturant titration suggests that structuring of the N-terminal domain proceeds from a completely unfolded state through an intermediate state where helices H1 to H4 are partially formed, followed by the subsequent packing and stabilization of the bundle (Fig. 3C). Observation of an intermediate configuration in the four-helix bundle confirms previous interpretation of ensemble data where an intermediate state was presumed (26). Contextually to the folding of the four-helix bundle, a perturbation occurs in the configurations of the hinge and in the N- and C-terminal tails. While folding of these domains remains largely independent, our data suggest that their structural organization is not disconnected. Indeed, even for labeling positions that do not sample the four-helix bundle, we identify transitions with a midpoint at approximately 2 M GdmCl accompanied by a similar change in free energy (from 5 to 7 RT, Fig. 3C and SI Appendix, Figs. S3 and S4). While folding of the C-terminal region is only captured by broadening of the distribution of transfer efficiencies (SI Appendix, Fig. S5), the observation of distinct populations in the N- and C-terminal tails and hinge region provides quantification of the energy difference between these distinct states. The similarity in the relative populations between the hinge and C-terminal regions (as measured by ApoE4<sub>182,241</sub> and ApoE4<sub>223,291</sub>) across different denaturant concentrations and the overlap between the sequence of the two regions suggest we are monitoring the same configurational change. Therefore, the emerging picture is of a folded four-helix bundle in equilibrium with at least three distinct populations of the C-terminal domain: closed, open, and extended. These three distinct configurations of the C-terminal domain are

---

**Fig. 7.** Single-molecule fluorescence experiments on lipid-bound ApoE4. (A) Distribution of radii for lipid-free ApoE4 (gray, df-FCS), lipid-bound ApoE4 (orange, FCS), and extruded DMPC liposomes (blue, cryo-TEM). (B–F) Comparison of transfer efficiency histograms for lipid-free and lipid-bound ApoE4 constructs. All histograms report on fluorescent species with a labeling stoichiometry ratio of 1D:1A. Lines are visual guides between the lipid-free (gray) and lipid-bound (orange) mean transfer efficiencies. (G) and (H) Representatives examples of conformations of ApoE4 in the expanded and compact lipid-bound states based on an ultracoarse-grained model that satisfies the mean transfer efficiency constraints.
characterized by internal dynamics on the hundreds of nanosecond timescale and are in a slow exchange, one with each other, on a timescale longer than milliseconds (SI Appendix, Fig. S21).

Monomeric ApoE4 Forms Heterogeneous Complexes with Lipids. Early EPR studies of ApoE4 suggested that helices in the N- and C-terminal domains remain in close contact in the lipid-bound state, whereas the four-helix bundle undergoes structural rearrangements (28). A competing model proposed that lipid binding favors a separation between the N- and C-terminal halves of the protein based on the ApoE3-like NMR structure (19, 22). Interestingly, our data indicate that such an open configuration is a constitutive state explored by the ApoE4 monomer and, therefore, does not require interaction with the lipids to occur. The open and extended configurations expose the required surface of the C-terminal domain making interaction with lipids possible (SI Appendix, Fig. S11). Indeed, the region between positions 165 and 270 has been identified as containing Class A amphipathic helices, which can promote lipid binding (39). Therefore, modulation of the abundance of the open state may impact the affinity of ApoE variants for lipids. Our measurements further indicate that monomeric ApoE can extract lipids and form smaller particles compared to the initial liposome preparation. This observation is compatible with previous measurements monitoring decrease of turbidity in liposome solutions upon addition of ApoE (40–42). The ability to extract lipids implies an intercalation of the amphipathic helices of the protein within the lipid bilayer. Indeed, amphipathic helices are known to play a key role in nonenzymatic membrane fusion (43), where the membrane fusion can be self-propelled by insertion of a first helix that favors insertion of subsequent helices (44–46). This same mechanism may be at play in the interaction of ApoE with liposomes, where insertion of the C terminus can then propagate through the hinge to the N terminus (22). This model explains how the hinge region, which locks the N-terminal domain in the four-helix bundle structure, can be displaced, leading to a rearrangement of the helices of the bundle and allowing for more expanded configurations. Our experiments indicate that the N-terminal domain adopts at least two different configurations, one where the helices H3 and H4 are in close proximity to one another and one in which the four helices are spread apart on the lipid particle (Fig. 7 G and H). This interpretation is fully compatible with the configurations identified by Henry et al. (29) using cross-linking, mass spectrometry, and simulations of ApoE4, although our data suggest a more expanded configuration of the N-terminal tail (as measured by ApoE4 341) and a larger separation between the N- and C-terminal halves of the protein (as measured by ApoE4 468). Interestingly, previous simulations of ApoE3 identify only a close configuration for helices H3 and H4, possibly suggesting a different structural organization of the two variants in their monomeric lipid-bound form (47). Future work will address the local organization of each ApoE region to test whether different isoforms adopt unique configurations in the lipid-bound state.

Conclusions

The realization that ApoE isoforms do not adopt one single stable structure but an intricate conformational ensemble opens the door to new explanations for the mechanism of function of the protein and its role in the context of AD. Our results demonstrate the potential of single-molecule approaches for investigating the relationship between structural ensemble and function of monomeric ApoE. This approach bypasses experimental complications due to protein oligomerization, setting the stage for exploring the impact of sequence variations and interaction with AD factors. Understanding how and why sequence mutations and environmental factors tune ApoE from being a risk factor to having neutral effects is key to identifying appropriate therapeutic strategies that can slow down or even arrest the progression of AD.

Materials and Methods

Protein Expression, Purification, and Labeling. All ApoE4 constructs were expressed in BL21-Gold (DE3) cells (Agilent). The thioredoxin–His6–ApoE protein fusion was purified using a HiTrap FF column (Cytiva). The tag was cleaved by Human rhinovirus (HRV) 3C protease and separated from ApoE4 using a heparin Sepharose FF column (Cytiva). Anion exchange chromatography (Q Sepharose HP FF column, Cytiva) was then used as the final polishing step. Correct mass of the constructs was analyzed using sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE) and/or electrospray ionization mass spectrometry. All constructs have been labeled with Alexa 488 and Alexa 594, which serve as donor and acceptor, respectively. For further details, see SI Appendix.

Single-Molecule Measurements. All single-molecule fluorescence measurements were performed on a Picoquant MT200 instrument (Picoquant). Single-molecule FRET and fluorescence correlation spectroscopy (FCS) were performed with labeled protein concentrations of 100 μM estimated from dilutions of samples with known concentration based on absorbance measurements. All single-molecule measurements were performed in 50 mM NaP, pH 7.4, 200 mM β-mercaptoethanol (for photoprotection), 0.001% Tween 20 (for surface passivation), and GdmCl at the reported concentrations, at a room temperature of 295 ± 0.5 K. Pulsed interleave excitation was used to ensure that each burst represents the transfer efficiency determined from a 1:1 donor-acceptor stoichiometry. Importantly, attachment of the probes across different labeling positions has a small impact on the overall protein conformations as measured by dual-focus FCS, which reveals variations across the different constructs of less than 10%. All data were analyzed using the Mathematica package ”Fretica” (https://schuler.bioc.uzh.ch/wp-content/uploads/2020/09/Fretica20200915.zip) developed by Daniel Nettels and Ben Schuler. Fluorescence lifetimes (SI Appendix, Fig. S22) are analyzed using a convolution with the instrument response function (SI Appendix, Fig. S23). Comparing transfer efficiency estimates from donor lifetimes (reporting about the nanosecond timescale) and from bursts of photons (reporting on the millisecond timescale) enables distinguishing whether the associated population represents a rigid configuration or a dynamic ensemble. In the case of a rigid configuration, the same transfer efficiency is recovered on both timescales and results in a constant value that follows the linear dependence of the lifetime on the mean transfer efficiency. In the case of a dynamic ensemble, a deviation from the linear dependence occurs, which depends on the sampled conformational distribution (31). Burst variance analysis (48) and nanosecond FCS (49) further provide information on interdyne dynamics (SI Appendix, Fig. S21). For further details, see SI Appendix.

MD Simulations. The NMR structure of ApoE3 (Protein Data Bank (PDB) ID: 2L7B) was used as a starting point for our simulations, with mutations performed in PyMOL to achieve the structures of ApoE4. We performed 20 rounds of directed sampling harnessing the FAST algorithm (50) to explore the conformational space of ApoE4 using the residue pairs: R92 and S263, G182 and A241, and S223 and A291, as a directed metric. The resulting simulations were clustered with similar simulations of ApoE2, ApoE3, and ApoE3ChristChurch (R136S) to a shared state (SI Appendix, Fig. S24). For further details, see SI Appendix.
Author affiliations: “Department of Biochemistry and Molecular Biophysics, Washington University in St. Louis, St. Louis, MO 63110; and “Center for Science and Engineering of Living Systems, Washington University in St. Louis, St. Louis, MO 63130
Supporting Information for
Apolipoprotein E4 has extensive conformational heterogeneity in lipid free and bound forms

Melissa D. Stuchell-Brereton\textsuperscript{1,2}, Maxwell I. Zimmerman\textsuperscript{1,2}, Justin J. Miller\textsuperscript{1,2}, Upasana L. Mallimadugula\textsuperscript{1,2}, J. Jeremias Incicco\textsuperscript{1,2}, Debjit Roy\textsuperscript{1,2}, Louis G. Smith\textsuperscript{1,2}, Jasmine Cubuk\textsuperscript{1,2}, Berevan Baban\textsuperscript{1}, Gregory T. DeKoster\textsuperscript{1}, Carl Frieden\textsuperscript{1}, Gregory R. Bowman\textsuperscript{1,2}, Andrea Soranno\textsuperscript{1,2,*}

\textsuperscript{1}Department of Biochemistry and Molecular Biophysics, Washington University in St Louis, 660 St Euclid Ave, 63110, Saint Louis, MO, USA
\textsuperscript{2}Center for Science and Engineering of Living Systems, Washington University in St Louis, 1 Brookings Drive, 63130, Saint Louis, MO, USA

*corresponding author
Andrea Soranno
Email: soranno@wustl.edu

This PDF file includes:
- Supporting text
- Figures S1 to S24
- Tables S1 to S15
- SI References
Supplementary information.

Extended authorship description.
M.D.S.B. and A.S. conceived the project and designed the experiments. M.D.S.B. expressed, purified, labeled all the constructs, and performed all single-molecule FRET and FCS measurements. M.D.S.B. and A.S. analyzed all single-molecule measurements. M.Z. performed FAST simulations and developed the FRET histogram algorithm. J.J.M. performed MD simulations on Folding@home. J.J.M., U.M., L.G.S., and M.Z. analyzed the MD simulation data and M.D.S.B. and A.S. contributed to the interpretation. A.S. performed and analyzed ultra-coarse grained MonteCarlo simulations of lipid-bound ApoE. J.J.I and D.R. performed lifetime and anisotropy calibration measurements. D.R. prepared liposomes and performed and analyzed TEM experiments. J.J.I. performed and analyzed 2d-FCS experiments. J.C. performed nanosecond-FCS experiments and contributed reagents. B.B. and G.T.D. contributed reagents. C.F. supervised the experiments. M.D.S.B., G.B., J.J.M., and A.S. wrote the manuscript.

Methods.

Protein expression, purification, and labeling.
Plasmid construct design.
Apolipoprotein E protein (NCBI Reference Sequence: NP_000032.1 – for e3 isoform) is expressed from a modified pET32a expression plasmid, which includes an N-terminal thioredoxin fusion protein, His6 tag, and HRV 3C protease site followed by the Apolipoprotein E gene (excluding the first 18 residues, or signal peptide) as previously described51. The ApoE e3 isoform gene was initially cloned into the BamHI EcoRI sites in the MCS of pET32a vector; however, the BamHI site was destroyed when the HRV 3C protease site was incorporated using site directed mutagenesis.

ATGAGCGGATAAAAATTATTCACCTGACTGACGACAGTTTTGACACGGATGTACTCAAAAGCGGACGG
GGCGATCTCTGTCGATTTCTGGGCAGAGTGGTGCGGTCCGTGCAAAATGATCGCCCCGATTCTGG
ATGAAATCGCTGACGAATATCAGGGCAAACTGACCGTTGCAAAACTGAACATCGATCAAAACCCT
GGCACTGCGCCGAAATATGGCATCCGTGGTATCCCGACTCTGCTGCTGTTCAAAAACGGTG
AAGTGGCGGCAACCAAAGTGGGTGCACTGTCTAAAGGTCAGTTGAAAGAGTTCCTCGACGCTAACCTGG
CCGGTTCTCTGTTCTGGCCCATATGCACCCATCATCATCATCATTTCTCTGTCCTGCTGTCCACGCTAACCCTG
TCTGGTATGAAAGAAACCCTGCTGCTGCTAATCAGCGACATGGACAGCCCGAGTCTCGGG
Site-directed mutagenesis was performed on the pET32a Apolipoprotein E expression vector to create the ApoE (ε4 isoform) constructs (Table S1). All mutations were verified using Sanger sequencing.

Protein expression and purification.

All ApoE4 constructs were expressed recombinantly in BL21-Gold (DE3) cells (Agilent). 2L cultures were grown in LB medium containing carbenicillin (100 µg/mL) to OD₆₀₀ ~0.8 and induced with 1 mM IPTG for 4 hours at 37 degrees C. Harvested cells were lysed with sonication at 4 degrees C in lysis buffer (50 mM Tris pH 8.0, 500 mM NaCl, 10 mg/mL lysozyme, 5 mM BME, and cOmplete™ EDTA-free Protease Inhibitor Cocktail (Roche)).
The supernatant was cleared by centrifugation (37000 rpm for 1 hour) and loaded onto a HisTrap FF column (Cytiva) in buffer A (20 mM Sodium Phosphate pH 7.5, 500 mM NaCl, 20 mM Imidazole, 5 mM BME). The Thioredoxin-His<sub>6</sub>-ApoE protein fusion was eluted with Buffer B (buffer A + 500 mM Imidazole) and dialyzed into HRV 3C protease cleavage buffer (50 mM HEPES pH 7.5, 100 mM NaCl, 5 mM BME) after adding HRV 3C protease to pooled fractions. Thus, the Thioredoxin-His<sub>6</sub> fusion protein was cleaved yielding full length ApoE4 with two additional N-term residues (GlyPro). FL ApoE4 was then bound to a Heparin Sepharose FF column (Cytiva) and eluted using a step gradient from 0 to 100% buffer B (buffer A: 50 mM HEPES pH 7.5, 100 mM NaCl, 5 mM BME, buffer B: buffer A + 1 M NaCl) over 140 min. Heparin elution fractions of FL ApoE4 were pooled and dialyzed overnight into buffer A (20 mM Sodium Phosphate pH 7.3, 2 M Urea, 5 mM BME) for subsequent ion exchange chromatography. Dialyzed FL ApoE4 was then bound to a Q Sepharose FF column (Cytiva) and eluted using a gradient of 0-40% buffer B (buffer A: (20 mM Sodium Phosphate pH 7.3, 2 M Urea, 5 mM BME, buffer B: buffer A + 1 M NaCl) over 80 min, then 40-100% buffer B over 60 min. Purified FL ApoE4 variants were analyzed using SDS-PAGE and verified by electrospray ionization mass spectrometry (LC-MS). Concentrations were determined spectroscopically in 50 mM Sodium Phosphate pH 7.5, 2 M Urea, 200 mM NaCl, 5 mM BME using an extinction coefficient = 44950 M<sup>-1</sup> cm<sup>-1</sup>.

**Choice of labeling positions.**

The choice of labeling positions has been designed as a compromise between flanking regions of interest (hinge, C-terminal domain, four-helix bundle) and a series of different criteria regarding the structural properties of the protein. As there are no full-length structures available for either monomeric ApoE4 or ApoE4 oligomers, the ApoE3-like NMR structure (PDB: 2L7B) was used as a guide for identifying structural similarities with the resolved structure. In particular, we avoided altering amino acids clearly involved in structurally relevant interactions (such as aromatic residues) or previously identified functional mutations of the protein. When choosing labeling positions within folded domains, we focused on surface exposed residues to maximize accessibility of the cysteine residues during labeling. The spacing of the fluorophores was optimized to ensure use of the entire FRET dynamic range when the protein is unfolded, so that a contrast between the distance in the folded structure and in a disordered state would provide different transfer efficiencies. We also attempted to reduce the effects of
quenching due to proximity between fluorophores and aromatic residues\textsuperscript{52,53}. With tryptophan residues identified as major quenchers for Alexa Fluor 488 and Alexa Fluor 594, fluorophores were positioned at least 10 residues away from its nearest tryptophan in the sequence.

**Protein labeling.**
All ApoE variants were labeled with Alexa Fluor 488 maleimide (Invitrogen Molecular Probes) under denaturing conditions in buffer A (20 mM Sodium Phosphate pH 7.3, 2 M Urea, 5 mM BME) at a dye/protein molar ratio of 0.7/1.0 at 4 degrees Celsius overnight. Single labeled ApoE protein was isolated using ion-exchange chromatography (Mono Q 5/50 GL, Cytiva – protein bound in buffer A and eluted with 0-40% buffer B (buffer A + 1 M NaCl) gradient over 80 min, then 40-100% buffer B over 20 min). UV-Vis spectroscopic analysis was used to identify fractions with 1:1 dye:protein labeling. Single labeled Alexa Fluor488 maleimide labeled N protein was then subsequently labeled with Alexa Fluor 594 maleimide at a dye/protein molar ratio of 1.2/1.0 overnight at 4 degrees Celsius. Double labeled (488:594) protein was then further purified using ion-exchange chromatography (Mono Q 5/50 GL, Cytiva – as listed above).

**Liposome preparation.**
Unilamellar vesicles were prepared from 1,2-dimyristoyl-sn-glycero-3-phosphocholine (DMPC) (Avanti Polar Lipids, USA) dissolved in Chloroform (25 mg/mL). The DMPC was used as received without further purification. DMPC solution containing 12.5 mg lipid was dried under a stream of dry nitrogen in glass scintillation vials. This lipid thin film was kept under vacuum overnight to remove traces of organic solvent. Liposomes were prepared by hydrating the film with 1 ml of 50 mM sodium phosphate buffer (pH 7.4) at 42 ± 1°C. The vial was shaken to form a rough liposomal suspension followed by five freeze-thaw cycles to obtain homogeneous preparation. Freezing was achieved by submersion into liquid nitrogen followed by thawing the sample in a water bath at 42 ± 1°C. The prepared vesicles were then extruded through an extruder set (Avanti Polar Lipids, USA) containing polycarbonate membranes with a pore size of 0.2 μm (Whatman plc., Cytiva Life Sciences). Finally, the extruded vesicles were flash frozen using liquid nitrogen and stored at -80°C for future use. Before the experiment, the frozen liposomes were thawed in a
water bath at 42 ± 1°C followed by five freeze/thaw cycles, according to the protocol discussed earlier. The liposomes were then re-extruded using Avanti Polar Lipids extruder set at room temperature. The extruded liposomes were kept at 4°C and used within 2 days after the preparation. In parallel to single-molecule experiments, 50 μL of the prepared liposome solutions were used for cryo-TEM imaging using JEOL JEM-1400(Plus) (JEOL USA Inc., USA) 120 kV Transmission Electron Microscope. The images were captured using an AMT XR111 CCD camera and analyzed using a Graphical User Interface developed in Mathematica (Wolfram research).

Data analysis.

Experimental setup and procedure for single-molecule fluorescence experiments. Single-molecule fluorescence measurements were performed on a Picoquant MT200 instrument (Picoquant, Germany). For single-molecule FRET measurements, a Pulsed Interleaved Excitation (PIE) scheme was obtained by synchronizing 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) and pulsed at 20 MHz. Emitted photons were collected through a 60x1.2 UPlanSApo Superapochromat water immersion objective (Olympus, Japan). For nanosecond 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 through a 60x1.2 UPlanSApo Superapochromat water immersion objective (Olympus, Japan), passed through a dichroich mirror (ZT568rpc, Chroma, USA), a long-pass filter (HQ500LP, Chroma Technology) to suppress scattering light, and a confocal pinhole (100 µm 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 PIE 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 μW). Single-molecule FRET efficiency histograms were acquired from samples
with protein concentrations between 50 pM and 100 pM, estimated from dilutions of samples with known concentration based on absorbance measurements. All measurements were performed in 50 mM NaPi pH 7.4, 143 mM β-mercaptoethanol (for photoprotection), 0.001% Tween 20 (for surface passivation) and GdmCl or TMAO at the reported concentrations. Tween 20 at this concentration is below the critical micelle concentration and we observe no evidence of interaction of Tween with the protein, neither interference with its ability of binding lipids. Measurements under aqueous buffer conditions are performed with PolyEthyleneGlycol (PEG)-passivated cuvettes\textsuperscript{54}, which significantly prevents adhesion of the protein to the surface. The exact concentration of denaturant 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).

**FRET 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 15 photons in each bin and 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 and a minimum common threshold across constructs has been used to minimize contributions to the width of transfer efficiency distributions due to different thresholds\textsuperscript{30}. Transfer efficiencies for each burst were calculated according to

\[
E = \frac{n_A}{n_A + n_D}
\]

Eq. S1

where \(n_D\) and \(n_A\) 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\textsuperscript{55}. The labeling stoichiometry ratio \(S\) was computed accordingly to:

\[
S = \frac{I_D}{(y_{PIE}I_A + I_D)}
\]

Eq. S2
where $I_D$ and $I_A$ represent the total intensities observed after donor and acceptor excitation and $\gamma_{PIE}$ 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 according to the criterion $0.3 < S < 0.7$. The final histogram of transfer efficiencies was constructed from these described selected bursts. 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 extract multiple populations from the transfer efficiency histograms, each population was approximated with either a Gaussian or a LogNormal distribution function. For fitting more than one peak, the histogram was analyzed with a sum of the abovementioned 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. A list of used parameters is provided in the following.

For ApoE$_{182,241}$, data were globally fitted with the same width for each population across all concentrations of GdmCl. While this does not account for small deviations in the width due to the transfer efficiency dependence of the shot noise$^{30}$, this option was made necessary because of the strong overlap between the two main populations across a large range of denaturant concentrations. The width was constrained between 0 and 0.09 based on the main width determined in high denaturant concentrations, where the region is completely unfolded. The upper limit value aligns with analogous values observed for other constructs.

For ApoE$_{223,291}$, we performed a global fit across all GdmCl concentrations with the following constraints: the mean transfer efficiency of the population at low transfer efficiency was shared across multiple GdmCl concentrations and constrained within a range of $0.06<E<0.18$; the mean transfer efficiency of the population at high transfer efficiency was shared across multiple GdmCl concentrations and constrained within a range of $E>0.85$; the width of the distribution associated with each population was constrained for the population at intermediate transfer efficiencies between 0.06 and 0.13, whereas the one for the population at high transfer efficiency was constrained between 0 and 0.1.
For ApoE4<sub>86,241</sub>, no global fit was necessary but the mean transfer efficiencies of the three populations were constrained such that 0.1<E<0.25, 0.3<E<0.75, and 0.8<E<0.9 for the low, intermediate, and high transfer efficiency population, respectively. All the widths were constrained between 0 and 0.12.

For ApoE4<sub>86,165</sub>, we initially fit the data using two Gaussian distributions and a Log-Normal distribution. After the initial fit suggests an almost constant value for the folded and intermediate population, the model has been further restrained to account for these fixed positions (though corrected for the refractive index change due to the increasing denaturant. Given the small amplitude, the width of the intermediate distribution has been enforced to be 0.1 across all denaturant concentration. Similarly, the width of the unfolded state has been constrained to 0.1 up to 2 M GdmCl, where the width of the distribution can start to be determined more robustly.

For ApoE4<sub>5,86</sub>, we use a single Gaussian distribution without any constraint for each denaturant concentration above 1 M GdmCl. From 1 M GdmCl, we use a combination of a Log-Normal and a Gaussian distribution. Based on an iterative fitting procedure, further constraints have been enforced, requiring that E<sub>1</sub> < 0.15 and fixing the asymmetry of the Log-Normal distribution to 1.875. This is made necessary to limit the error in the estimate of the area of the low transfer efficiency population.

Folding equilibrium.
Folding equilibrium across the identified populations is described in terms of two-state and three-state models depending on the observable of the specific construct. Here, we exemplify the two-state model for the case of an unfolded (U) and folded (F) states and the three-state model as the equilibrium between native (N), intermediate (I) and unfolded (U) states. The same models can be applied to describe the folding equilibrium between different native states, as in the case of ApoE4<sub>223,281</sub>.

For the two-state equilibrium N ⇌ U, the corresponding fraction folded and unfolded can be written in terms of the equilibrium constant $K_{UN}$ as

$$f_U = \frac{1}{1+K_{UN}}, f_N = \frac{K_{UN}}{1+K_{UN}}$$

Eq. S3
where

\[ K_{UN} = \exp[-\Delta G^U_0/RT + m_c] = \exp [m/RT(c - c_{1/2})] = \exp [\Delta G^U_0/c_{1/2}(c - c_{1/2})] \]

where \( c \) is the denaturant concentration, \( c_{1/2} \) represents the denaturant concentration at which the folding and unfolding fraction curves cross each other, e.g., when the two states have an equal abundance, \( \Delta G^U_0 \) is the free energy difference between the N and U states extrapolated at zero denaturant concentration, R is the ideal gas constant and T is the temperature.

For the three-state equilibrium \( N \rightleftharpoons I \rightleftharpoons U \), the corresponding fraction folded, intermediate, and unfolded can be written in terms of the equilibrium constant \( K_{UI} \) and \( K_{IN} \) as

\[
\begin{align*}
    f_U &= \frac{1}{1 + K_{UI} + K_{UI} K_{IN}}, \\
    f_I &= \frac{K_{UI}}{1 + K_{UI} + K_{UI} K_{IN}}, \\
    f_N &= \frac{K_{UI} K_{IN}}{1 + K_{UI} + K_{UI} K_{IN}} \quad \text{Eq. S4}
\end{align*}
\]

where \( K_{UI} = \exp [\Delta G^U_0/c_{1/2}^I(c - c_{1/2}^I)] \) and \( K_{IN} = \exp [\Delta G^I_0/c_{1/2}^U(c - c_{1/2}^U)] \) with \( \Delta G^U_0 \) and \( \Delta G^I_0 \) represent the free energy difference extrapolated to aqueous buffer conditions between the U and I and I and N states, respectively, whereas \( c_{1/2}^U \) and \( c_{1/2}^I \) are the concentration where the corresponding fraction curves cross each other. In this case, it is important to notice that the crossing point does not necessarily occur in the midpoint (50%) of the transition. Finally, when the folding transition is not reported in terms of fractions but is estimated based on a specific signal (e.g., variation in mean transfer efficiency or width of the transfer efficiency distribution), the total signal is described as combination of specific fraction with the corresponding signal of each population, where the signals associated to each population are considered further parameters.

**IRF determination.** Instrument response function (IRF) for each 4 detectors and for the 2 lasers was obtained recording the TCSPC histogram for the fluorescence of Rhodamine B (RhB, 100 µM) in presence of high concentration of potassium ferricyanide (95% of a saturated solution of K₃Fe(CN)₆ at 23°C in water), an electron-transfer fluorescence quencher.\(^{56}\)

As the rate of non-radiative relaxation pathways of the fluorophore increases with addition of quencher, the overall fluorescence time decay gets faster and eventually it is fast enough in comparison with the IRF—effectively a delta function—, the observed TCSPC histogram approaches the actual IRF.\(^{57}\) This is due to the fact the observed TCSPC
histogram is a time convolution of the IRF and the fluorescence relaxation kinetics of the fluorophore,

\[ f(t) = \int_0^t IRF(t - \tau)I(\tau) d\tau \quad \text{Eq. S5} \]

and the convolution of a function with a δ-function is equal to the function itself.

Laser power was set at 0.5 µW after the main dichroic at a repetition rate of 20 MHz and decreased 100 times before measurement (final power ~ 5 nW). Data was collected with 16 ps resolution. The collected TCSPC histograms were corrected by subtracting dark counts - which appear uniformly distributed in the histogram - and normalized over the total number of counts.

For consistency, we checked if at the ferricyanide concentration used its quenching effect on the TCSPC histogram was already maximum in the conditions of our measurements. On the one hand, we found that the histograms converged and became sharper as the concentration was increased and from 0.8 M up, they were indistinguishable. On the other hand, we also found that, as the mean fluorescence intensity decreased, the computed steady-state anisotropy converged to a value of 0.37-coincident with previously reported values of fundamental anisotropy of RhB derivatives \(^5^8\), as expected for emission taking place much faster than rotational relaxation from randomly oriented fluorophores with almost parallel excitation and emission transition dipolar moments\(^5^9\).

We found stronger quenching and faster fluorescence decay of RhB with ferricyanide than with potassium iodide, a quencher commonly used for this purpose, both tested at the concentration of a 95% saturated solution at room temperature (~23°C). Additionally, in presence of quencher, we found no detectable differences in the normalized TSCPC histograms obtained with Alexa594 upon 589 nm excitation with 640 nm detection, but slightly slower fluorescence decay with Alexa488 in comparison to RhB, upon 485 nm excitation and 525 nm detection.

**Correction factors for quantitative anisotropy determination.** In order to obtain estimates of the correction coefficients \(L_p\) and \(L_s\) accounting for scrambling of polarization due to tight focusing optics we employed the technique developed by Koshioka and coworkers\(^6^0\). The methodology consists in performing a measurement of fluorescence intensity in two solutions of a fluorophore, in low and high viscosity media, recording the
signal in 2 detectors collecting light polarized parallel (p) and perpendicular (s) to that of the excitation laser. The time correlated single photon counting (TCSPC) histogram of p- and s- polarization detection for each medium is fitted as a convolution with the Instrument Response Function (IRF) and a fluorescence lifetime function containing fluorescence and rotational relaxation kinetic terms. The parameters $L_p$ and $L_s$ are obtained through global fitting of the TCSPC histograms from the two solutions and two polarizations.

To this end, we used maleimide-C5 Alexa488 to determine $L_p$ and $L_s$ in detectors for donor dye upon excitation at 485 nm, and maleimide-C5 Alexa594 dyes to determine $L_p$ and $L_s$ in acceptor-channel detectors upon excitation at 590 nm. The concentration of dye solutions was 5 µM in water for the low viscosity solution and in 59% v/v glycerol in water (high viscosity solution), supplemented with 240 mM beta-mercaptoethanol. Measurements were performed at 23 ± 1°C. Laser power was set at 0.5 µW after the main dichroic at a repetition rate of 20 MHz and decreased 100 times before measurement (final power ~ 5 nW). Data was collected with 16 ps resolution.

The following equations for parallel (p) and perpendicular (s) detection were simultaneously fit to the corresponding pair of experimental TCSPC histograms:

$$f^d_p(t) = \Delta \tau \sum_{j=1}^{k} IRF^d_{p, \lambda}(\tau_j - i) L_p(\tau_j)$$

$$f^d_s(t) = \Delta \tau \sum_{j=1}^{k} IRF^d_{s, \lambda}(\tau_j - i) L_s(\tau_j)$$

Eq. S6

where $d$ denotes the detector, $\lambda$ denotes the excitation laser employed, $IRF^d_{p, \lambda}$ is the corresponding experimentally determined instrument response function, and $I_p$ and $I_s$ are given by,

$$I_p(\tau) = a_p e^{-\frac{\tau}{\tau_F}} \left[ 1 + r_0(2-3L_p) e^{-\frac{\tau}{\tau_R}} \right]$$

$$I_s(\tau) = a_s e^{-\frac{\tau}{\tau_F}} \left[ 1 - r_0(1-3L_s) e^{-\frac{\tau}{\tau_R}} \right]$$

Eq. S7

where $a_p$ and $a_s$ are amplitude coefficients, $\tau_F$ is the fluorescence emission lifetime parameter, $\tau_R$ is a rotational diffusion characteristic time parameter, $r_0$ is the fundamental anisotropy parameter, and $L_p$ and $L_s$ are the corresponding scrambling correction coefficients for each polarization. Parameters $a_p$, $a_s$, $\tau_F$, $\tau_R$, $L_p$ and $L_s$ were floating fitting
parameters while \( r_0 \) was fix at the previously reported value for Alexa488 and Alexa594, \( \sim 0.38 \) \( ^{61} \). Simultaneous fitting of Eq. S7 to the pairs of TCSPC histograms obtained in water and in 59% v/v glycerol was performed sharing only parameters \( L_p \) and \( L_s \), employing weighted least squares, using \( 1/f(t) \) as weights.

Computing the time resolved anisotropy, \( r(t) \),

\[
    r(t) = \frac{F_p(t) - G F_s(t)}{\left(1 - 3L_s\right)F_p(t) + \left(2 - 3L_p\right)G F_s(t)} \tag{Eq. S8}
\]

also requires estimation of the relative efficiencies of detection in both detectors, parallel and perpendicular, denoted by the \( G \) factor, i.e., the efficiency of detection compared between parallel and perpendicular polarization.

To this end, for each pair of detectors, we used 3 different approaches which show consistency with each other:
- the ratio of amplitude parameters \( a_p \) and \( a_s \) in Eq. S7;
- solving Eq. S8 for \( G \) using independent measurements of steady-state anisotropy together with measurements in our microscope of time-average fluorescence intensity \( <F_p> \) and \( <F_s> \) and parameters \( L_p \) and \( L_s \) obtained as described above.
- fitting the value of \( G \) such that the time trace of \( r(t) \) in Eq. S8 obtained for free Alexa dyes in water at room temperature, decreases asymptotically to 0 and effectively reaches that value in less than 10 ns.

The values obtained of \( L_p, L_s \) and \( G \) for each excitation laser and each pair of detectors and employed in the analysis of the results in this work are shown in Table S5.

**Fluorescence lifetimes analysis.** We estimated the fluorescence lifetime of the donor and acceptor by globally fitting the histogram of photon arrival times with Eq. S7, where \( a_p \) and \( a_s \) are treated as free parameters, \( r_0 \) is fixed to 0.38 as an average value based on previous estimates for Alexa Fluor 488 and Alexa Fluor 594\(^ {61} \) and \( \tau_L \) and \( \tau_r \) are the lifetime of the fluorophore and the rotational component of the dye, respectively. To this end, we time-gate the fluorescence arrival time to separate photons derived from donor and acceptor excitation and we select for stoichiometry and transfer efficiency to restrict the analysis to each subpopulation. We found that addition of an additional fitting parameter
representing background counts is particularly useful when analyzing high transfer efficiency populations, such as for ApoE86,165. Obtained values are reported in Table S6 and S7 for lipid-free and lipid-bound constructs. The lifetime of donor in presence or absence of acceptor is used to further compute the FRET rate. This provides an overall measurement of the characteristic timescales at play, enabling to compare whether the energy transfer occurs on timescale longer than the rotation of the fluorophores, which is commonly assumed to interpret transfer efficiency according to:

$$E = \int_{b_0}^{l_c} E(r)P(r)dr$$ \hspace{1cm} \text{Eq. S9}$$

Where $b_0$ is the shortest contact distance and $l_c$ is the contour length of the chain assuming all amino acids are in part of a disordered polymer (i.e., equal to the number of amino acids times the Cα-Cα distance)62.

**Anisotropy analysis.** Anisotropy of the donor-only, acceptor-only, and acceptor after donor excitations are computed based on Eq. S8 from photon arrival times histograms detected on the $p$- and $s$-polarization. Time-resolved anisotropy decays are then fitted to a double exponential decay according to:

$$\left( (r_0 - r_\infty) e^{-t/\tau_r} + r_\infty \right) e^{-t/\tau_M}$$ \hspace{1cm} \text{Eq. S10}$$

where $r_0$ is fixed to 0.38 as described above, $r_\infty$ is the residual anisotropy at long times (as indicated by the limiting time between two excitations pulses), $\tau_r$ represents the rotational component of the dye, and $\tau_M$ reports about tumbling of the overall molecule. Paralleling the analysis of lifetime fluorescence, values for $r_\infty$ and $\tau_r$ are reported in Tables S8 and S9. $\tau_r$ can be compared to the analogous quantity obtained from lifetime fits and is always faster than the corresponding inverse of the FRET rate or donor lifetime. $r_\infty$ are used to compute the orientation factor $\kappa^2$. Finally, we estimated also steady-state anisotropies by computing anisotropy based on the burst-determined photons in the different polarizations. The final calculation is equivalent to the one presented in Eq. S8, where the time-dependent factors are replaced by the number of photons in each selected burst.
**Orientation Factor $\kappa^2$.** While commonly assumed to be equal to $2/3$, the orientation factor $\kappa^2$ can be estimated using residual fluorescence anisotropies, which provide boundaries on the angles sampled by fluorophores. Though this is subject to various approximations, it does provide a useful test to quantify whether the accessible space sampled by the dyes is hindered by a folded domain. Following previous treatments where tumbling of the dye is described as a wobble-in-a-cone\(^{61,63}\), $\kappa^2$ can be estimate as:

\[
\kappa^2 = \left(1 - \frac{r_A}{r_0}\right)\left(\frac{r_D}{r_0}\left(\cos \frac{\theta_+ - \theta_-}{2}\right)^2 + \frac{1}{3}\right) + \\
\left(1 - \frac{r_D}{r_0}\right)\left(\frac{r_A}{r_0}\left(\cos \frac{\theta_+ + \theta_-}{2}\right)^2 + \frac{1}{3}\right) + \frac{1}{r_0} \sqrt{r_A r_D} \left(\frac{3}{2} \cos \theta_+ + \cos \theta_- - \cos \beta\right)
\]

Eq. S11

where $\cos \beta = \frac{2 r_{A(D)}}{\sqrt{3} r_{A(D)}} + \frac{1}{3}$. Based on previous determinations, $r_0$ is set to 0.38. The only two unknowns are the angles $\theta_+$ and $\theta_-$, whose boundaries are estimated as:

\[
\cos^{-1}(\min(\cos \beta , 1)) < \theta_+ < 2\pi - \cos^{-1}(\min(\cos \beta , 1)) \\
-\cos^{-1}(\min(\cos \beta , 1)) < \theta_- < \cos^{-1}(\min(\cos \beta , 1))
\]

Eq. S12a

Eq. S12b

with $\min$ indicating the smaller among the two values. Evaluating the solution across all the possible angles in these two intervals enables reconstructing a distribution of the possible $\kappa^2$ values, which deviates from the expected distribution for freely rotating fluorophores. Given that the tumbling of fluorophores occurs on a time scale faster the fluorophore lifetime and that of the inverse of the rate of energy transfer (see Table S6-S9), we can compute a mean value of $\kappa^2$ from the distribution, which can be compared to the expected value of $2/3$. All the measured mean values are within the range of 0.74-1.01, with a bias toward higher values (stronger hindrance) for the lipid-bound states. It is important to note that $\kappa^2$ enters in the Forster radius at the power of $1/6$ and therefore such deviations from $2/3$ have an impact of about 10% on the estimate value of $R_0$ (Table S10-S11).
From the distributions of $\kappa^2$, we can further evaluate the precision and accuracy associated to the estimate of mean $\kappa^2$, as well as the minimum and maximum of the distribution:

$$\text{Precision} = \int_{\kappa^2_{\text{min}}}^{\kappa^2_{\text{max}}} d\kappa^2 \left( 1 - \left( \frac{3}{2} \kappa^2 \right)^{-1/6} \right) P(\kappa^2) / \int_{\kappa^2_{\text{min}}}^{\kappa^2_{\text{max}}} d\kappa^2 P(\kappa^2)$$

Eq. S13

$$\text{Accuracy} = \text{var} \left( \left( \frac{3}{2} \kappa^2 \right)^{-1/6} \right)$$

Eq. S14

$$\kappa^2_{\text{min}} = \frac{2}{3} \left( 1 - \left( \frac{r_D}{r_0} + \frac{r_A}{r_0} \right) / 2 \right)$$

Eq. S15

$$\kappa^2_{\text{max}} = \frac{2}{3} \left( 1 + \sqrt{\frac{r_D}{r_0}} + \sqrt{\frac{r_A}{r_0}} + 3 \sqrt{\frac{r_D}{r_0}} \sqrt{\frac{r_A}{r_0}} \right)$$

Eq. S16

The minimum and maximum of $\kappa^2$ are analytically computed assuming there is no knowledge of $r_{\infty-A(D)}$. A computational value can be obtained by evaluating all the possible angles allowed by Eq. S12$^{61}$. The terms $r_A, r_D, r_{A(D)}$ reflects the residual anisotropy of the acceptor only, donor only and acceptor after donor excitation, respectively, and can be estimated from the values of residual anisotropy associated with the dye relaxation $r_{\infty-Aonly}, r_{\infty-Donly}, r_{\infty-A(D)}$ (see previous section). As validation, we further compared the values obtained with this method with analogous estimates obtained from the corresponding steady-state anisotropies, which is often used as a first approximation for the estimate of the residual anisotropies. Importantly, steady-state anisotropies are computed as an average across population specific bursts (donor-only, acceptor-only, donor-acceptor) and therefore provide an independent estimate compared to the time resolved ones. We found a general good agreement between these two methods of estimate, suggesting that the quantification of anisotropies and $\kappa^2$ are robust.

For the conversion of transfer efficiency to distances, we then use the value of the Förster radius for Alexa Fluor 488 and Alexa Fluor 594 previously determined and reported in literature, $R_0 = 5.4$ nm$^{64}$, corrected by the variation in solution refractive index and $\kappa^2$.

**Lifetime vs FRET.** The fluorescence lifetime is related to the mean transfer efficiency$^{30}$ through:
\[ \tau_{DA}/\tau_D = 1 - \langle E \rangle + \frac{\sigma^2}{1-\langle E \rangle} \tag{Eq. S17} \]

where \( \sigma^2 \) is related to the variance of the sampled distribution via:

\[ \sigma^2 = \int_0^\infty E(r)^2 P(r)dr - \langle E \rangle^2 \tag{Eq. S18} \]

For \( \sigma^2 \) equal to zero, Eq. S17 reduces to the linear trend expected for a rigid distance. For fitting experimental data, we choose the \( P(r) \) of a wormlike chain, as described by:

\[ P_{WLC}(r, l_p, l_c) = \frac{4\pi(r/l_c)^2C(l_p,l_c)}{l_c(1-(r/l_c)^2)^{3/2}} \exp \left[ -\frac{3l_c}{4l_p(1-(r/l_c)^2)} \right] \tag{Eq. S19a} \]

\[ C(l_p, l_c) = \frac{1}{\pi^{3/2}e^{-a^2(1/3+a+15/(4a^2))}} \tag{Eq. S19b} \]

with \( a = \frac{3l_c}{4l_p} \), where \( l_p \) is the persistence length. In analyzing the lifetime vs transfer efficiency data, both \( l_p \) and \( l_c \) are varied. Whereas for a given polymer the contour length is fixed, using the contour length as a fitting parameter for a region in which some parts are folding reflects the inherent change in the nature of polymer: the polymer cannot be anymore stretched to the say contour length because of the folded regions. It is an approximation used to provide an empirical quantification of the underlying distance distribution as well as to provide a qualitative measurement that accounts for the conformational changes within the protein. The limit of a completely dynamic chain is recovered for a Gaussian distribution:

\[ P_G(r, \langle r^2 \rangle) = 4\pi r^2 \left( \frac{3}{2\pi \langle r^2 \rangle} \right)^{3/2} \exp \left[ -\frac{3}{2} \frac{r^2}{\langle r^2 \rangle} \right] \tag{Eq. S20} \]

**Nanosecond Fluorescence Correlation Spectroscopy.** Autocorrelation curves of acceptor and donor channels and cross-correlation curves between acceptor and donor channels were calculated as described previously. All samples have been measured at
a concentration between 100 pM and 1 nM and bursts with a transfer efficiency larger than the donor only state have been selected to eliminate the contribution of donor-only molecules to the correlation amplitude. Finally, the correlation was computed over a time window of 5 μs and characteristics timescales were extracted according to:

\[ g_{ij}(\tau) = 1 + 1/N(1 - c_{AB}Exp[-(\tau - \tau_0)/\tau_{AB}])(1 + c_bExp[-(\tau - \tau_0)/\tau_b])(1 + c_TExp[-(\tau - \tau_0)/\tau_T]) \]  

Eq. S21

where \( N \) is the mean number of molecules in the confocal volume and \( i \) and \( i \) indicate the type of signal (either from the Acceptor or Donor channels). This allows to estimate the timescale and amplitudes related to photon antibunching \( \tau_{AB} \) and \( c_{AB} \), chain dynamics \( \tau_b \) and \( c_b \), and triplet blinking of the dyes \( \tau_T \) and \( c_T \).

**Burst Variance analysis.** To further test for dynamics within the C-terminal region, where we observe a broadening of the distribution of distances for ApoE4_{223,291}, we performed burst variance analysis\(^{48}\). In short, we compute for define a proximity ratio \( E^* \) by substituting in Eq. 1 the uncorrected counts \( n_{\lambda}^U \) and \( n_{\lambda}^D \) from the acceptor and donor channel:

\[ E^* = \frac{n_{\lambda}^U}{n_{\lambda}^U + n_{\lambda}^D} \]  

Eq. S22a

and a standard deviation by using:

\[ \sigma E^* = \sqrt{E^*(1-E^*)/n} \]  

Eq. S22b

where \( n = n_{\lambda}^U + n_{\lambda}^D \), i.e., the total number of selected photons.

These two equations provide the theoretical expectations for the proximity ratio and its standard deviation based on a given photon-window of \( n \) consecutives photons.

We then compute the experimental expectations for each burst \((E^*_i, s_i)\) by defining:

\[ E^*_i = \frac{1}{M_i}\sum_{j=1}^{M_i} E^*_{ij} \]  

Eq. S23a

\[ s_i = \sqrt{\frac{1}{M_i}(E^*_{ij} - E^*_i)^2} \]  

Eq. S23b

where \( E^*_{ij} \) is the proximity ratio of the photon-window \( j \) in the burst \( i \), \( M_i \) is the number of photon-windows of \( n \) photons in the burst \( i \).
We use a photon-window of 5 photon and report the estimated values from each burst as well as the average based on binning all proximity ratios in bins of width 0.02. See Figure S21.

Small deviations from the theoretical line suggests some degree of dynamics on the microsecond timescale for the intermediate transfer efficiency population of \textbf{ApoE4}_{223,291}. To understand whether these dynamics propagates also to the hinge, we compared the results with the ones for construct \textbf{ApoE4}_{182,241}, finding a similar change in the main population.

**MD simulations of lipid-free ApoE4.** The NMR structure of ApoE3 (2L7B) was used as a starting model for all our simulations. Using pymol’s mutagenesis wizard, we restored the native, mature, sequence of ApoE3 and made subsequent mutations from this model to acquire structures for ApoE2 (R158C), ApoE4 (C112R), and ApoE3 Christchurch (R136S). Each model was solvated in a dodecahedron box with 2.1 nm between the protein and the edge of the box. Systems were subsequently solvated, Na$^+$ and Cl$^-$ ions were added to a final concentration of 0.1 M with no net system charge, and the system was energy minimized with a steepest descents algorithm until the maximum force fell below 100 kJ mol$^{-1}$ nm$^{-1}$, using a step size of 0.01 nm and a cut-off distance of 1.2 nm for the Coulomb interactions, van der Waals interactions, and neighbors list. The AMBER03 force field$^{66}$ and explicit TIP3P solvent$^{67}$ were used for all simulations. Systems were subsequently equilibrated for 1.0 ns, where all bonds were constrained with the LINCS algorithm$^{68}$ and virtual sites to allow for a 4.0 fs timestep. Cut-offs of 1.1 nm were used for the neighbor list with 0.9 for Coulomb and van der Waals interactions. We deployed the particle mesh Ewald method for treatment of long-range interactions with a Fourier spacing of 0.12 nm. The Verlet cut-off scheme was used for the neighbor list. The stochastic velocity rescaling (v-rescale) thermostat$^{69}$ was used to hold the temperature at 300K. All simulations were prepared using Gromacs$^{70}$ 2020.

We ran our FAST adaptive sampling algorithm$^{50}$ on each of these models to explore distances between the following residue pairs: R92-W264, P183-K242, and R224-A292. The FAST algorithm balances directed exploration and unbiased simulations to efficiently explore conformational space of proteins. The algorithm proceeds as follows: 1) run initial simulations, 2) build a Markov State Model (MSM) for the aggregate simulation time, 3) rank each observed state in the MSM based on the exploration parameter, 4) restart
simulations from the top ranked states, 5) repeat steps 2-4 until the specified number of FAST rounds is complete. For each model, we ran our FAST-sampling algorithm at 300 K for 12 rounds, followed by another 8 rounds of FAST-string (20 total) with 10 simulations per round\(^{71}\). Each simulation was 40 ns in length, for a total simulation time of 8 \(\mu\)s for each variant. For ApoE4 we performed 2 independent rounds of the above FAST sampling. All FAST simulations were performed using Gromacs.

As ApoE is highly flexible, we wanted to ensure we had robust sampling to adequately calculate the probability of each state. Using enspara’s clustering app\(^{72}\), we clustered the FAST simulations from all mutants into a shared state space model coarse-grained to a final RMSD of 3.5 Å. This clustering resulted in a total of 18,182 discrete states. As each cluster center could come from any ApoE variant, we utilized MODELLER\(^{73}\) to mutate all sequences back to ApoE4. Subsequently, we solvated each state in a dodecahedron box whose edges extend 1.0 nm beyond ApoE. The remainder of the system preparation followed that of the FAST system preparation. We next launched five independent simulations from each structure on the distributed computing platform, Folding@home. Each trajectory ran for a maximum of 100 ns, though the average trajectory length was 37 ns, for an aggregate simulation time of 3.43 ms. Simulations were performed using OpenMM\(^{74}\), using the Langevin Integrator, a step size of 2 fs, and a temperature of 300K. During our simulations, some trajectories resulted in protein unfolding and caused ApoE to interact with its periodic image. These trajectories, along with trajectories that had improper periodic boundary condition removal, were excluded from our analysis resulting in a total of 3.37 ms of aggregate simulation time analyzed.

As the C-terminal domain of ApoE is highly flexible, we were concerned that clustering on all-atom features such as RMSD would yield a poorly connected state-space. Accordingly, we clustered ApoE on the distances between 15 different residue pairs distributed throughout ApoE. We selected these pairs to both ensure that we had an overall understanding of various aspects of ApoE movement, including extension of the C-terminal domain, movement of the N-terminal domain, motions between each domain, and the five distances measured in our single-molecule FRET experiments. The total list of residue pairs is in Table S15. Following coarse-graining, we generated a Markov State Model (MSM) using enspara’s MSMBuilder\(^{72}\) with a lag time of 10 ns (Figure S13)
Post-hoc calculation of FRET histograms. FRET histograms were calculated through a kinetic Monte Carlo simulation using the constructed Markov State Model (MSM). We designed this simulation to replicate the experimental process of single molecule FRET. First, we ran nearly a microsecond of MD simulations for both cysteine-Alexa 488 and cysteine-Alexa 594 using the same parameters as our all-atom ApoE simulations. Next, we aligned the cysteine backbones for all frames and the atomic coordinates of the center of mass for the dye's photon acceptor/donor region were taken as coordinates to form a dye point cloud. Accordingly, each point in the point cloud represents a position that the dye head was located in the simulation. Then, we align the center of the dye point clouds onto each structure in the MSM at both the labeling sites for a given FRET construct, removing any dye positions that touch or overlap with the protein. Finally, we compute the distribution of distances between each potential dye position.

Next, we simulate smFRET by recoloring an experimental photon burst. To do this, we select the photon arrival times from an experimental FRET trace and simulate a synthetic trajectory from our MSM of length \( \alpha t \), where \( \alpha \) is a time rescaling factor and \( t \) is the total length of time of the experimental FRET burst. The states visited during the synthetic trajectory are governed by the transition probabilities of our MSM model. At each photon arrival time in the experimental burst, we select the corresponding state in our synthetic trajectory and randomly choose a position of the FRET dyes based on the previously calculated distribution probability for that state. To determine whether the observed photon is the result of a FRET event, or a donor emission, we convert the inter-dye distance into a FRET transfer probability based on the following equation:

\[
E(r) = \frac{R_0^6}{(R_0^6 + r^6)}
\]

\textbf{Eq. S24}

In each case, the transfer probability is compared to a random number between 0 and 1 to determine if the photon was emitted as a donor or acceptor photon. This process is repeated for each photon event in the trajectory, and the resulting acceptor photons are summed and divided by the total number of photons to yield the FRET efficiency for the simulated burst. This process is repeated for all bursts in our experimental photon trace and the resulting FRET efficiencies are summed and plotted as the displayed FRET histograms. As smFRET is highly sensitive to changes in timescale and MD simulations can be faster than experimental time, we fit the time rescaling factor, \( \alpha \), using the
Experimental FRET distributions (see Figure S24). The code for this kinetic Monte Carlo simulation has been developed as a command line app and is distributed via the enspara github (https://github.com/bowman-lab/enspara).

Interpretation of discrepancies between experimental and simulated single-molecule FRET data. As described in the previous section, calculation of single-molecule FRET data based upon MD simulations requires important assumptions. It is important to note that large discrepancies in transfer efficiencies do not directly transform in large discrepancies on distances, especially with transfer efficiency near 0.5. For example, assuming a fixed distance and a Förster radius of 5.4 nm, the difference between a mean value of transfer efficiency of 0.4 and 0.62 is equal to a change in distance from 5 to 5.7 nm. The following list represents current challenges in experiment and simulations that are commonly recognized in the field and that may lead to minor discrepancies such as the ones observed in our experiments. Indeed, it is often common to rescale simulations for matching experimental results or to use experimental results to constrain simulated conformational ensembles. While these are both valid alternatives, they introduce a bias due to the experimentally determined quantities. Few cases report an unbiased comparison of simulated transfer efficiencies and experiments and provide examples where mean transfer efficiencies can be as far apart even of 0.15 or 0.2, such as in ref. 75.

The role of $\kappa^2$. Not only does the mean transfer efficiency per burst depend on the average of multiple photons detected from different conformations of the protein, but the value of the Förster radius itself can be influenced by various factors. In Fig. 3, we compared simulated and experimental transfer efficiencies assuming the same Förster radius for all the constructs (5.4 nm). However, we know that the dipole orientation term $\kappa^2$ can vary across each construct and even across each subpopulation. We report the effect of variations of $\kappa^2$ across the experimentally determined boundaries (see corresponding section) on the simulated histogram in Fig. S8. Note that the distribution of values of $\kappa^2$ represents a distribution of possible solutions to Eq. S11-S16, but do not indicate a measure of correctness of a given solution. In other words, the experimental constrains can be skewed toward one of the extremes of the distribution. Variations in $\kappa^2$ represent a change in the accessible volume and relative orientation of the dyes. Overall, alterations in $\kappa^2$ are sufficient to explain all the mean transfer efficiencies discrepancies observed between experiments and simulations (Fig. S8). Small alterations in the Förster radius due to $\kappa^2$ can explain the larger overlap of populations observed in the experiments for
ApoE5,86 compared to simulations. The more significant deviations occur for constructs ApoE182,241 and ApoE86,241 and suggest that the shared position 241 may be the one involved in a significant alteration of the angles explored by the dyes.

**Donor quantum yield and dye quenching.** Another element entering in the Forster radius is the quantum yield of the dye. Though we observe no significant variations in the lifetime, we cannot exclude modulations of the quantum yield, which would directly impact the estimate of the Forster radius. In addition, in ns-FCS experiments (see Fig. S21), we observe minor correlated amplitude in the donor-acceptor cross correlation for ApoE4162,241 and ApoE4223,291 that hint to a possible contribution of quenching due to dye-quencher or dye-dye interactions. This contribution would result in a lower experimental transfer efficiency than the one expected in absence of quenching, though the contribution is probably minor.

**The role of cysteine mutations.** Introduction of dyes via cysteine-maleimide chemistry required the mutation of amino acids in the wild-type sequence to cysteines. Some of these mutations may alter some of the torsional angle properties of the chain. Indeed, while we chose residues that are part of unstructured regions, substitution of a glycine to cysteine (e.g., G165C and G182C) may result in an alteration of the dihedral angles and impact corresponding conformations. This can contribute to the larger discrepancy observed between experimental and simulated ApoE182,241 data (explaining why this construct requires a larger $\kappa^2$ than ApoE86,241), while it has no impact on ApoE86,165 because position 165 is at the bottom of the four-helix bundle.

**The role of the physico-chemical properties of the fluorophores.** Addition of explicit dyes in the simulation is not a feasible strategy, largely because it requires running 5 independent simulations (of several millisecond each) to test the effect of dyes on each position. We opted for a more effective strategy to insert dyes post-hoc. However, this simplification neglect to account for any specific chemistry and physical properties of the dyes (such as electrostatics).

**Simulations force-field.** While force-fields are constantly improved, there is still a lack of convergence on force-field that correctly reproduce disordered conformational ensembles. Small variations in the interaction energies between residues may influence the local organization of the protein and introduce discrepancies with the measured data, which can impact both conformations and the relative abundance of a specific subpopulation. Indeed, previous test of various force fields and water models on disordered proteins have reported very different results compared to experimentally determined configuration by
FRET (or other methods). Though ApoE is not completely unfolded, the choice of the specific force field may play a substantial role. Finally, while the simulation explored a large portion of the energy landscape, it may not have explored all the very low populated states, such as the intermediate and unfolded state reported for ApoE$_{86,165}$.

**Coarse-grained model of lipid-bound states.** While an atomically-detailed description of the protein requires all-atom simulations that are beyond the scope of this work, to provide a possible interpretation of the lipid-bound states observed with FRET measurements, we performed high coarse-grained MonteCarlo simulations of the protein that satisfy the FRET constraints. The scope of these simulations is purely illustrative and does not account for specific interactions with lipids or within the protein; instead, we aim only to provide a possible visual interpretation of the FRET data. To this end, we simplify the structure of the protein to a set of random coils (corresponding to the disordered regions) and rigid sticks (corresponding to the alpha helices), under the assumption that these elements are maintained in the lipid-bound state. Random coils are simulated by identifying the length of the segment from a distribution that follows the statistics of:

$$P(r, N) = (1/0.18 N^{-3/2}) 4\pi r^2 \text{Exp}[-3/2 r^2/(0.38 * 0.4 * N)]$$  \hspace{1cm} \text{Eq. S23}

based on previous estimates for a disordered region. Since disordered segments are often significantly shorts, their main effect is to contribute to protein flexibility and, therefore, we neglect effects of local compaction or expansion of disordered regions due to differences in sequence composition. To enforce bound geometries, we reject all configurations that are outside a spherical corona of radius 8.67 nm and thickness 0.1 nm. For disordered regions this entails checking whether the beginning or end of the segment are within the spherical corona. For the rigid regions, this is performed for the end, middle point, and beginning of the rigid region. We further allow for overlapping configurations of the protein to facilitate the search of configurations in which helices could come in contact. We start the simulation by simulating the protein region between 86 and 241, since it displays heterogeneity in the experimental data and entails three experimental constraints (as given by ApoE$_{486,165}$, ApoE$_{4182,241}$, and ApoE$_{486,241}$). Out of 500,000 productive results of the simulation, only 1 configuration is found to satisfy the experimental results for the expanded state of ApoE$_{486,165}$ and only 2 configurations satisfy the other state. We
then append to the resulting states an additional 50,000 configurations of the residues between 1 and 86 and 50'000 configurations of the residues between 241 and 299. By using the additional experimental constraints, we obtain a total of 39249 configurations of the protein, where 4361 elements differ for the N-terminal tail and 9 for the C-terminal domain. The lowest number of configurations identified in the C-terminal domain stems from the larger number of constraints imposed on position 241. Two examples of configurations are reported in Fig. 5 to provide an interpretative model of the lipid-bound state of the protein.

ApoE oligomerization.

Previous work on oligomerization kinetics of ApoE using ensemble FRET has provided estimates for the dimerization and tetramerization rates of ApoE4 and corresponding equilibrium dissociation constants for dimer and tetramer$^{18}$. The dissociation constant between the monomer and dimer state is determined to be $K_{D}^{mon-dim} \sim 80$ nM, whereas the dissociation constant between dimers and tetramers is $K_{D}^{dim-tet} \sim 20$ nM. As a result, previous analysis (see Fig. S1a and compare with Fig. 8d in $^{18}$) suggests that:
- at 10 nM of total protein concentration, $\sim$20% of ApoE4 molecules form dimers;
- at 100 nM of total protein concentration, $\sim$30% of ApoE4 molecules form dimers and an additional $\sim$30% is assembled in tetramers;
- at 1 µM of total protein concentration, $\sim$80% and $\sim$15% of ApoE4 molecules are in the tetramer and dimer forms, respectively;
- at 10 µM of total protein concentration, $\sim$95% of ApoE4 molecules are part of tetramer species.

To verify whether our measurements are performed in a regime under which the protein is monomeric, we evaluate changes of the labeling stoichiometry ratio as function of GdmCl. Increasing denaturant concentration shifts the equilibrium stability of dimer and tetramers and at high GdmCl concentration the protein is only monomeric. We quantify changes in labeling stoichiometry across five different regions of the stoichiometry ratio: below 0.125 (acceptor-only), between 0.125 and 0.375 (3A:1D), between 0.375 and 0.625 (1D:1A), between 0.625 and 0.875 (3D:1A), and above 0.875 (acceptor-only). In all measurements (e.g., Fig. S1b) across all denaturant concentrations we observe only three populations centered at 0.5 stoichiometry ratio (1D:1A) and at donor- and acceptor-only stoichiometry ratio.
Note that for the labeling procedure, we do expect to have saturated labeling of the cysteines for donor and acceptor only populations (2 donor dyes for donor-only, 2 acceptor dyes for acceptor-only) and therefore we expect 4 fluorophores in a labeled dimer. Based on the labeling procedure, the fraction of unlabeled molecules is negligible. Fraction of molecules at different stoichiometries (3D:1A or 3A:1D) represents symmetric tails of the central distribution at 0.5 stoichiometry and the small deviations reported in Fig. S1c should be interpreted as modulations in the brightness or quantum yield of the labeled species. If oligomers were present in solution, we would expect to observe a decrease of intermediate stoichiometries in favor of other stoichiometries, in particular donor- and acceptor-only. If any percentage of oligomers is present is within the ~5% variation reported in the fluctuations in Fig. S1c.
Supplementary Figures.
Figure S1: Oligomerization of Apolipoprotein E. 

a. Estimates of the concentration-dependent fractions of monomer (blue), dimer (cyan), and tetramer (black) for ApoE4 based on ref. 16. 

b. Example of labeling stoichiometry ratio for ApoE4223,291. 

c. Fractions associated with the different labeling stoichiometry ratios across multiple GdmCl concentration. Across all concentrations of denaturant, the fraction of population with 1D:1A is mainly constant supporting the absence of oligomer formation. In the case of oligomer formation, dimers and tetramers of acceptor-only (Aonly) or donor-only (Donly) population would contribute to an increase in the 1D:1A, which is not observed (overall deviations are within a few percent).
Figure S2: Lifetime vs transfer efficiency plot and corresponding distance distributions for each of the five ApoE4 constructs measured in aqueous buffer conditions (50 mM NaPi pH 7.4). The solid red line indicates the expected result for a rigid distance. The solid black line reports the expected trend for a Gaussian chain distribution whose mean square end-to-end distance satisfies the measured mean transfer efficiency. The colored line represents the best fit to a wormlike chain, where both persistence length and contour length are fitted to satisfy the constraints imposed by the measured mean transfer efficiency and mean lifetime.
Figure S3. Folding free energy and midpoint of the transition determined from the folding equilibrium analysis. Plot of fitting parameters in Table S2 obtained from the folding equilibrium analysis of Fig. 2 with Eq. S4. Blue, red, and green line represent the corresponding states from which the $\Delta G_0$ are computed. Gray (ApoE4$_{5,86}$), teal (ApoE4$_{86,165}$), green (ApoE4$_{182,241}$), purple (ApoE4$_{223,291}$) and violet bars (ApoE4$_{86,241}$) connect the states associated with the computed $\Delta G_0$. Width of the bars represents the associated error on the corresponding $c_{1/2}$, whereas vertical error bars quantify the error (from the fit) associated with $\Delta G_0$. In some cases, error bars may be occluded by the data points. Dashed vertical line in ApoE4$_{223,291}$ identifies the transition in Fig. S5 when analyzed with a two-state model.
Figure S4. Folding stability of individual constructs. Free energy diagram of identified states in the 4-helix bundle, hinge, N- and C- terminus, and from long range measurements. Source for the composed picture reported in Fig. 3c.
Figure S5: Widths of transfer efficiency distributions attributed to independent populations in each construct. Solid lines are polynomial fits to help guiding the eyes. The widths of ApoE4_{5,86} and ApoE4_{86,241} exhibit an increase between 1.25 and 2.5 M GdmCl suggesting coexistence of two independent configurations under the same peak, which we interpret as a conformational shift occurring during the folding of the four-helix bundle. In ApoE4_{182,241} the widths of the two populations are best described using a shared parameter for each population in the global fit of the denaturant titration. In ApoE4_{223,291}, the width of the main population increases when decreasing GdmCl concentration from 1 M to 0 M, which is compatible with the range where previous experiments identified formation of structure in the C-terminal domain. Due to the small amplitudes of the low and high transfer efficiencies populations in ApoE4_{223,291}, the corresponding widths were fixed across all concentrations. Similarly, in ApoE4_{86,165}, due to the small amplitude of the unfolded and intermediate state, below 3 M GdmCl the widths of the corresponding transfer efficiency distribution were fixed. Errors associated with repeated measurements are reported in Table S1.
Figure S6. Analysis of the ApoE45,86 construct using one single population in the range from 0.5 and 6 M GdmCl. When fitting the transfer efficiency with one single population, we observe a jump in transfer efficiencies (lower panel) and a broadening of the width (see Fig. S5) in the range between 1.5 and 2.5 M GdmCl. The range of the transition corresponds to the folding transition of the four-helix bundle. Corresponding values of the fit are reported in Table S2.
**Figure S7: Comparison of FAST simulations of ApoE4 and ApoE3.** Distance distributions obtained from MSMs constructed by clustering the simulations on distances between 15 residue pairs as described in Methods for A) 3.43 ms of aggregate simulations of ApoE4 performed on Folding@Home with a lag time of 10 ns. B) 8 us of FAST simulations of ApoE4 performed as described in Methods with a lag time of 10 ns and C) 8 us of FAST simulations of ApoE3 performed in the same method as ApoE4. **Upper panels:** ApoE4_{223,291} **Lower panels:** ApoE4_{86,241} FAST simulations capture only a small portion of the complex energy landscape of the protein observed in Folding@Home simulations.
Figure S8. Simulated transfer efficiency histograms as function of $\kappa^2$. Simulated transfer efficiency histograms with $\kappa^2$ values across the range established from anisotropy measurements. Colored vertical lines identify mean values of transfer efficiency for each population in single-molecule FRET experiments. Note that for each subpopulation measured across the same construct, a different $\kappa^2$ can be adopted.
Figure S9. Correlations across different pair distances in lipid-free ApoE4. Distance pair correlations from MD simulations contrasting all labeled distance pairs.
Figure S10. Contacts in lipid-free ApoE4. Probability contact map from all structures from MD simulations contrasted against contacts found in the ApoE3-like NMR structure (PDB:2L7B). Contacts defined as any heavy atom between two residues being within 3 Å. Contacts that are present only in the NMR structure are reported in orange. Black boxes identify contacts that are due to domain-domain interactions.
Figure S11. Solvent Accessible Surface Area and structural content of the C-terminal region. a. Solvent Accessible Surface Area. Specific residues between 230 and 280 are more protected in the extended conformation. b. Secondary structure content in the C-terminal domain across the three closed, open, and extended subpopulations supports the presence of a structured helix with some degree of flexibility because of disordered regions.
Figure S12. Graphical representation of the MSM states belonging to the open, closed and extended configurations identified in Fig 4. The nodes representing each state are sized proportional to the sum of the equilibrium probabilities of all the MSM states belonging to that node. The arrows indicate aggregate transition probabilities between the nodes.
Figure S13: Size distribution of extruded liposomes determined from cryo-TEM images (inset). The bar on the inset figure represents 100 nm.
Figure S14: Lifetime vs transfer efficiency plot and corresponding distance distributions for each of the five ApoE4 constructs measured in the lipid-bound state (100 µg/mL extruded liposomes, 50 mM NaPi pH 7.4). The solid red line indicates the expected result for a rigid distance. The solid black line reports the expected trend for a Gaussian distribution whose mean square end-to-end distance satisfies the measured mean transfer efficiency. The colored line represents the best fit to a wormlike chain, where both persistence length and contour length are fitted to satisfy the constraints imposed by the measured mean transfer efficiency and mean lifetime.
Figure S15: Fractions of acceptor only (red), donor only (green), and acceptor-donor (orange) labeled molecules as identified by Pulsed Interleaved Excitation (PIE) of lipid-free (dark shaded areas) and DMPC-bound (light shaded areas) full-length ApoE constructs. Error bars are standard deviations from multiple measurements (see Table S13 for details). In all cases, binding to lipids results in a small decrease (up to about 15%) of the acceptor-donor labeled molecules often accompanied by a small increase in the fraction of donor and acceptor only population. These data support that only one labeled molecule of ApoE4 is binding to lipids since binding of two molecules would result in an increase of the donor-acceptor population.
Figure S16: Estimation of distances for the lipid-bound state assuming a fixed distance is adopted.
**Figure S17**: Hydrodynamic radii ($R_H$) of lipid-bound full-length ApoE4 constructs as measured by FCS. Mean $R_H$ calculated from both donor-donor and acceptor$_{PIE}$-acceptor$_{PIE}$ correlations for each full-length construct assuming one diffusing species and two triplet components. Donor-donor and acceptor$_{PIE}$-acceptor$_{PIE}$ correlated data points are represented by dark and light blue dots, respectively. Box whisker plots report the median value as a white line, the 25% and 75% quantile, and the max and min values of the distribution for the combined donor-donor and acceptor$_{PIE}$-acceptor$_{PIE}$ datasets. Mean $R_H$ values for lipid-bound (solid gray line) and lipid-free (dashed gray line) full length ApoE4 (compare with Figure S18).
Figure S18: Hydrodynamic radii of lipid-free full-length ApoE4 constructs as measured by df-FCS assuming one diffusing species and two triplet components. Variations across each construct are within a 10% of relative error and they are indistinguishable from each other according to a one-tail t-student test with significance level ≤ 0.1. Black diamonds indicate the mean $R_H$ value for each construct, and the error bars span the standard error of the mean. Dashed line denotes the overall weighted mean, and the darker shaded area spans the overall weighted SE of the mean –using individual $1/\text{SE}^2$ values as weights. Lighter shaded area denotes the overall 90% confidence interval around the overall weighted mean.
Figure S19. Comparison of Folding@home simulation results ApoE3-like NMR structure, 2L7B. Displayed are the probabilities of a contact (<3 Å between any heavy atom on the two sidechains) occurring between each pairwise residue in ApoE4 according to the Folding@home equilibrium probabilities. Above the diagonal, orange marks indicate a distance predicted in the ApoE3-like NMR structure, 2L7B, but are less than 10% likely in our simulations. Below the diagonal any contacts that occur in 2L7B and are more than 10% likely in our simulations are displayed in orange.
**Figure S20.** Predicted salt bridge interactions observed in our simulations. In blue are the equilibrium probabilities of any heavy atom in the residue sidechains being within 6Å of the other sidechain. In orange are interactions predicted by the ApoE3-like NMR structure, 2L7B, which are less than 10% likely in our simulations.
Figure S21. Dynamics in the C-terminal tail. a. Correlation time and amplitudes from global fit of nanosecond FCS data for ApoE4_{182,241} and ApoE4_{223,291}. The measurements highlight dynamics in the 100 ns timescale. Note that at 0M a positive amplitude is observed for the cross-correlation of both constructs suggesting a possible contribution of quenching of the fluorophores. The small effect disappears at 0.5 M GdmCl and higher.
denaturant concentrations. The associated timescale appears on trend with the ones measured at 0 M GdmCl. Error bars represents the error of the fit. 

b. Burst variance analysis of ApoE4_{182,241} and ApoE4_{223,291}, as described in the Supplementary Information. Red dashed line represents the theoretical expectation of the standard deviation of the proximity ratio E*, blue solid line represents the mean value in proximity ration (binned in steps of 0.02), and red dots represents the mean value of the three major identified populations. Small deviations are observed for the low transfer efficiency peak of ApoE4_{182,241} and for the intermediate population of ApoE4_{223,291}, suggesting a possible contribution of dynamics in the microsecond timescale.

c. Analysis of identified transfer efficiency distribution as function of different time binning choices, from 0.25 ms (blue), to 0.50 ms (magenta), 0.75 ms (yellow), and 1.00 ms (green). No significant variation is observed in the normalized histogram of transfer efficiencies suggesting that interconversion dynamics between the different states would occur on longer timescales, larger than 1 ms.
**Figure S22**: GdmCl dependent changes in donor-only and acceptor-only lifetimes. Error bars are standard deviation from multiple measurements (Table S6).
Figure S23: Instrument Response Function (IRF) determination for donor and acceptor detection. a-c. Fluorescence lifetime histogram of Rhodamine B in water with normalized counts against the maximum of the distribution ($I/I_{\text{max}}$) at different concentrations of Potassium Ferrocyanide (a) and corresponding steady-state anisotropies (b) and mean photon arrival time $\tau$ (c). d-e. Normalized fluorescence lifetime histograms of Alexa Fluor dyes and Rhodamine B in presence of saturated K$_3$Fe(CN)$_6$ with normalized counts against the total area $I/I_{\text{tot}}$ for donor (d) and acceptor (e) excitations wavelengths.
Figure S24. Optimization of time correction for simulations. a. Comparison between experimental and simulated transfer efficiency histograms with different time correction factors. b. The time correction factors in a. represent the minima obtained by optimizing the factor using the sum of the residuals for the first four moments of the distributions (upper panel, time factor equals to 225), the sum of the squared residuals across each histogram normalized by maximum excursion (central panel, time factor equals to 375), and the sum of the squared residuals for only the ApoE4223,291 construct (lower panel, time factor equals to 525).
Supplementary Tables

Table S1. Mean transfer efficiencies, widths, and relative areas obtained from the fit of lipid-free ApoE4 smFRET histograms. $E_1$, $E_2$, and $E_3$ represent the mean value of transfer efficiencies for each identified population in the transfer efficiency histograms determined at 0, 1, 2, and 5 M GdmCl (compare with Fig. 2) for lipid-free ApoE4. $W_1$, $W_2$, and $W_3$ and $A_1$, $A_2$, and $A_3$ reports about the corresponding width of the distribution and relative area. Errors are standard deviation of “n” independent repeats of the experiments.

<table>
<thead>
<tr>
<th>GdmCl (M)</th>
<th>$E_1$ (%)</th>
<th>$E_2$ (%)</th>
<th>$E_3$ (%)</th>
<th>$W_1$ (%)</th>
<th>$W_2$ (%)</th>
<th>$W_3$ (%)</th>
<th>$A_1$ (%)</th>
<th>$A_2$ (%)</th>
<th>$A_3$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.21±0.05</td>
<td>0.61±0.02</td>
<td>-</td>
<td>0.14±0.01</td>
<td>0.118±0.008</td>
<td>-</td>
<td>12±5</td>
<td>88±5</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>0.46±0.05</td>
<td>0.57±0.04</td>
<td>-</td>
<td>0.097±0.009</td>
<td>0.097±0.009</td>
<td>-</td>
<td>9±5</td>
<td>91±5</td>
<td>-</td>
</tr>
<tr>
<td>2*</td>
<td>-</td>
<td>0.5±0.1</td>
<td>-</td>
<td>-</td>
<td>0.113±0.006</td>
<td>-</td>
<td>-</td>
<td>100</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>-</td>
<td>0.23±0.05</td>
<td>-</td>
<td>-</td>
<td>0.074±0.003</td>
<td>-</td>
<td>-</td>
<td>100</td>
<td>-</td>
</tr>
</tbody>
</table>

*fit obtained assuming a single population for the main transfer efficiency.

<table>
<thead>
<tr>
<th>GdmCl (M)</th>
<th>$E_1$ (%)</th>
<th>$E_2$ (%)</th>
<th>$E_3$ (%)</th>
<th>$W_1$ (%)</th>
<th>$W_2$ (%)</th>
<th>$W_3$ (%)</th>
<th>$A_1$ (%)</th>
<th>$A_2$ (%)</th>
<th>$A_3$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.345</td>
<td>0.62</td>
<td>0.98±0.01</td>
<td>0.10</td>
<td>0.10</td>
<td>0.14±0.03</td>
<td>0.84</td>
<td>5.8±0.2</td>
<td>9±2</td>
</tr>
<tr>
<td>1</td>
<td>0.345</td>
<td>0.608</td>
<td>0.978±0.005</td>
<td>0.10</td>
<td>0.10</td>
<td>0.15±0.01</td>
<td>0.84</td>
<td>6±2</td>
<td>12±3</td>
</tr>
<tr>
<td>2</td>
<td>0.41±0.02</td>
<td>0.63±0.02</td>
<td>0.973±0.005</td>
<td>0.099±0.003</td>
<td>0.10</td>
<td>0.15±0.02</td>
<td>0.84</td>
<td>51±10</td>
<td>24±3</td>
</tr>
<tr>
<td>5</td>
<td>0.23±0.02</td>
<td>-</td>
<td>-</td>
<td>0.076±0.009</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GdmCl (M)</th>
<th>$E_1$ (%)</th>
<th>$E_2$ (%)</th>
<th>$E_3$ (%)</th>
<th>$W_1$ (%)</th>
<th>$W_2$ (%)</th>
<th>$W_3$ (%)</th>
<th>$A_1$ (%)</th>
<th>$A_2$ (%)</th>
<th>$A_3$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.67±0.02</td>
<td>0.83±0.02</td>
<td>-</td>
<td>0.078</td>
<td>0.09</td>
<td>-</td>
<td>61±6</td>
<td>38±6</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>0.50±0.02</td>
<td>0.69±0.03</td>
<td>-</td>
<td>0.079</td>
<td>0.09</td>
<td>-</td>
<td>77±6</td>
<td>23±6</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>0.44±0.02</td>
<td>0.62±0.04</td>
<td>-</td>
<td>0.074±0.009</td>
<td>0.09</td>
<td>-</td>
<td>89±1</td>
<td>11±1</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>0.32±0.02</td>
<td>0.56±0.02</td>
<td>-</td>
<td>0.074±0.009</td>
<td>0.09</td>
<td>-</td>
<td>94±3</td>
<td>6±3</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GdmCl (M)</th>
<th>$E_1$ (%)</th>
<th>$E_2$ (%)</th>
<th>$E_3$ (%)</th>
<th>$W_1$ (%)</th>
<th>$W_2$ (%)</th>
<th>$W_3$ (%)</th>
<th>$A_1$ (%)</th>
<th>$A_2$ (%)</th>
<th>$A_3$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.13±0.04</td>
<td>0.61±0.02</td>
<td>0.851±0.003</td>
<td>0.105±0.004</td>
<td>0.13</td>
<td>0.104±0.001</td>
<td>27±4</td>
<td>58±4</td>
<td>15±5</td>
</tr>
<tr>
<td>1</td>
<td>0.15±0.06</td>
<td>0.46±0.02</td>
<td>0.85</td>
<td>0.104</td>
<td>0.095±0.003</td>
<td>0.104</td>
<td>10±6</td>
<td>85±6</td>
<td>5±1</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>0.37±0.02</td>
<td>-</td>
<td>-</td>
<td>0.093±0.009</td>
<td>-</td>
<td>-</td>
<td>100</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>-</td>
<td>0.27±0.02</td>
<td>-</td>
<td>-</td>
<td>0.079±0.006</td>
<td>-</td>
<td>-</td>
<td>100</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GdmCl (M)</th>
<th>$E_1$ (%)</th>
<th>$E_2$ (%)</th>
<th>$E_3$ (%)</th>
<th>$W_1$ (%)</th>
<th>$W_2$ (%)</th>
<th>$W_3$ (%)</th>
<th>$A_1$ (%)</th>
<th>$A_2$ (%)</th>
<th>$A_3$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Lipid-free ApoE4
<table>
<thead>
<tr>
<th>GdmCl (M)</th>
<th>E₁</th>
<th>E₂</th>
<th>E₃</th>
<th>W₁</th>
<th>W₂</th>
<th>W₃</th>
<th>A₁ (%)</th>
<th>A₂ (%)</th>
<th>A₃ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 n=5</td>
<td>0.24±0.01</td>
<td>0.59±0.02</td>
<td>0.87±0.02</td>
<td>0.12</td>
<td>0.113±0.007</td>
<td>0.09±0.01</td>
<td>7 ± 1</td>
<td>81 ± 1</td>
<td>12 ± 1</td>
</tr>
<tr>
<td>1 n=5</td>
<td>0.16±0.02</td>
<td>0.40±0.03</td>
<td>0.80</td>
<td>0.06±0.02</td>
<td>0.089±0.003</td>
<td>0.12</td>
<td>2.2 ± 0.3</td>
<td>92 ± 1</td>
<td>6 ± 1</td>
</tr>
<tr>
<td>2 n=4</td>
<td>-</td>
<td>0.25±0.04</td>
<td>-</td>
<td>-</td>
<td>0.106±0.008</td>
<td>-</td>
<td>-</td>
<td>100</td>
<td>-</td>
</tr>
<tr>
<td>5 n=4</td>
<td>-</td>
<td>0.10±0.03</td>
<td>-</td>
<td>-</td>
<td>0.06±0.01</td>
<td>-</td>
<td>-</td>
<td>100</td>
<td>-</td>
</tr>
</tbody>
</table>

Table S2. Folding free energy and midpoint of the transition determined from folding equilibrium analysis.

<table>
<thead>
<tr>
<th>3-state model</th>
<th>ApoE₄₆₆,₁₆₅</th>
</tr>
</thead>
<tbody>
<tr>
<td>c₁/₂ UN (M)</td>
<td>1.91±0.015</td>
</tr>
<tr>
<td>ΔG₀ UN (RT)</td>
<td>-8.3±4.2</td>
</tr>
<tr>
<td>c₁/₂ NI (M)</td>
<td>3±0.2</td>
</tr>
<tr>
<td>ΔG₀ NI (RT)</td>
<td>2.7±0.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3-state model</th>
<th>ApoE₄₅₈₆</th>
<th>ApoE₄₁₈₂,₂₄₁</th>
<th>ApoE₄₂₂₃,₂₉₁</th>
<th>ApoE₄₆₆,₂₄₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>c₁/₂ UN₁ (M)</td>
<td>2.06±0.01</td>
<td>1.9±0.2</td>
<td>-0.8±0.1</td>
<td>-1.00±0.07</td>
</tr>
<tr>
<td>ΔG₀ UN₁ (RT)</td>
<td>-5.2±0.2</td>
<td>-8±7</td>
<td>1.15±0.08</td>
<td>1.47±0.04</td>
</tr>
<tr>
<td>c₁/₂ NI₂ (M)</td>
<td>-0.6±0.4</td>
<td>-0.3±0.1</td>
<td>1.40±0.09</td>
<td>-1.2±0.7</td>
</tr>
<tr>
<td>ΔG₀ NI₂ (RT)</td>
<td>2.1±0.4</td>
<td>1.0±0.2</td>
<td>1.6±0.1</td>
<td>0.7±0.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2-state model</th>
<th>ApoE₄₅₈₆</th>
</tr>
</thead>
<tbody>
<tr>
<td>c₁/₂ UL (M)</td>
<td>2.07±0.03</td>
</tr>
<tr>
<td>ΔG₀ UL (RT)</td>
<td>-5.4±0.4</td>
</tr>
<tr>
<td>c₁/₂ NI₂ (M)</td>
<td>-0.41±0.06</td>
</tr>
<tr>
<td>ΔG₀ NI₂ (RT)</td>
<td>1.76±0.08</td>
</tr>
</tbody>
</table>

Table S3. Folding of ApoE₄ as reported in literature. c₁/₂ and ΔG₀ UN previously reported in literature.

<table>
<thead>
<tr>
<th>Region</th>
<th>Residues</th>
<th>c₁/₂ (M)</th>
<th>ΔG₀ UN (kcal mol⁻¹)</th>
<th>reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Four-helix bundle</td>
<td>1-183</td>
<td>2</td>
<td>-</td>
<td>Weers et al. 2003²⁷</td>
</tr>
<tr>
<td>Four-helix bundle</td>
<td>1-191</td>
<td>2</td>
<td>-</td>
<td>Morrow et al. 2000²⁵</td>
</tr>
<tr>
<td>N-terminal domain</td>
<td>full-length</td>
<td>2.6 ± 1.3</td>
<td>5.25 ± 0.20</td>
<td>Dolai et al. ²⁶</td>
</tr>
<tr>
<td>C-terminal domain</td>
<td>full-length</td>
<td>0.73 ± 0.07</td>
<td>3.29 ± 0.20</td>
<td>Dolai et al. ²⁶</td>
</tr>
<tr>
<td>N-terminal domain (1-191)</td>
<td>fragment</td>
<td>-</td>
<td>4.56 ± 0.03</td>
<td>Dolai et al. ²⁶</td>
</tr>
<tr>
<td>C-terminal domain (192-299)</td>
<td>fragment</td>
<td>-</td>
<td>1.58 ± 0.08</td>
<td>Dolai et al. ²⁶</td>
</tr>
</tbody>
</table>
Table S4. Mean transfer efficiencies, widths, and relative areas obtained from the fit of lipid-free ApoE4 smFRET histograms. $E_1$, $E_2$, and $E_3$ represent the mean value of transfer efficiencies for the each identified population in the transfer efficiency histograms determined at 0, 1, 2, and 5 M GdmCl (compare with Fig. 7.) for lipid-bound ApoE4. $W_1$, $W_2$, and $W_3$ and $A_1$, $A_2$, and $A_3$ reports about the corresponding width of the distribution and relative area. Errors are standard deviation of “n” independent repeats of the experiments.

<table>
<thead>
<tr>
<th>Construct</th>
<th>$E_1$ (%)</th>
<th>$E_2$ (%)</th>
<th>$E_3$ (%)</th>
<th>$W_1$ (%)</th>
<th>$W_2$ (%)</th>
<th>$W_3$ (%)</th>
<th>$A_1$ (%)</th>
<th>$A_2$ (%)</th>
<th>$A_3$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5, 86</td>
<td>0.111±0.006</td>
<td>-</td>
<td>-</td>
<td>0.073±0.002</td>
<td>-</td>
<td>-</td>
<td>100</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>86, 165</td>
<td>0.037±0.006</td>
<td>0.894±0.004</td>
<td>-</td>
<td>0.17±0.01</td>
<td>0.080±0.009</td>
<td>-</td>
<td>51 ± 16</td>
<td>49 ± 16</td>
<td>-</td>
</tr>
<tr>
<td>182, 241</td>
<td>0.09±0.01</td>
<td>0.744</td>
<td>-</td>
<td>0.21±0.03</td>
<td>0.137</td>
<td>-</td>
<td>91 ± 2</td>
<td>9 ± 2</td>
<td>-</td>
</tr>
<tr>
<td>223, 291</td>
<td>0.091±0.008</td>
<td>-</td>
<td>-</td>
<td>0.072±0.002</td>
<td>-</td>
<td>-</td>
<td>100</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>86, 241</td>
<td>0.05±0.003 (1.57±0.04)</td>
<td>-</td>
<td>-</td>
<td>0.215±0.007</td>
<td>-</td>
<td>-</td>
<td>100</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

(=Asymmetry of lognormal distribution)

Table S5. Correction parameters for anisotropy determinations. Anisotropy determination through confocal detection requires correction factors accounting for the detection efficiency on different parallel and perpendicular detectors (G-factor) as well as for scrambling of polarization due to tight focusing optics ($L_p$ and $L_s$ factors). Each parameter has been determined for Donor and Acceptor excitation and for corresponding detection.

<table>
<thead>
<tr>
<th>Excitation</th>
<th>Detection</th>
<th>$L_s$</th>
<th>$L_p$</th>
<th>$G$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Donor</td>
<td>Acceptor</td>
<td>0.036</td>
<td>0.218</td>
<td>0.776</td>
</tr>
<tr>
<td>Donor</td>
<td>Donor</td>
<td>0.072</td>
<td>0.126</td>
<td>0.992</td>
</tr>
<tr>
<td>Acceptor</td>
<td>Acceptor</td>
<td>0.042</td>
<td>0.169</td>
<td>1.29</td>
</tr>
</tbody>
</table>
Table S6. Subpopulation-specific analysis of time-resolved fluorescence decays for lipid-free ApoE. Lifetime components obtained by fitting lifetime decays for donor only subpopulation, donor-acceptor population, and acceptor only population after PIE excitation. For each subpopulation, the lifetime component $\tau$ and the rotational component $\tau_{rot}$ are identified. From the comparison between the lifetime decay $\tau_{DA}$ decay and the lifetime of the donor in absence of acceptor, $\tau_{Donly}$, we estimated the energy transfer characteristic time $\tau_{FRET}$. Data are further subdivided based on the corresponding mean transfer efficiency (as identified in Table S1) and error are standard deviations of “n” independent repeats of the measurement.

<table>
<thead>
<tr>
<th>ApoE4_{5, 86}</th>
<th>GdmCl (M)</th>
<th>$\tau_{Donly}$ (ns)</th>
<th>$\tau_{rot-Donly}$ (ns)</th>
<th>$\tau_{DA}$ (ns)</th>
<th>$\tau_{rot-DA}$ (ns)</th>
<th>$\tau_{FRET}$ (ns)</th>
<th>$\tau_{Aonly}$ (ns)</th>
<th>$\tau_{rot-Aonly}$ (ns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 E_{1} n=6</td>
<td>4.06±0.08</td>
<td>0.43±0.07</td>
<td>3.28±0.08</td>
<td>0.7±0.1</td>
<td>17±3</td>
<td>3.98±0.02</td>
<td>0.54±0.04</td>
<td></td>
</tr>
<tr>
<td>0 E_{2} n=6</td>
<td>4.06±0.08</td>
<td>0.44±0.06</td>
<td>2.13±0.07</td>
<td>0.62±0.06</td>
<td>4.5±0.3</td>
<td>3.98±0.02</td>
<td>0.53±0.04</td>
<td></td>
</tr>
<tr>
<td>1 E_{1} n=3</td>
<td>4.1±0.2</td>
<td>0.52±0.06</td>
<td>2.28±0.04</td>
<td>0.58±0.02</td>
<td>5.1±0.4</td>
<td>3.98±0.004</td>
<td>0.7±0.1</td>
<td></td>
</tr>
<tr>
<td>2 E_{1} n=3</td>
<td>3.9±0.2</td>
<td>0.51±0.01</td>
<td>2.45±0.03</td>
<td>0.61±0.02</td>
<td>6.6±0.6</td>
<td>3.89±0.03</td>
<td>0.79±0.03</td>
<td></td>
</tr>
<tr>
<td>5 E_{1} n=3</td>
<td>3.67±0.08</td>
<td>0.60±0.03</td>
<td>2.90±0.01</td>
<td>0.58±0.01</td>
<td>16±2</td>
<td>3.78±0.01</td>
<td>0.99±0.09</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ApoE4_{86, 165}</th>
<th>GdmCl (M)</th>
<th>$\tau_{Donly}$ (ns)</th>
<th>$\tau_{rot-Donly}$ (ns)</th>
<th>$\tau_{DA}$ (ns)</th>
<th>$\tau_{rot-DA}$ (ns)</th>
<th>$\tau_{FRET}$ (ns)</th>
<th>$\tau_{Aonly}$ (ns)</th>
<th>$\tau_{rot-Aonly}$ (ns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 E_{3} n=3</td>
<td>3.78±0.03</td>
<td>0.32±0.04</td>
<td>0.68±0.05</td>
<td>0.225±0.008</td>
<td>0.83±0.07</td>
<td>3.86±0.01</td>
<td>0.40±0.01</td>
<td></td>
</tr>
<tr>
<td>1 E_{3} n=3</td>
<td>3.77±0.06</td>
<td>0.48±0.02</td>
<td>0.82±0.02</td>
<td>0.30±0.01</td>
<td>1.05±0.03</td>
<td>3.77±0.02</td>
<td>0.51±0.02</td>
<td></td>
</tr>
<tr>
<td>2 E_{3} n=2</td>
<td>3.67±0.02</td>
<td>0.61±0.03</td>
<td>0.86±0.09</td>
<td>0.29±0.01</td>
<td>1.1±0.2</td>
<td>3.79±0.02</td>
<td>0.75±0.08</td>
<td></td>
</tr>
<tr>
<td>2 E_{3} n=2</td>
<td>3.67±0.02</td>
<td>0.61±0.03</td>
<td>2.30±0.04</td>
<td>0.578±0.009</td>
<td>6.2±0.3</td>
<td>3.79±0.02</td>
<td>0.75±0.08</td>
<td></td>
</tr>
<tr>
<td>5 E_{3} n=3</td>
<td>3.55±0.03</td>
<td>0.72±0.009</td>
<td>3.03±0.006</td>
<td>0.56±0.01</td>
<td>21±1</td>
<td>3.77±0.01</td>
<td>1.03±0.05</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ApoE4_{182, 241}</th>
<th>GdmCl (M)</th>
<th>$\tau_{Donly}$ (ns)</th>
<th>$\tau_{rot-Donly}$ (ns)</th>
<th>$\tau_{DA}$ (ns)</th>
<th>$\tau_{rot-DA}$ (ns)</th>
<th>$\tau_{FRET}$ (ns)</th>
<th>$\tau_{Aonly}$ (ns)</th>
<th>$\tau_{rot-Aonly}$ (ns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 E_{2} n=4</td>
<td>4.11±0.04</td>
<td>0.62±0.02</td>
<td>1.24±0.07</td>
<td>0.31±0.01</td>
<td>1.8±0.1</td>
<td>3.91±0.02</td>
<td>0.52±0.03</td>
<td></td>
</tr>
<tr>
<td>0 E_{1} n=4</td>
<td>4.11±0.04</td>
<td>0.62±0.02</td>
<td>2.07±0.04</td>
<td>0.52±0.02</td>
<td>6.2±0.2</td>
<td>3.91±0.02</td>
<td>0.52±0.03</td>
<td></td>
</tr>
<tr>
<td>1 E_{2} n=3</td>
<td>4.02±0.05</td>
<td>0.68±0.06</td>
<td>2.47±0.02</td>
<td>0.65±0.05</td>
<td>6.4±0.2</td>
<td>3.925±0.008</td>
<td>0.9±0.01</td>
<td></td>
</tr>
<tr>
<td>1 E_{1} n=3</td>
<td>4.02±0.05</td>
<td>0.68±0.06</td>
<td>2.63±0.01</td>
<td>0.74±0.04</td>
<td>7.6±0.2</td>
<td>3.925±0.008</td>
<td>0.9±0.01</td>
<td></td>
</tr>
<tr>
<td>2 E_{2} n=3</td>
<td>3.91±0.04</td>
<td>0.71±0.05</td>
<td>2.55±0.03</td>
<td>0.63±0.03</td>
<td>7.3±0.3</td>
<td>3.93±0.03</td>
<td>1.0±0.1</td>
<td></td>
</tr>
<tr>
<td>2 E_{1} n=3</td>
<td>3.91±0.04</td>
<td>0.71±0.05</td>
<td>2.77±0.02</td>
<td>0.775±0.009</td>
<td>9.5±0.3</td>
<td>3.93±0.03</td>
<td>1.0±0.1</td>
<td></td>
</tr>
<tr>
<td>5 E_{1} n=3</td>
<td>3.69±0.07</td>
<td>0.75±0.04</td>
<td>2.85±0.01</td>
<td>0.72±0.01</td>
<td>13±1</td>
<td>3.83±0.04</td>
<td>1.15±0.05</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ApoE4_{223, 291}</th>
<th>GdmCl (M)</th>
<th>$\tau_{Donly}$ (ns)</th>
<th>$\tau_{rot-Donly}$ (ns)</th>
<th>$\tau_{DA}$ (ns)</th>
<th>$\tau_{rot-DA}$ (ns)</th>
<th>$\tau_{FRET}$ (ns)</th>
<th>$\tau_{Aonly}$ (ns)</th>
<th>$\tau_{rot-Aonly}$ (ns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 E_{3} n=4</td>
<td>4.1±0.1</td>
<td>0.6±0.2</td>
<td>1.59±0.08</td>
<td>0.4±0.04</td>
<td>2.6±0.2</td>
<td>3.99±0.02</td>
<td>0.68±0.02</td>
<td></td>
</tr>
<tr>
<td>0 E_{2} n=4</td>
<td>4.1±0.1</td>
<td>0.6±0.2</td>
<td>2.33±0.01</td>
<td>0.59±0.03</td>
<td>5.4±0.2</td>
<td>3.99±0.02</td>
<td>0.68±0.02</td>
<td></td>
</tr>
<tr>
<td>0 E_{1} n=4</td>
<td>4.1±0.1</td>
<td>0.6±0.2</td>
<td>3.35±0.02</td>
<td>0.6±0.1</td>
<td>18±2</td>
<td>3.99±0.02</td>
<td>0.68±0.02</td>
<td></td>
</tr>
<tr>
<td>1 E_{2} n=3</td>
<td>3.9±0.1</td>
<td>0.7±0.1</td>
<td>2.65±0.03</td>
<td>0.74±0.02</td>
<td>8.3±0.5</td>
<td>3.95±0.02</td>
<td>0.92±0.03</td>
<td></td>
</tr>
</tbody>
</table>
Apoe4, 241
GdmCl (M) | Construct | E | n | \(\tau_{\text{Only}}\) (ns) | \(\tau_{\text{rot-Only}}\) (ns) | \(\tau_{\text{DA}}\) (ns) | \(\tau_{\text{rot-DA}}\) (ns) | \(\tau_{\text{FRET}}\) (ns) | \(\tau_{\text{AllOnly}}\) (ns) | \(\tau_{\text{rot-AllOnly}}\) (ns)
--- | --- | --- | --- | --- | --- | --- | --- | --- | --- | ---
0 | E1 | n=3 | 3.93±0.06 | 3.33±0.04 | 1.33±0.03 | 2.01±0.09 | 4.04±0.01 | 0.34±0.01
0 | E2 | n=3 | 3.93±0.06 | 3.33±0.04 | 1.33±0.03 | 2.01±0.09 | 4.04±0.01 | 0.34±0.01
0 | E1 | n=3 | 3.84±0.08 | 3.87±0.02 | 0.64±0.03 | 3.95±0.02 | 0.82±0.02
2 | E2 | n=3 | 3.84±0.08 | 3.87±0.02 | 0.64±0.03 | 3.95±0.02 | 0.82±0.02
5 | E2 | n=3 | 3.68±0.07 | 3.32±0.02 | 0.64±0.03 | 3.95±0.02 | 0.82±0.02

Table S7. Subpopulation-specific analysis of time-resolved fluorescence decays for lipid-bound ApoE. Lifetime components obtained by fitting lifetime decays for donor only subpopulation, donor-acceptor population, and acceptor only population after PIE excitation. For each subpopulation, the lifetime component \(\tau\) and the rotational component \(\tau_{\text{rot}}\) are identified. From the comparison between the lifetime decay \(\tau_{\text{DA}}\) decay and the lifetime of the donor in absence of acceptor, \(\tau_{\text{Only}}\), we estimated the energy transfer characteristic time \(\tau_{\text{FRET}}\). Data are further subdivided based on the corresponding mean transfer efficiency (as identified in Table S4) and error are standard deviations of “n” independent repeats of the measurement.

Table S8. Steady-state and time resolved anisotropies in aqueous buffer conditions. Steady-state anisotropy \(r_{ss}\), \(r_{ss}\) and rotational component \(r_{rot}\) for donor-only, acceptor-only, and acceptor-donor subpopulations (compare Eq. S10 and corresponding section). Errors represent standard deviations from “n” independent repeats of the measurement.
Table S9. Steady-state and time resolved anisotropies for lipid-bound ApoE4. Steady-state anisotropy $r_{ss}$, $r_{\infty}$ and rotational component $\tau_{rot}$ for donor-only, acceptor-only, and acceptor-donor subpopulations (compare Eq. S10 and corresponding section). Errors represent standard deviations from “n” independent repeats of the measurement.

<table>
<thead>
<tr>
<th>Construct</th>
<th>E</th>
<th>$r_{ss}$-Donly</th>
<th>$r_{\infty}$-Donly</th>
<th>$\tau_{rot}$-Donly (ns)</th>
<th>$r_{ss}$-Aonly</th>
<th>$r_{\infty}$-Aonly</th>
<th>$\tau_{rot}$-Aonly (ns)</th>
<th>$r_{ss}$-A(D)</th>
<th>$r_{\infty}$-A(D)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ApoE488, 86</td>
<td>E₁</td>
<td>0.078±0.002</td>
<td>0.10±0.02</td>
<td>0.34±0.02</td>
<td>0.07±0.01</td>
<td>0.053±0.007</td>
<td>0.66±0.02</td>
<td>0.121±0.002</td>
<td>0.14±0.01</td>
</tr>
<tr>
<td>ApoE488, 165</td>
<td>E₁</td>
<td>0.16±0.01</td>
<td>0.149±0.009</td>
<td>0.64±0.09</td>
<td>0.10±0.02</td>
<td>0.11±0.03</td>
<td>0.57±0.02</td>
<td>0.17±0.01</td>
<td>0.24±0.03</td>
</tr>
<tr>
<td>ApoE488, 165</td>
<td>E₂</td>
<td>0.16±0.01</td>
<td>0.149±0.009</td>
<td>0.64±0.09</td>
<td>0.10±0.02</td>
<td>0.11±0.03</td>
<td>0.57±0.02</td>
<td>0.050±0.006</td>
<td>0.051±0.004</td>
</tr>
<tr>
<td>ApoE4182, 241</td>
<td>E₁</td>
<td>0.13±0.009</td>
<td>0.14±0.01</td>
<td>0.47±0.08</td>
<td>0.139±0.006</td>
<td>0.147±0.009</td>
<td>0.71±0.04</td>
<td>0.131±0.003</td>
<td>0.17±0.03</td>
</tr>
<tr>
<td>ApoE4182, 241</td>
<td>E₂</td>
<td>0.13±0.009</td>
<td>0.14±0.01</td>
<td>0.47±0.08</td>
<td>0.139±0.006</td>
<td>0.147±0.009</td>
<td>0.71±0.04</td>
<td>0.083±0.006</td>
<td>0.085±0.004</td>
</tr>
<tr>
<td>ApoE4232, 291</td>
<td>E₁</td>
<td>0.13±0.01</td>
<td>0.15±0.01</td>
<td>0.44±0.04</td>
<td>0.14±0.01</td>
<td>0.16±0.01</td>
<td>0.75±0.04</td>
<td>0.141±0.002</td>
<td>0.208±0.009</td>
</tr>
<tr>
<td>ApoE488, 241</td>
<td>E₁</td>
<td>0.14±0.02</td>
<td>0.14±0.02</td>
<td>0.54±0.06</td>
<td>0.17±0.02</td>
<td>0.20±0.01</td>
<td>0.61±0.04</td>
<td>0.152±0.003</td>
<td>0.22±0.01</td>
</tr>
</tbody>
</table>

Table S10. Estimates of deviations in the orientational $\chi^2$ factor in aqueous buffer conditions. See corresponding section in the Supplementary Information for definition of the different terms.

<table>
<thead>
<tr>
<th>Construct</th>
<th>E</th>
<th>$\chi^2$ min</th>
<th>$\chi^2$ max</th>
<th>$\chi^2$ mean</th>
<th>precision</th>
<th>accuracy</th>
<th>$R_0$,mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>ApoE488, 86</td>
<td>E₂</td>
<td>0.36 (0.35)</td>
<td>1.75 (1.77)</td>
<td>0.79 (0.79)</td>
<td>0.01 (0.01)</td>
<td>0.06 (0.07)</td>
<td>5.55</td>
</tr>
<tr>
<td>ApoE488, 86</td>
<td>E₁</td>
<td>0.36 (0.35)</td>
<td>1.75 (1.77)</td>
<td>0.88 (0.91)</td>
<td>0.03 (0.03)</td>
<td>0.07 (0.07)</td>
<td>5.66</td>
</tr>
<tr>
<td>ApoE488, 165</td>
<td>E₃</td>
<td>0.29 (0.30)</td>
<td>2.17 (2.12)</td>
<td>0.73 (0.74)</td>
<td>0.006 (0.002)</td>
<td>0.07 (0.07)</td>
<td>5.48</td>
</tr>
<tr>
<td>ApoE4182, 241</td>
<td>E₂</td>
<td>0.29 (0.30)</td>
<td>2.22 (2.09)</td>
<td>0.76 (0.76)</td>
<td>0.0002 (0.0002)</td>
<td>0.07 (0.07)</td>
<td>5.72</td>
</tr>
<tr>
<td>ApoE4182, 241</td>
<td>E₁</td>
<td>0.29 (0.30)</td>
<td>2.22 (2.09)</td>
<td>0.77 (0.76)</td>
<td>0.0008 (0.0008)</td>
<td>0.07 (0.07)</td>
<td>5.53</td>
</tr>
<tr>
<td>ApoE4232, 291</td>
<td>E₃</td>
<td>0.24 (0.28)</td>
<td>2.35 (2.13)</td>
<td>0.76 (0.76)</td>
<td>0.005 (0.0006)</td>
<td>0.08 (0.08)</td>
<td>5.52</td>
</tr>
<tr>
<td>ApoE4232, 291</td>
<td>E₂</td>
<td>0.24 (0.28)</td>
<td>2.35 (2.13)</td>
<td>0.75 (0.76)</td>
<td>0.008 (0.001)</td>
<td>0.08 (0.08)</td>
<td>5.51</td>
</tr>
<tr>
<td>ApoE4232, 291</td>
<td>E₁</td>
<td>0.24 (0.28)</td>
<td>2.35 (2.13)</td>
<td>0.98 (0.88)</td>
<td>0.03 (0.02)</td>
<td>0.1 (0.08)</td>
<td>5.76</td>
</tr>
<tr>
<td>ApoE488, 241</td>
<td>E₃</td>
<td>0.25 (0.26)</td>
<td>2.26 (2.23)</td>
<td>0.76 (0.75)</td>
<td>0.004 (0.005)</td>
<td>0.08 (0.08)</td>
<td>5.52</td>
</tr>
<tr>
<td>ApoE488, 241</td>
<td>E₂</td>
<td>0.25 (0.26)</td>
<td>2.26 (2.23)</td>
<td>0.76 (0.75)</td>
<td>0.004 (0.0009)</td>
<td>0.08 (0.08)</td>
<td>5.54</td>
</tr>
</tbody>
</table>

Table S11. Estimates of deviations in the orientational $\chi^2$ factor for lipid-bound ApoE4. See corresponding section in the Supplementary Information for definition of the different terms.

<table>
<thead>
<tr>
<th>Construct</th>
<th>E</th>
<th>$\chi^2$ min</th>
<th>$\chi^2$ max</th>
<th>$\chi^2$ mean</th>
<th>precision</th>
<th>accuracy</th>
<th>$R_0$,mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>ApoE488, 86</td>
<td>E₁</td>
<td>0.37 (0.37)</td>
<td>1.78 (1.66)</td>
<td>0.85 (0.86)</td>
<td>0.03 (0.03)</td>
<td>0.05 (0.06)</td>
<td>5.62</td>
</tr>
<tr>
<td>ApoE488, 165</td>
<td>E₁</td>
<td>0.28 (0.28)</td>
<td>2.23 (2.28)</td>
<td>0.92 (0.95)</td>
<td>0.03 (0.03)</td>
<td>0.07 (0.08)</td>
<td>5.70</td>
</tr>
<tr>
<td>ApoE4182, 241</td>
<td>E₂</td>
<td>0.26 (0.26)</td>
<td>2.23 (2.28)</td>
<td>0.74 (0.74)</td>
<td>0.005 (0.005)</td>
<td>0.07 (0.07)</td>
<td>5.49</td>
</tr>
<tr>
<td>ApoE4182, 241</td>
<td>E₁</td>
<td>0.26 (0.26)</td>
<td>2.16 (2.21)</td>
<td>0.98 (0.92)</td>
<td>0.03 (0.02)</td>
<td>0.09 (0.09)</td>
<td>5.76</td>
</tr>
<tr>
<td>ApoE4232, 291</td>
<td>E₂</td>
<td>0.26 (0.26)</td>
<td>2.22 (2.21)</td>
<td>0.80 (0.79)</td>
<td>0.002 (0.003)</td>
<td>0.08 (0.08)</td>
<td>5.56</td>
</tr>
<tr>
<td>ApoE4232, 291</td>
<td>E₁</td>
<td>0.24 (0.27)</td>
<td>2.30 (2.15)</td>
<td>0.99 (0.99)</td>
<td>0.03 (0.03)</td>
<td>0.09 (0.09)</td>
<td>5.77</td>
</tr>
<tr>
<td>ApoE488, 241</td>
<td>E₁</td>
<td>0.22 (0.24)</td>
<td>2.29 (2.25)</td>
<td>1.01 (0.99)</td>
<td>0.03 (0.03)</td>
<td>0.10 (0.10)</td>
<td>5.79</td>
</tr>
</tbody>
</table>

(*values estimated from steady-state anisotropies.*
Table S12. $\sigma^2$ values (Eq. 18) for ApoE4 under aqueous buffer and lipid-bound conditions. Comparison of $\sigma^2$ computed with Eq. 18 between measured values and estimates from Gaussian distribution. The Gaussian distribution serves as a reference for a completely disordered and highly dynamic ensemble.

<table>
<thead>
<tr>
<th>Construct</th>
<th>E1-Measured</th>
<th>E1-Gaussian</th>
<th>E2-Measured</th>
<th>E2-Gaussian</th>
<th>E3-Measured</th>
<th>E3-Gaussian</th>
</tr>
</thead>
<tbody>
<tr>
<td>ApoE45, 86 lipid-free</td>
<td>n=6</td>
<td>0.02±0.01</td>
<td>0.08±0.02</td>
<td>0.055±0.007</td>
<td>0.107±0.004</td>
<td>-</td>
</tr>
<tr>
<td>DMPC</td>
<td>n=3</td>
<td>0.01±0.01</td>
<td>0.050±0.002</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ApoE486, 165 lipid-free</td>
<td>n=3</td>
<td>0.14±0.03</td>
<td>0.115</td>
<td>0.09±0.01</td>
<td>0.107</td>
<td>0.0010±0.0006</td>
</tr>
<tr>
<td>DMPC</td>
<td>n=6</td>
<td>0.003±0.003</td>
<td>0.017±0.003</td>
<td>0.021±0.001</td>
<td>0.028±0.001</td>
<td>-</td>
</tr>
<tr>
<td>ApoE4182, 241 lipid-free</td>
<td>n=4</td>
<td>0.061±0.002</td>
<td>0.089±0.002</td>
<td>0.022±0.003</td>
<td>0.035±0.007</td>
<td>-</td>
</tr>
<tr>
<td>DMPC</td>
<td>n=4</td>
<td>0</td>
<td>0.039±0.005</td>
<td>0.095±0.005</td>
<td>0.0769</td>
<td>-</td>
</tr>
<tr>
<td>ApoE4223, 291 lipid-free</td>
<td>n=4</td>
<td>0.01±0.01</td>
<td>0.076±0.0003</td>
<td>0.075±0.007</td>
<td>0.104±0.001</td>
<td>0.035±0.004</td>
</tr>
<tr>
<td>DMPC</td>
<td>n=3</td>
<td>0</td>
<td>0.041±0.003</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ApoE486, 241 lipid-free</td>
<td>n=3</td>
<td>0.07±0.01</td>
<td>0.097</td>
<td>0.091±0.003</td>
<td>0.1094±0.0003</td>
<td>0.0255±0.0009</td>
</tr>
<tr>
<td>DMPC</td>
<td>n=3</td>
<td>0</td>
<td>0.025±0.001</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table S13. Fraction of molecules corresponding to acceptor only Aonly, donor-acceptor (DA) and donor only (Donly) for ApoE4 under aqueous buffer and lipid-bound conditions. All errors represent standard deviations based on n independent replicates of the experiments.

<table>
<thead>
<tr>
<th>Construct</th>
<th>Aonly</th>
<th>DA</th>
<th>Donly</th>
</tr>
</thead>
<tbody>
<tr>
<td>ApoE45, 86 lipid-free</td>
<td>n=6</td>
<td>0.23±0.04</td>
<td>0.59±0.08</td>
</tr>
<tr>
<td>DMPC</td>
<td>n=3</td>
<td>0.31±0.02</td>
<td>0.46±0.02</td>
</tr>
<tr>
<td>ApoE486, 165 lipid-free</td>
<td>n=3</td>
<td>0.460±0.009</td>
<td>0.323±0.009</td>
</tr>
<tr>
<td>DMPC</td>
<td>n=6</td>
<td>0.46±0.04</td>
<td>0.30±0.03</td>
</tr>
<tr>
<td>ApoE4182, 241 lipid-free</td>
<td>n=4</td>
<td>0.22±0.01</td>
<td>0.57±0.03</td>
</tr>
<tr>
<td>DMPC</td>
<td>n=4</td>
<td>0.29±0.01</td>
<td>0.50±0.02</td>
</tr>
<tr>
<td>ApoE4223, 291 lipid-free</td>
<td>n=4</td>
<td>0.13±0.02</td>
<td>0.77±0.04</td>
</tr>
<tr>
<td>DMPC</td>
<td>n=3</td>
<td>0.227±0.005</td>
<td>0.62±0.02</td>
</tr>
<tr>
<td>ApoE486, 241 lipid-free</td>
<td>n=3</td>
<td>0.131±0.006</td>
<td>0.79±0.02</td>
</tr>
<tr>
<td>DMPC</td>
<td>n=3</td>
<td>0.20±0.02</td>
<td>0.71±0.03</td>
</tr>
</tbody>
</table>
**Table S14. Analysis of Fluorescence Correlation Spectroscopy under lipid-bound conditions.** Hydrodynamic radius $R_h$ of ApoE4 constructs bound to DMPC liposomes as determined by donor-donor (DD) and acceptor-acceptor (AA) correlation. $N$ reports about the corresponding number of molecules as determined by the amplitude of the correlation. All errors represent standard deviations based on “n” independent replicates of the experiments.

<table>
<thead>
<tr>
<th>Construct</th>
<th>$R_h^{DD}$ (nm)</th>
<th>$R_h^{AA}$ (nm)</th>
<th>$N^{DD}$</th>
<th>$N^{AA}$</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>ApoE45, 86</td>
<td>6.3±0.7</td>
<td>9±1</td>
<td>0.4±0.1</td>
<td>0.13±0.02</td>
<td>n=3</td>
</tr>
<tr>
<td>ApoE466, 165</td>
<td>9±2</td>
<td>10±2</td>
<td>0.25±0.04</td>
<td>0.097±0.003</td>
<td>n=6</td>
</tr>
<tr>
<td>ApoE4182,241</td>
<td>9±1</td>
<td>9.2±0.5</td>
<td>0.3±0.1</td>
<td>0.100±0.006</td>
<td>n=4</td>
</tr>
<tr>
<td>ApoE4223,291</td>
<td>11.5±0.4</td>
<td>14±2</td>
<td>0.27±0.04</td>
<td>0.109±0.009</td>
<td>n=4</td>
</tr>
<tr>
<td>ApoE486, 241</td>
<td>12±1</td>
<td>12±3</td>
<td>0.212±0.009</td>
<td>0.17±0.04</td>
<td>n=3</td>
</tr>
</tbody>
</table>

**Table S15. Residue pairs used for computing distances used in clustering protein conformations.**

REFERENCES


