Primary Structure and Solution Conditions Determine Conformational Ensemble Properties of Intrinsically Disordered Proteins

Albert Hsuan-Han Mao
Washington University in St. Louis

Follow this and additional works at: http://openscholarship.wustl.edu/etd

Recommended Citation
Mao, Albert Hsuan-Han, "Primary Structure and Solution Conditions Determine Conformational Ensemble Properties of Intrinsically Disordered Proteins" (2014). All Theses and Dissertations (ETDs). 1250.
http://openscholarship.wustl.edu/etd/1250

This Dissertation is brought to you for free and open access by Washington University Open Scholarship. It has been accepted for inclusion in All Theses and Dissertations (ETDs) by an authorized administrator of Washington University Open Scholarship. For more information, please contact digital@wumail.wustl.edu.
Primary Structure and Solution Conditions Determine Conformational Ensemble

Properties of Intrinsically Disordered Proteins

by

Albert Hsuan-Han Mao

A dissertation presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2014

St. Louis, Missouri
Contents

List of Figures ......................................................... ix

List of Tables ......................................................... xiii

Acknowledgments ....................................................... xiv

Abstract ................................................................. xvi

1 Intrinsically disordered proteins (IDPs) have distinctive sequence
characteristics and exhibit an ensemble of conformations ........ 1

1.1 The importance of sequence-ensemble relationships ............... 5

1.2 Limiting models from polymer physics as descriptors of conformational
heterogeneity ......................................................... 8

1.3 Toward more realistic polymer models for IDPs ..................... 11

1.4 Aqueous solutions are poor solvents for generic polypeptide backbones 15

1.5 Polar tracts form compact globules in aqueous solutions ........... 17

1.6 IDPs with charged residues constitute a distinct archetype from polar
tracts ................................................................. 18

1.7 References ......................................................... 19

2 Net charge per residue modulates conformational ensembles of in-
trinsically disordered proteins ........................................ 30

2.1 Introduction ......................................................... 31

2.2 Results and discussion ............................................. 35

2.2.1 Characteristics of protamine sequences used in this study ... 35
2.2.2 Net charge per residue segregates protamine conformations along a globule-to-coil transition ........................................ 37
2.2.3 $\langle R_g \rangle$ exhibits minor sensitivity to sequence permutation .... 39
2.2.4 Additional tests support the proposed globule-to-coil transition 41
2.2.5 Explicit solvent simulations do not exhibit salt dependence ... 41
2.2.6 Fluorescence experiments corroborate implicit solvent simulations 44
2.2.7 Increased swelling of charged protamines is realized through local conformational preferences for P$_{II}$ helices ............. 51
2.2.8 Conformational properties of protamines exhibit sequence specificity ................................................................. 54
2.2.9 Classification of different collapsed states ...................... 59
2.2.10 Electrostatic attraction drives collapse of polyampholyte sequences ............................................................... 60
2.2.11 Anion contact numbers are mildly enhanced around basic residues 62
2.2.12 Simulation and experimental results are consistent with theories for polyelectrolytes and polyampholytes in poor solvents ... 63
2.2.13 Toward a sequence-space phase diagram for proteins ......... 67
2.3 Materials and methods .................................................. 68
2.3.1 Sequence selection .................................................. 68
2.3.2 Simulations ............................................................ 70
2.3.3 Experiments ........................................................... 84
2.4 References .................................................................... 87

3 Counterion-mediated collapse of acid-rich intrinsically disordered proteins in simulations depends on salt ion models ........... 97
3.1 Intrinsically disordered regions modeled in isolation ............ 98

iii
3.1.1 Simulation methods .............................................. 99
3.1.2 Disordered regions are typically short and have moderate net
charge per residue .................................................. 100
3.1.3 Most IDRs adopt ensembles of collapsed, globular conformations101
3.1.4 Acidic and basic sidechains exhibit distinct levels of counterion
accumulation ......................................................... 105
3.2 Synthetic polyelectrolyte IDPs ........................................ 110
3.2.1 Polyglutamate and polyaspartate collapse while entraining
counterions in simulations ........................................... 110
3.2.2 Development of a simplified symmetric model for salt ions in
ABSINTH .............................................................. 112
3.2.3 Polyelectrolyte IDPs simulated with simplified salt ions do not
collapse ............................................................... 114
3.3 Discussion ............................................................. 116
3.3.1 Solution conditions modulate the conformational ensembles of
IDPs ................................................................. 116
3.3.2 Can the original salt ion parameters used in ABSINTH be im-
proved? ............................................................... 122
3.4 References ............................................................. 124

4 Crystal lattice properties fully determine short-range interaction pa-
parameters for alkali and halide ions ................................. 128
4.1 Introduction ........................................................... 129
4.2 Methods ............................................................... 131
4.2.1 Calibration targets ............................................... 131
4.2.2 Ion interaction models .......................................... 132
4.2.3 Lattice energy and interionic distance calculations 134
4.2.4 Calibration objective function and parameter constraints 137
4.2.5 Constrained nonlinear numerical optimization 139

4.3 Results 141
4.3.1 Calibrated ion parameters 141
4.3.2 Attained values of calibration observables 142
4.3.3 Lattice structure prediction 144
4.3.4 Hydration free energies 145

4.4 Discussion 147

4.5 Conclusion 150

4.6 References 150

5 Sequence patterning modulates composition in determining conformational ensembles of polyampholytic intrinsically disordered proteins 157

5.1 Simulation methods 158

5.2 Results 160
5.2.1 Both acid-rich and base-rich polyelectrolytes adopt swollen coil ensembles 160
5.2.2 Synthetic glutamine/glutamate IDPs exhibit a globule-to-coil transition with net charge per residue 162
5.2.3 Conformational ensembles of proline-rich and strongly polyampholytic IDRs disagree with phase diagram predictions 164

5.3 Discussion 167

5.4 References 169
6 Summary of contributions and future directions ................. 172

6.1 References ............................................. 176

Appendix A Exact recording of Metropolis-Hastings-class Monte Carlo simulations using one bit per sample ............... 177
A.1 Introduction ............................................ 178
A.2 Description of the method ................................ 180
A.3 Demonstration of the method ................................ 181
A.4 Discussion ............................................. 184
A.5 Conclusion ............................................. 185
A.6 References ............................................. 186

Appendix B Mixing Extent: a quantifier of sequence patterning . 188
B.1 Introduction ............................................ 188
B.2 Definition of the Mixing Extent ......................... 188
B.3 Algorithms for calculating the Mixing Extent .......... 190
B.4 Both folded and disordered proteins exhibit the full range of Mixing Extents ............................................ 196
B.5 Conclusions ............................................. 198
B.6 References ............................................. 199

Appendix C Calculating activity coefficients of ABSINTH electrolytes using grand canonical Monte Carlo simulations .... 200
C.1 Introduction ............................................ 200
C.2 Methods ................................................ 200
C.2.1 Fundamental thermodynamic definitions ............... 201
C.2.2 Definition of the activity coefficient and standard states ... 202
C.2.3 Relationship between whole salt and individual ion quantities 203
C.2.4 Conversion between concentration scales 205
C.2.5 Searching for target concentrations using inverse grand canonical Monte Carlo simulations 206
C.2.6 Formulation of grand canonical Monte Carlo for whole salt insertion and deletion 209
C.2.7 Ensemble sampling and analysis methods 213
C.3 Results 215
C.3.1 Simulations in ABSINTH overestimate the chemical potential of electrolytes at high concentrations 215
C.3.2 ABSINTH solvation parameters can be adjusted to improve agreement with experimental measurements 217
C.3.3 Decreasing the effective radius of solutes improves agreement with experiments 219
C.4 Conclusions 220
C.5 References 222

Appendix D Contributions to other scientific works 225
D.1 Forces for restraining the radius of gyration during molecular dynamics simulations 225
D.2 A reflection move for Monte Carlo simulations that connects kinetically distant proline pucker states 226
D.3 Simulations of the C-terminal tail of single-stranded DNA binding protein (SSB C-t) from Escherichia coli and its fragments 227
D.4 An algorithm for performing inverse Maxwell constructions 228
D.5 Strategy for elucidating the role of sequence contexts in modulating functional motifs within intrinsically disordered regions 229
D.6 References 229
List of Figures

2.1 Inventory of protamine sequences ........................................ 36
2.2 Normalized \langle R_g \rangle plotted against net charge per residue .......... 38
2.3 Inventory of permuted sequences using the composition of protamine 4 ........ 39
2.4 Comparison of ensemble average radii of gyration for the globule and coil reference states, protamine 4, and its permutants ....................... 40
2.5 Asphericity \delta^* plotted against net charge per residue .................... 42
2.6 Scaling of the ensemble average internal distances, \langle R_{ij} \rangle, between residues i and j plotted against chain separation, |j - i| ..................... 43
2.7 Comparison of mass-weighted radius of gyration histograms between different salt concentrations and solvation models ......................... 44
2.8 Comparison of translational diffusion coefficients obtained from analysis of simulation results and FCS experiments ......................... 46
2.9 Correlation between measured and calculated translational diffusion coefficients .............................................................. 47
2.10 Steady-state fluorescence anisotropy r plotted against the net charge per residue .............................................................. 48
2.11 Correlation between fluorescence anisotropy and radius of gyration .... 49
2.12 Correlation between fluorescence anisotropy and asphericity .......... 50
2.13 Calculated Kratky profiles for all protamines ............................. 51
2.14 Ensemble average values for P_{II} propensity plotted against net charge per residue .............................................................. 52
2.15 Plot of \alpha-helical propensity against net charge per residue for the protamines .............................................................. 53
<table>
<thead>
<tr>
<th>Section Number</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.16</td>
<td>Plot of β-strand propensity against net charge per residue for the protamines</td>
<td>54</td>
</tr>
<tr>
<td>2.17</td>
<td>Selected ensemble average contact maps</td>
<td>56</td>
</tr>
<tr>
<td>2.18</td>
<td>Ensemble average residue contact maps for protamines 1–12</td>
<td>57</td>
</tr>
<tr>
<td>2.19</td>
<td>Ensemble average residue contact maps for protamines 12–21</td>
<td>58</td>
</tr>
<tr>
<td>2.20</td>
<td>Inventory of additional simulated sequences with low net charge per residue and hydrophobicity</td>
<td>60</td>
</tr>
<tr>
<td>2.21</td>
<td>Normalized $\langle R_g \rangle$ plotted against net charge per residue of additional polyampholyte sequences</td>
<td>61</td>
</tr>
<tr>
<td>2.22</td>
<td>$\delta^*$ plotted against net charge per residue for the additional polyampholyte sequences</td>
<td>62</td>
</tr>
<tr>
<td>2.23</td>
<td>Scaling of the ensemble average internal distances, $\langle R_{ij} \rangle$ plotted against chain separation, $</td>
<td>j - i</td>
</tr>
<tr>
<td>2.24</td>
<td>Combined pairwise distance histograms for sidechain tips of polyampholytic sequences</td>
<td>64</td>
</tr>
<tr>
<td>2.25</td>
<td>Combined pairwise distance histograms for sidechain tips of polyampholytic sequences simulated with a Lennard-Jones reference Hamiltonian</td>
<td>65</td>
</tr>
<tr>
<td>2.26</td>
<td>Illustrative structures from the ensembles of polyampholytic globules</td>
<td>66</td>
</tr>
<tr>
<td>2.27</td>
<td>Residue-resolved ensemble average chloride ion contact numbers</td>
<td>67</td>
</tr>
<tr>
<td>2.28</td>
<td>Proposed schematic phase diagram for the single chain phase behavior of unbound, single domain proteins</td>
<td>69</td>
</tr>
<tr>
<td>2.29</td>
<td>The decision-tree used by the simulation software to select a Monte Carlo move at each step</td>
<td>75</td>
</tr>
</tbody>
</table>
2.30 Listing of the aligned backbone atom RMSD between pairs of starting conformations .......................................................... 78
2.31 Operational definitions for local structure designations used in this work 80

3.1 Length histogram of intrinsically disordered regions from DisProt .... 101
3.2 Phase diagram annotated with IDRs from DisProt ..................... 102
3.3 Normalized radius of gyration versus net charge per residue for IDRs from DisProt .......................................................... 104
3.4 Example conformation showing collapse of PEEKKEEGSANRPEDELES LSA with associated K$^+$ ions .................................................. 107
3.5 Pooled ion-sidechain distance histograms for all IDRs from DisProt . 108
3.6 Individual ion-sidechain pair correlation functions for two example IDRs 109
3.7 Example conformation showing collapse of (Glu)$_{34}$ with entrained counterions ............................................................... 111
3.8 Example conformation showing (Arg)$_{34}$ with associated counterions . 113
3.9 Reference hydration free energy as a function of diameter for ABSINTH RPM ions .............................................................. 115
3.10 Radius of gyration for acidic synthetic polyelectrolytes in ARPM ion solution ............................................................... 117
3.11 Radius of gyration for basic synthetic polyelectrolytes in ARPM ion solution ............................................................... 118
3.12 Ion-sidechain distance histograms for (Asp)$_{17}$ .......................... 119
3.13 Ion-sidechain distance histograms for (Glu)$_{17}$ .......................... 120

5.1 Normalized $\langle R_g \rangle$ for (Asp)$_{50}$, (Glu)$_{50}$, (Lys)$_{50}$, and (Arg)$_{50}$ ........... 161
5.2 Inventory of synthetic Q/E IDP sequences .................................. 162
5.3 Normalized \( \langle R_g \rangle \) as a function of net charge per residue for synthetic Q/E IDPs .................................................. 164

5.4 Inventory of naturally occurring acid-rich IDPs and IDRs .......................... 165

5.5 Normalized \( \langle R_g \rangle \) as a function of net charge per residue for naturally occurring acid-rich IDPs and IDRs ........................................ 166

5.6 Example expanded conformation of \textit{WPPDRGHDKSDRDGERGYDKVDREERDR}
\textit{DRGERDGDREKERRHHHHRREEC} with associated salt ions ............ 168

B.1 Bondarenko’s algorithm for calculating Mixing Extents ........................ 195

B.2 Normalized histograms of Mixing Extents across folded and disordered proteomes ........................................................... 197

B.3 Normalized histograms of reduced alphabet Mixing Extents across folded and disordered proteomes ............................................ 198

C.1 Comparison of simulated and experimental activity coefficients for NaCl and KCl as a function of molarity ............................... 216

C.2 Activity coefficients for NaCl and KCl as a function of molarity from experiments and simulations with \( \tau_d = 0.059 \) and \( \chi_d = 0.6 \) .......... 218

C.3 Activity coefficients for NaCl and KCl as a function of molarity from experiments and simulations using a constant dielectric model .... 221
List of Tables

4.1 Lattice energy and interionic distance measurements used as calibration targets ........................................... 133
4.2 Short-range interaction parameters derived from crystal lattice properties ...................................................... 142
4.3 Lattice energies and interionic distances attained using lattice-derived primitive model parameters ............... 143
4.4 Lattice energies and interionic distances attained using lattice-derived Lennard-Jones parameters with Lorentz-Berthelot mixing rules ................................................................. 144
4.5 Hydration free energies attained using lattice-derived Lennard-Jones parameters and three water models .... 146

A.1 Comparison of efficiencies for different recording methods .............................................................................. 183
A.2 Comparison of total run times for generating $10^6$ samples of the demonstration system ............................... 183
B.1 Permutations of the string $\text{CABBAB}$ organized by number of runs ............................................................. 189
Acknowledgments

It is often difficult to predict or control the course of events and relationships that influence a student’s training. In light of this uncertainty, I am constantly aware of my great fortune in being surrounded by colleagues who make science enjoyable, social, and enriching. Most significantly, I acknowledge my mentor Rohit Pappu for his tireless commitment to his trainees’ development as scientists and human beings. Rohit’s unquenchable curiosity and ability to forge connections between scientific disciplines as well as the humans who work in them have been truly inspiring. His collaborative and communicative personality fosters a laboratory culture of mutual respect and intellectual stimulation.

This culture has been strengthened by the postdocs and fellow students, from alumnus to neophyte, whom I am proud to call labmates. Hoang Tran, Xiaoling Wang, Andreas Vitalis, Alan Chen, Scott Crick, Matt Wyczalkowski, Nick Lyle, Tim Williamson, Emma Morrison, Rahul Das, Adam Steffen, Aditya Radhakrishnan, Ronak Patel, Kiersten Ruff, Anuradha Mittal, Siddique Khan, and Tyler Harmon have all shaped me through their camaraderie and insight. In particular, I acknowledge Alan and Andreas for their long hours spent coaching me during my novice days, Nick for his many heroic acts of system administration, and Scott for his willingness to engage with computationalists. I am also humbled by how much I learned while mentoring Amanda Spencer, Kelly Culhane, and Tony Wang during their undergraduate research programs.

I thank my past and present thesis committee members Nathan Baker, Donald Elbert, Elliot Elson, Roberto Galletto, James Havranek, and Timothy Lohman for their insights and guidance. I am also grateful to the National Institutes of Health (NIGMS 5T32GM008802), the National Science Foundation (MCB 0718924 and MCB

xiv
1121867), the Medical Scientist Training Program at Washington University, and by extension the American taxpayer, for funding my training in a time of national economic hardship.

This dissertation is dedicated to my parents Su-Chen and Yu-Hua. It is a direct outcome of the love for learning and discovery they instilled in me from my earliest days. Their steady infusion of support and encouragement has been a sanctuary during throughout my training. Together with my parents-in-law Chian Fen and Shing Chung and sister-in-law Stephanie Hsu, our shared experiences of warm companionship, beautiful journeys, and delicious food have given me memories to last a lifetime.

Finally, I acknowledge my dear wife Kimberly for being my true love and best friend. Being with you each day gives me perspective and happiness which make responsibilities lighter and challenges easier to overcome. I will always look back upon my years as a student with joy because it was time we spent together.

Albert Hsuan-Han Mao

Washington University in St. Louis

May 2014
Intrinsically disordered proteins (IDPs) are a class of proteins that do not exhibit well-defined three-dimensional structures. The absence of structure is intrinsic to their amino acid sequences, which are characterized by low hydrophobicity and high net charge per residue compared to folded proteins. Contradicting the classic structure-function paradigm, IDPs are capable of interacting with high specificity and affinity, often acquiring order in complex with protein and nucleic acid binding partners. This phenomenon is evident during cellular activities involving IDPs, which include transcriptional and translational regulation, cell cycle control, signal transduction, molecular assembly, and molecular recognition. Although approximately 30% of eukaryotic proteomes are intrinsically disordered, the nature of IDP conformational ensembles remains unclear. In this dissertation, we describe relationships connecting characteristics of IDP conformational ensembles to their primary structures and solution conditions.

Using molecular simulations and fluorescence experiments on a set of base-rich IDPs, we find that net charge per residue segregates conformational ensembles along a globule-to-coil transition. Speculatively generalizing this result, we propose a phase
diagram that predicts an IDP’s average size and shape based on sequence composition and use it to generate hypotheses for a broad set of intrinsically disordered regions (IDRs). Simulations reveal that acid-rich IDRs, unlike their oppositely charged base-rich counterparts, exhibit disordered globular ensembles despite intra-chain repulsive electrostatic interactions. This apparent asymmetry is sensitive to simulation parameters for representing alkali and halide salt ions, suggesting that solution conditions modulate IDP conformational ensembles. We refine the ion parameters using a calibration procedure that relies exclusively on crystal lattice properties. Simulations with these parameters recover swollen coil behavior for acid-rich IDRs, but also uncover a dependence on sequence patterning for polyampholytic IDPs.

These contributions initiate an endeavor to elucidate general principles that enable prediction of an IDP’s conformational ensemble based on primary structure and solution conditions, a goal analogous to structure prediction for folded proteins. Such principles would provide a molecular basis for understanding the roles of IDPs in physiology and pathophysiology, guide development of agents that modulate their behavior, and enable their rational design from chosen specifications.
Chapter 1

Intrinsically disordered proteins (IDPs) have distinctive sequence characteristics and exhibit an ensemble of conformations

This chapter is adapted from a review article in preparation for submission to Biochemical Journal. The candidate, Albert H. Mao, co-authored the article along with Nicholas Lyle and Rohit V. Pappu.

In 1894, Emil Fischer discovered that enzymes distinguish stereoisomers in their fermentation of glucosides [1–4]. He had already demonstrated the accuracy of van’t Hoff and Le Bel’s then-controversial tetrahedral carbon model [5, 6] by using it to deduce the stereochemistry of the carbohydrates [7, 8]. Building on the three-dimensional basis for molecular structure provided by this model, Fischer realized that enzyme specificity could be explained by shape complementarity. He proposed the metaphor of a lock and key [2] to illustrate how the three-dimensional arrangements of atoms comprising an enzyme and its substrate could enable them to fit together and prevent non-specific catalysis. Interpreted liberally and generally, this metaphor implied that each protein possesses a rigid structure that determines its function. It became a scientific meme and spread throughout chemistry, biology, and medicine [9, 10] at a time when physicists were only beginning to explore the implications of X-ray diffraction, the quantum hypothesis, and the existence of atomic nuclei. Therefore,
the protein structure-function paradigm not only emerged over half a century before
the first protein structure was elucidated, it preceded all innovations that would make
structure determination even possible.

The assumption that proteins possess one rigid structure was challenged by
biochemical evidence. Protein rigidity was unsatisfactory at explaining noncompeti-
tive inhibition and could not account for enzymes where binding of one reactive group
increased the exposure of another [11]. Fischer himself did not advocate this assump-
tion and felt that popular interpretations of his lock-and-key metaphor exceeded its
scope and experimental justification [12]. In fact, the operation of a mechanical lock is
dependent on the ability of its constituent parts to move relative to one another [13].
The concepts of allosteric linkage [14] and induced fit [15] invoked the ability of a
protein’s conformation to change in response to the approach and binding of an in-
teraction partner. Rather than overthrowing the structure-function paradigm, these
ideas augmented it by introducing the nuance that structures could change during
execution of a function. They remained consistent with Anfinsen’s thermodynamic
hypothesis [16], which precisely states that a protein’s structure represents the global
free energy minimum of an entire physical system [17], since a system is characterized
by solution conditions such as ligand concentration in addition to protein amino acid
sequences.

Advancements on all scientific and technological fronts during the ensuing
decades falsified the assumption of protein rigidity and revealed the ubiquity of confor-
mational fluctuations. The advent of protein crystallography directly confirmed that
proteins could adopt distinct conformations in complex with different ligands [18, 19].
Even in the absence of ligands, many multi-domain proteins exhibit multiple crystal
structures where the domains have similar folds but are sheared [20] or rotated [21].
relative to one another, demonstrating that conformational multiformity can be intrinsic to the protein. A panoply of methods including fluorescence anisotropy [22], Raman spectroscopy [23], fluorescence quenching [24], hydrogen exchange [25], kinetic analysis [26], and patch-clamp current recording [27] provided evidence that fluctuations in structure exist across a broad range of length and time scales and are functionally relevant. Early biological applications of nuclear magnetic resonance spectroscopy [28, 29] and molecular dynamics simulations [30] showed that the atoms of even small, stable, single-domain proteins are in incessant fluctuation. Despite these findings, the proliferation of solved protein structures ushered in a “golden age of structural biology” [31]. Conformational fluctuations could be understood as perturbations about a structurally-defined minima of the energy landscape, or as transitions between such minima, and were therefore secondary to structures in importance. The visual clarity, mathematical simplicity, and high information content of static high-resolution structures made them the standard of knowledge for understanding a protein and its function.

Despite the apparent importance and explanatory power of protein structures, there exist biologically functional proteins that do not exhibit well-defined three-dimensional structure under native physiological conditions. Instead, they adopt an ensemble of conformations for which no single structure or finite collection of structures is representative. Conformational heterogeneity is a central and defining characteristic of these proteins, which transcend the structure-function paradigm. This class of proteins has engendered a variety of names, which reflects the diversity of opinion regarding the proper way to think about proteins when the mental picture of a static, rigid structure is absent. The proliferation of nomenclature can be summarized by the combination of an adverb such as “intrinsically”, “natively”, or “inherently” with a
succeeding adjective such as “disordered”, “unstructured”, “unfolded”, “denatured”, or “flexible” to indicate that the lack of folded structure is not caused by an external perturbation. For consistency and simplicity, this dissertation uniformly adopts the choice of “intrinsically disordered protein” or IDP to refer to this class; the adjective “disordered” refers to the lack of ordered structure and not to disease or pathology. We define an IDP to be a protein whose amino acid sequence encodes a preference for heterogeneous ensembles of conformations as the thermodynamic ground state under standard physiological conditions (aqueous solutions, 150 mM monovalent salt, low concentrations of divalent ions, pH 7.4, and temperature in the range of 298–310 K) [32, 33].

Although evidence for the existence of IDPs has accumulated for nearly as long as protein structures have been available, relatively little is known about them. Glucagon and fetuin were discovered to be examples of IDPs as early as 1959 [34], just one year after publication of the first protein crystal structure [35]. In addition to recognition of the prevalence of conformational fluctuations, many small polypeptide hormones were known to be disordered under certain solution conditions [36], and proteins which recognize specific nucleic acid sequences were occasionally observed to fold only upon binding [37]. The obscurity of IDPs compared to folded proteins is due to the difficulty of characterizing biological polymers with no stable structure as well as the dominance of structural biology. IDPs are not only unresolvable in crystal structures by definition, they are difficult to even isolate from tissue or cell systems because the process of homogenization exposes them to proteases [38] which rapidly and preferentially degrade disordered proteins [39]. Only with a synergistic confluence of methodological advances in the 1990s including the maturation of genetic engineering and bioinformatics, the exponential growth of computational power,
and the availability of genomic databases have IDPs been subject to systematic and detailed characterization. Therefore, at the time of this writing in 2012, intrinsically disordered proteins are an adolescent, relatively unexplored, and rapidly evolving subject of inquiry.

The existence of IDPs raises fundamental physical and biological questions about the nature and interactions of proteins. What intrinsic features of a protein cause it to be disordered instead of folded? What is the nature of conformational ensembles, and how do they differ between IDPs? How can specificity in interactions and biological function be achieved in the context of intrinsic disorder? What features of conformational ensembles are relevant to function, and how should the structure-function paradigm be revised? Answers to these questions would provide a molecular basis for understanding the roles of IDPs in physiology and pathophysiology, guide development of agents that modulate their behavior, and enable their rational design from chosen specifications. As part of a broad scientific effort to elucidate these answers, this dissertation describes results from an ongoing search for general principles that connect IDP conformational ensemble properties to their amino acid sequence.

1.1 The importance of sequence-ensemble relationships

Most IDPs are actually disordered regions that fail to fold autonomously to specific three-dimensional structures. Many though not all disordered regions can adopt singular ordered structures in specific bound complexes [37, 40]. The intrinsic heterogeneity in their unbound forms is reflected in their ability to adopt different folds in the context of different complexes [41]. Transcription factors represent striking examples of molecules that undergo disorder-to-order transitions as they bind to their cognate DNA partners [42–46]. Disorder in the unbound forms is proposed to
be important for lowering the overall affinity, which in turn increases the off-rates of protein-DNA complexes [47]. Interestingly, there is growing evidence for “fuzzy complexes” whereby conformational heterogeneity prevails in binary and multimolecular complexes [48, 49]. IDPs can also self-assemble to form ordered, supramolecular assemblies, although the degree of order within these assemblies is variable. Intermediates that seem to be obligatory for self-assembly are characterized by significant conformational heterogeneity that can be modulated to alter the mechanisms of self-assembly and the stabilities of supramolecular structures [50–57].

Sequence-structure relationships are well documented for proteins whose individual amino acid sequences fold autonomously into specific three-dimensional structures [58, 59]. Specificity for a well-defined fold is the result of information encoded in the amino acid sequence. In direct analogy, information encoded at the sequence level keeps IDPs from folding autonomously into singular, well-defined three-dimensional structures [60, 61]. The information content of IDP sequences is such that acquisition of a folded conformation, if present, is deferred by coupling folding to either binding or self-assembly provided that the interactions in trans can stabilize the IDP in a specific fold. From a thermodynamic standpoint, the stabilities of complexes and mechanisms of binding/assembly are linked to the conformational properties IDPs in their unbound forms. Hence, sequence-ensemble relationships are central to understanding how disorder is used in IDP function.

Quantitative descriptions regarding sequence-ensemble relationships require biophysical characterization. IDPs present challenges for characterization by traditional biophysical and spectroscopic methods. The signals are often highly averaged and by definition these systems resist crystallization unless they can be forced into specific folded structures. IDPs also present biochemical challenges because they can
be difficult to isolate from tissue or cell systems because the process of homogenization exposes them to proteases that rapidly and preferentially degrade disordered proteins. Efforts to characterize and quantify conformational heterogeneity and understand its role in protein function have gained prominence over the past 10–15 years. These efforts have required a systematic integration of biophysical, biochemical, and bioinformatics methods. Here, we focus on advances made in describing sequence-ensemble relationships in terms of quantitative connections between IDP sequence characteristics and their coarse grain conformational descriptors such as average shapes, sizes, amplitudes of conformational fluctuations, and time scales associated with these fluctuations.

Quantitative characterization of conformational heterogeneity requires an appropriate descriptive framework as well as biophysical data. The major biophysical methodologies include nuclear magnetic resonance (NMR) spectroscopy [62–66], steady state and time-resolved fluorescence spectroscopies [67–73], electron paramagnetic resonance spectroscopy (EPR), small angle X-ray scattering (SAXS) [74–78], and molecular simulations that are used either de novo [79–84] or in synergy with data collected from spectroscopic investigations of IDPs [76, 77, 85–95]. Methodological advances are maturing and evolving to enable comparative assessments of sequence-ensemble relationships for IDPs. Our focus here is on a set of polymer physics concepts [96] that provide a unifying framework for quantitative analysis and concise descriptions of sequence-ensemble relationships of IDPs. This framework is useful for analyzing and interpreting data obtained either from experiments or from molecular simulations. The latter affords voluminous information that, in the limit of extensive sampling and accurate energy functions, allows for detailed characterization and classification of conformational heterogeneity.
1.2 Limiting models from polymer physics as descriptors of conformational heterogeneity

A single set of position coordinates (and uncertainties in these coordinates) helps relate sequence and structure for a protein that folds autonomously into a distinct three-dimensional structure. Such coordinate sets are generated as models that fit either the electron density data from X-ray diffraction through ordered protein crystals or NMR data that report on the chemical environments of backbone and sidechain atomic nuclei in solution. The Protein Data Bank [97] provides a comprehensive archive of coordinate sets for a range of crystallizable proteins and proteins that are amenable for structure determination by NMR. This rich data set has lead to systematic classification of folds and fold families thus yielding an improved understanding of sequence-structure relationships and insights regarding the evolution of protein folds. IDPs are not amenable to descriptions by a single or even a small number of distinct coordinate sets. Instead, statistical descriptors are required to provide a concise classification of conformational ensembles. These descriptors form the language of polymer physics.

The two most popular statistical descriptions based on polymer physics are the Flory random coil and worm-like chain models. In the rotational isomeric approximation to the Flory random coil model, the conformational partition function for the polypeptide is written as a product of individual residue partition functions. This is feasible because all interactions between non-nearest neighbor residues are explicitly ignored while the intrinsic conformational preferences of individual residues are captured in terms of weights for each of the possible rotational isomers. Each conformation for residue $x$ is annotated by an intrinsic energy value that is calculated using an empirical potential function of one’s choosing. The conformations are binned
into rotational isomeric states based on the similarities of the backbone and sidechain
dihedral angles. Residue $x$ might have $m$ rotational isomers whereas residue $y$ might
have $n$ rotational isomers. For a given residue, each rotational isomer is assigned a
weight that is proportional to the Boltzmann factor computed from its conformational
energy. Given an amino acid sequence of $N$ residues, one can calculate, \textit{a priori}, the
probabilities associated with all combinations of rotational isomers. For the sequence
of interest the number of rotational isomers per residue, their statistical weights, and
the sequence composition dictates the total number of conformational possibilities
and the likelihoods associated with each conformation. These likelihoods make up
the predicted conformational distribution function and can be used to calculate a
variety of conformational properties including the average end-to-end distance, the
average radius of gyration, the average hydrodynamic size, the average distance be-
tween residues $i$ and $j$, and any observable that can be cast as a function of a moment
of the conformational distribution function. The Flory random coil model is often
used to calibrate measured observables such as NMR chemical shifts, NMR para-
magnetic relaxation enhancement effects, EPR measurements of spin-spin distances
and spin label mobilities, and structure factors from small angle X-ray scattering
measurements.

The advantage of the Flory random coil is its inherent simplicity and the as-
sumption of conformational independence of residues that makes statistical descrip-
tions analytically tractable. Of course, the assumptions underlying the model are
rather severe, but the model itself is useful because observables can be calibrated
as deviations from the Flory random coil. This helps quantify the contributions
of spatial interactions between residues that are distal in the linear sequence and
sequence-ensemble relationships can be classified in terms of deviations of confor-
mational properties from the Flory random coil model. This approach is decidedly one-sided because deviations from the Flory random coil tell us what an ensemble is not and this has limited value for achieving a comprehensive understanding of sequence-ensemble relationships for IDPs.

An alternative approach is to analyze experimental data, specifically data from fluorescence or force spectroscopy that are functions of end-to-end distances using variants of the worm-like chain model. This model assumes that the polymer of interest can be described as a continuously deformable entity. The persistence length $l_p$ is the length scale over which local orientations of the chain are correlated. Fluctuations from rod-like behavior are minimal for spatial separations that are smaller than $l_p$ and the stiff segments become uncorrelated and orient randomly about each other for separations longer than $l_p$. Worm-like chain models afford simple analytical expressions that can be fit to data. Estimates of $l_p$ values for different sequences, studied under similar solution conditions, lead to comparative assessments of sequence-ensemble relationships by quantifying how the average lengths of stiff segments change with sequence. The worm-like chain model allows a continuous interpolation between the maximally heterogeneous Flory random coil model and the uniformly stiff extended rod-like model, which is achieved through continuous changes to $l_p$ values. This class of models is appealing for its simplicity and ease of use and is commonly used to interpret experimental data for IDPs and denatured proteins.

The two limiting models discussed above are analogous to limiting models or laws in other branches of physics. These include the ideal gas law for dilute gases, the Debye-Hückel equation for calculating activity coefficients of electrolytes, and the Hildebrand/Flory-Huggins expressions for the free energies of ideal mixtures. Limiting laws or models provide a route for interpreting experimental data as deviations
from ideal behavior. This helps in quantifying the magnitudes of non-idealities such as the strengths of non-local interactions, which can lead to an improved understanding of sequence-ensemble relationships when combined with additional experiments.

1.3 Toward more realistic polymer models for IDPs

Despite their simplicity, the Flory random coil and worm-like chain models are valid only if the effects of interactions that are explicitly ignored are sufficiently weak. Sequences of IDPs are deficient in hydrophobic residues and enriched in polar and charged amino acids. Electrostatic and polar interactions are typically quite large, even when screened by the surrounding solvent. This raises serious questions regarding the applicability of limiting models for describing IDP conformations. With a few exceptions such as proline- or glycine-rich sequences, the intrinsic flexibilities of all polypeptides are roughly equivalent. This implies that the value of \( l_p \) is essentially fixed for a wide range of sequences, even if it is erroneously treated as a free parameter when fitting experimental data.

The conformational statistics are dictated by the interplay between chain-solvent and intrachain (intra-backbone, backbone-sidechain, and sidechain-sidechain) interactions. Quantities such as the average radius of gyration \( \langle R_g \rangle \), the average hydrodynamic radius \( \langle R_h \rangle \), and the average end-to-end distance \( \langle R_{ee} \rangle \), are different measures of chain size that can be used to respectively quantify the average density, intrinsic viscosity, and concentration of one end of the chain around the other. In addition to measures of chain size, one can also calculate the average shapes of polymers. The average asphericity \( \delta^* \) quantifies the extent of deviation from a perfect sphere \( (\delta^* = 0) \). For ellipsoids, \( \delta^* \approx 0.4 \), and this quantity attains its maximum value of 1 for a perfect rod. This quantity is calculated from the ensemble average
eigenvalues of the gyration tensor.

One can also calculate the average distances between residues $i$ and $j$. The quantity $\langle R_{ij} \rangle$ represents the ensemble average of spatial separations calculated as averages over all residue pairs with sequence separations of $|j - i|$. It is worth noting that multiple pairs of residues $i$ and $j$ will have similar sequence separations $|j - i|$. The profile of $\langle R_{ij} \rangle$ plotted against sequence separation $|j - i|$ quantifies the local concentration of chain segments around each other and provides the most detailed information regarding the so-called link density, which is a formal order parameter that underlies modern formalisms of polymer theories that are based on the Lifshitz approach. These theories explicitly consider the balance between chain-chain and chain-solvent interactions and can be combined with assessments of solvent quality to analyze experimental data and simulation results. In addition to ensemble averages, one can also calculate the one- and two-parameter distribution functions such as $P(R_g), P(Ree), P(Rh), P(\delta), P(R_{ij}, |j - i|)$, and $P(R_g, \delta)$. The latter quantifies the joint distribution of sizes and shapes. Importantly, all of the quantities listed above are accessible to the appropriate combination of experiments and can be calculated using coordinates for simulated ensembles. This enables quantitative comparisons between simulation results and experiments thus facilitating both testing of simulation predictions and incorporation of experimental data as restraints within simulations. Both approaches are important and have enabled the development of quantitative sequence-ensemble relationships for IDPs.

The balance between chain-solvent interactions and intra-chain interactions can be quantified using a parameter $v_{ex}$ that quantifies the volume excluded by individual residues, on average, for favorable interactions with the surrounding solvent. This parameter provides a measure of the strengths of pairwise inter-residue interac-
tions, on average, and can be estimated using light scattering to measure the second virial coefficient for a given sequence and solution condition.

In a good solvent, $v_{\text{ex}} > 0$, and the chain expands to maximize the polymer-solvent interface. As a result, quantities such as the average radius of gyration $\langle R_g \rangle$, the average hydrodynamic radius $\langle R_h \rangle$, and the average end-to-end distance $\langle R_{\text{ee}} \rangle$ scale as $N^{0.59}$ with the chain length $N$ while $\langle R_{ij} \rangle$ scales as $|j - i|^{0.59}$ with the sequence separation $|j - i|$. Aqueous solutions with high concentrations (8 M) of urea are presumed to be reasonable mimics of good solvents for generic polypeptides because urea, a carbonyl diamide, is chemically equivalent to polypeptide backbone amides. Expanded unfolded states are sampled \textit{in vitro} in high concentrations of chemical denaturants such as urea and guanidinium chloride. Consequently, the sizes of these chemically denatured proteins as quantified using $\langle R_h \rangle$ or $\langle R_g \rangle$ scale as $N^{0.59}$. This scaling results because proteins expand to make favorable contacts with the surrounding solvent.

In good solvents, the inter-residue pair interaction coefficient and intra-chain interactions are repulsive on average. The sizes of self-avoiding random walks also scale as $N^{0.59}$. Conformational ensembles for polymers in good solvents and self-avoiding random walks belong to the same “universality class”. Accordingly, ensembles generated in atomistic detail for proteins in the excluded volume (EV) limit are useful reference states for expanded unfolded states. In the EV limit, ensembles are generated using atomistic descriptions of proteins and all non-bonded interactions besides steric repulsions are ignored. Increased accuracy of residue-specific and local context-dependent dihedral angle preferences is obtained by including statistical potentials based on coil libraries to augment the steric only description for non-bonded interactions. Although these approaches provide a more realistic description of con-
formational heterogeneity, they cannot be used unquestioningly for IDPs because this imposes, a priori, the assumption that aqueous solutions are good solvents for IDPs.

If the effects of chain-solvent and intra-chain interactions exactly counterbalance, then $v_{ex} = 0$ and the chain is said to be in a theta solvent. Under such conditions, the chain statistics are consistent with those of a Flory random coil model. It is important to note that this behavior comes about due to counterbalancing of interactions rather than explicit ignoring of non-local interactions. In a theta solvent, $\langle R_g \rangle$, $\langle R_h \rangle$, and $\langle R_{ee} \rangle$ scale as $N^{0.5}$ and $\langle R_{ij} \rangle$ scales as $|j - i|^{0.5}$. The parameter $v_{ex}$ can change continuously going from positive values in a good solvent, through zero in a theta solvent, to negative values in a poor solvent.

In a poor solvent, $v_{ex} < 0$ and the chain prefers compact, globular conformations that efficiently minimize the polymer-solvent interface and $\langle R_g \rangle$ and $\langle R_h \rangle$ scale as $N^{0.33}$. The poorer the solvent, the more negative the value of $v_{ex}$. Statistics of inter-residue distances change fundamentally in a poor solvent. Instead of following power laws, the distances $\langle R_{ij} \rangle$ plateau to a fixed value dictated by the average density of the globule for all values of $|j - i|$ that are larger than a so-called blob length. For typical polypeptide sequences the blob length is between 5 and 7 residues. As $v_{ex}$ approaches zero, the plateauing behavior of the $\langle R_{ij} \rangle$ profile changes continuously toward the appropriate power law behavior depending on the value of $v_{ex}$.

Clearly, the consideration of the details of the balance between chain-chain and chain-solvent interactions affords a richer description of conformational statistics. Which of these models apply for describing IDPs? To answer this question, we need a systematic approach that first asks if typical physiological milieus are good, theta, or poor solvents for polypeptide backbones. This knowledge allows us to understand how sidechains in IDP sequences modulate the intrinsic backbone preferences.
Recent studies of archetypal IDPs using a combination of spectroscopic experiments and molecular simulations have yielded clear insights regarding sequence-ensemble relationships. The following sections provide a concise description of these findings and the implications for predicting sequence-ensemble relationships for IDP sequences.

1.4 Aqueous solutions are poor solvents for generic polypeptide backbones

The free energy of hydration for N-methylacetamide (NMA) at 298 K is $-10$ kcal/mol, indicating that the transfer of NMA from the gas phase into water is highly favorable. Naive extrapolation from the transfer model suggests that polyglycine — a poly-secondary-amide — should prefer structures that maximize the interface with the aqueous solvent. However, results based on molecular dynamics simulations show that polyglycine forms a heterogeneous ensemble of compact globules [98]. Similar simulations showed that polyglycine samples expanded coil-like structures in 8 M urea. There is minimal overlap between conformational ensembles sampled in water versus 8 M urea. These results supported by analysis of a series of order parameters drawn from polymer physics predict that water is a poor solvent for polyglycine, which is a mimic of polypeptide backbones. Recent fluorescence correlation spectroscopy experiments and solubility measurements of polyglycine peptides confirmed these predictions [99].

In a polymer, each residue has reduced translational entropy when compared to free analogs diffusing in solvent. Simple mean-field models show that the entropy of mixing between solute and solvent molecules is reduced by a factor of $N$ if we compare the situation of mixing $N$ freely diffusing solute molecules versus the same $N$ molecules concatenated into a linear polymer. Because of this diminution in mixing entropy, polymers, unlike small molecules, can undergo intramolecular phase separa-
tion to form globules or prefer the coil phase that is well mixed with solvent. Within globules, the concentration of residues around each other is independent of $N$, whereas in coils it decreases as $N^{-0.77}$. As a result, each residue can make collective contacts within globules whereas the coil state is characterized by a combination of negligible intra-chain interactions and additive interactions of individual residues with the solvent. Even for polyamides, where individual amides can be favorably solvated, it is the competition between collective self-interactions within a globule and additive interactions of individual residues with the solvent in the coil state that determines the stable phase. Model compounds do not account for the diminution in translational entropy or the competition between collective intra-polymer and additive polymer-solvent interactions. This partially explains the preference of polypeptide backbones for compact globules despite the favorable free energy of hydration for NMA. Questions remain regarding the balance between chain/solvation entropy and enthalpy, the interplay between backbone hydration and self-solvation of amides, and the comparative roles of hydrogen bonding versus van der Waals interactions in giving rise to the observed phase behavior of polypeptide backbones in dilute and concentrated aqueous solutions. To resolve these issues, we need a systematic investigation of the temperature and cosolute dependencies of the preference for collapsed states in dilute solutions and the solubility boundary in concentrated solutions. Together, these studies will provide insights regarding the effective two-body interactions, which will allow an inference of the sign and magnitude of $v_{ex}$. In addition, comparative studies of constructs with substituted amides such as amide to ester substitutions (to probe the effect of weakened hydrogen bond donors and stronger acceptors) and secondary amide to primary/tertiary amide substitutions (to probe the effect of hydrogen bond donors) will be necessary for quantifying the role of hydrogen bonding in driving
polypeptide backbone collapse.

1.5 Polar tracts form compact globules in aqueous solutions

Among the polar amino acids, IDPs are enriched in histidine (H), glutamine (Q), serine (S), and threonine (T) and are relatively deficient in asparagine (N). Q/N-rich regions are the hallmark of prion forming domains. Q-rich linkers were among the first disordered segments identified from sequence analysis. They are abundant in transactivation domains of transcription factors and in RNA-binding proteins that play important roles in post-transcriptional regulation. Fluorescence correlation spectroscopy measurements revealed that \( \langle R_h \rangle \) scales as \( N^{0.34} \) for monomeric polyglutamine molecules[100]. These results confirmed predictions from steady state Forster resonance energy transfer (FRET) measurements and molecular dynamics simulations, which also show that the densities of polyglutamine globules are similar to those of folded proteins. The low sequence complexity implies a lack of specificity for a single compact conformation and instead heterogeneous ensembles of compact conformations are energetically equivalent. Furthermore, the internal friction is uncharacteristically high for these molecules suggesting glassy dynamics for the conversions between distinct compact conformations. Simulation results obtained using the ABSINTH implicit solvation model show evidence for continuous globule-to-coil transitions for polyglutamine [101]. In accord with polymer physics theories, the stabilities of globular conformations and the sharpness of the globule-to-coil transitions increase with increasing chain length.

Single molecule atomic force spectroscopy studies showed that polyglutamine molecules form compact globules that are mechanically resistant to forces as large as 180 pN [102]. Interestingly, the introduction of proline residues within polyglutamine
tracts increases the mechanical compliance of these polar tracts. The preference for heterogeneous ensembles of compact globular conformations has also been observed using single molecule FRET measurements for Q/N-rich tracts and for glycine-serine block copolypeptides using a combination of time resolved FRET measurements and molecular dynamics simulations. Taken together, it is now clear that specific archetypes — namely sequences enriched in polar residues — form heterogeneous ensembles of compact globules as measured by a range of properties that quantify sizes, shapes, the scaling of inter-segmental distances, and the responses of these sequences to applied force. The preference for collapsed states can be traced, at least partially, to the intrinsic preferences of polypeptide backbones whose conformational properties in aqueous solutions are consistent with water being a poor solvent for polyamides. These results might seem surprising since the preference for collapsed states is realized despite the absence or deficiency of canonical hydrophobic residues in these polar tracts. The results are, however, consistent with the poor solubility profiles of IDP sequences that are enriched in polar residues and highlight the weaknesses underlying additivity assumptions, which suggest that conformational properties and solubility profiles of polymers can be inferred exclusively from the properties of their building blocks. In fact, the assumption of additivity has been questioned in the protein literature and has proven to be invalid for synthetic polymers.

1.6 IDPs with charged residues constitute a distinct archetype from polar tracts

Low hydrophobicity is a defining hallmark of IDP sequences [61, 103]. Given that archetypal polar IDPs form heterogeneous ensembles of collapsed structures in aqueous solutions and that the driving force for collapse originates in the intrinsic preference of polypeptide backbones for collapsed structures in water, is this behavior
ag eneri at tribu te o fa llI D Ps e q u e n ce s? Poly em physics theories suggest that even in poor solvents, polyelectrolytes — sequences enriched in charged residues of one kind — can reverse the preference for collapsed structures. Instead, the preferential solvation of charged residues and intra-chain electrostatic repulsions leads to chain expansion, and depending on the charge content, chain sizes can go beyond expectations for self-avoiding random walks in so-called good solvents. This behavior results without alterations to the solvent properties, but instead is the consequence of the interplay between chain-chain interactions and chain-solvent interactions whereby the former essentially override the latter due to the large electrostatic energies involved. In the next chapter, we explore the dependence of polymeric descriptors described above on the content of charged residues as part of our broader effort to elucidate sequence-ensemble relationships.

1.7 References


corps organiques et le pouvoir rotatoire de leurs dissolutions,” Bulletin de la

1891.


Fischer Use this Analogy?,” Angewandte Chemie International Edition, vol. 33,


1995.


[14] J. Monod, J.-P. Changeux, and F. Jacob, “Allosteric proteins and cellular con-

Synthesis.,” Proceedings of the National Academy of Sciences of the United

tion of native ribonuclease during oxidation of the reduced polypeptide chain,”
1961.


[28] A. Allerhand, D. Doddrell, V. Glushko, D. W. Cochran, E. Wenkert, P. J. Lawson, and F. R. N. Gurd, “Conformation and segmental motion of native and
denatured ribonuclease A in solution. Application of natural-abundance carbon-
13 partially relaxed Fourier transform nuclear magnetic resonance,” Journal of

[29] K. Wüthrich and G. Wagner, “NMR investigations of the dynamics of the
aromatic amino acid residues in the basic pancreatic trypsin inhibitor,” FEBS


[31] S. E. Ealick, “Synchrotron Radiation and the Golden Age of Structural Bi-

[32] P. Wright and J. Dyson, “Intrinsically unstructured proteins: re-assessing the
protein structure-function paradigm,” Journal of Molecular Biology, vol. 293,

[33] V. Uversky, J. Gillespie, and A. Fink, “Why are ”natively unfolded” proteins
unstructured under physiologic conditions?,” Proteins: Structure, Function,

[34] C. M. Kay and M. M. Marsh, “Some optical properties of fetuin and glucagon,”

Phillips, “A three-dimensional model of the myoglobin molecule obtained by

polypeptide hormones,” Annual Review of Biochemistry, vol. 51, pp. 123–154,
1982.

[37] R. S. Spolar and M. T. Record, “Coupling of local folding to site-specific binding

[38] J. Dyson and P. Wright, “Intrinsically unstructured proteins and their func-

“Operational definition of intrinsically unstructured protein sequences based
on susceptibility to the 20S proteasome,” Proteins: Structure, Function, and


Chapter 2

Net charge per residue modulates conformational ensembles of intrinsically disordered proteins

Intrinsically disordered proteins (IDPs) adopt heterogeneous ensembles of conformations under physiological conditions. Understanding the relationship between amino acid sequence and conformational ensembles of IDPs can help clarify the role of disorder in physiological function. Recent studies revealed that polar IDPs favor collapsed ensembles in water despite the absence of hydrophobic groups — a result that holds for polypeptide backbones as well. By studying highly charged polypeptides, a different archetype of IDPs, we assess how charge content modulates the intrinsic preference of polypeptide backbones for collapsed structures. We characterized conformational ensembles for a set of protamines in aqueous milieus using molecular simulations and fluorescence measurements. Protamines are arginine-rich IDPs involved in the condensation of chromatin during spermatogenesis. Simulations based on the ABSINTH implicit solvation model predict the existence of a globule-to-coil transition, with net charge per residue serving as the discriminating order parameter. The transition is supported by quantitative agreement between simulation and experiment. Local conformational preferences partially explain the observed trends of polymeric properties. Our results lead to the proposal of a schematic protein phase diagram that should enable prediction of polymeric attributes for IDP conformational ensembles using easily calculated physicochemical properties of amino acid sequences.
Although sequence composition allows the prediction of polymeric properties, interresidue contact preferences of protamines with similar polymeric attributes suggest that certain details of conformational ensembles depend on the sequence. This provides a plausible mechanism for specificity in the functions of IDPs.

This chapter is adapted from an article [1] published in the Proceedings of the National Academy of Sciences and therefore includes significant contributions from the article’s co-authors. Albert H. Mao (AHM), the candidate, Scott L. Crick (SLC), Andreas Vitalis (AV), and Rohit V. Pappu (RVP) set the specific aims for the research and designed the simulations and experiments. AHM performed the simulations and analyzed their results using software and computational methods that were largely implemented by AV. SLC and Caitlin L. Chicoine (CLC) performed the fluorescence experiments. AHM produced all figures and wrote the paper along with SLC and RVP. We are also grateful to Carl Frieden and Timothy Lohman for assistance with fluorescence measurements, and to the National Science Foundation (MCB 0718924), National Institutes of Health - National Institute of General Medical Sciences 5T32GM008802, and McKelvey Scholars program for funding. The article may be accessed via its digital object identifier (DOI) name 10.1073/pnas.0911107107. Note that the online supporting information accompanying the article has been incorporated into this chapter.

2.1 Introduction

Intrinsically disordered proteins (IDPs) are a class of proteins that fail to fold autonomously in aqueous solutions to well-defined three-dimensional structures [2, 3]. This “intrinsic disorder” has been implicated in a range of regulatory functions that require IDPs to interact with other macromolecular ligands [4–12]. Many of these interactions promote disorder-to-order transitions within IDPs [4, 10], and different
mechanistic models [4, 13, 14] have been proposed for coupled folding and binding.
To develop a better understanding of how disorder is used in function [3], we have
pursued quantitative, polymer-physics-based descriptions [15–17] for conformational
ensembles of IDPs [18–21].

Low hydrophobicity is a defining characteristic of IDP sequences [22, 23].
This suggests that IDPs cannot collapse to form compact, globular conformations
in aqueous solutions [22]. However, spectroscopic [24–28] and computational inves-
tigations [18, 29] have shown that polar tracts form heterogeneous ensembles of col-
lapsed structures in aqueous solutions. These sequences are rich in uncharged, polar
amino acids and are devoid of canonical hydrophobic residues. Collapse of polar
tracts has been observed for polyglutamine [18, 21, 24, 27–29], the N domain of the
yeast prion Sup35 [26], and glycine-serine block copolypeptides [20, 25]. Studies of
polyglycine [20] have shown that polypeptide backbones also form heterogeneous en-
sembles of collapsed structures in water, suggesting that water is a poor solvent for
these constructs and other polar tracts.

For homopolymers, the thermodynamic preference for collapsed globules ver-
sus swollen coils can be inferred from expressions for the free energy of a polymer
in solvent written in terms of chain length ($N$) and effective two- and three-body
terms ($\omega_2$ and $\omega_3$, respectively) that quantify average, solvent-mediated interactions
between chain monomers [30]. Due to excluded volume considerations, $\omega_3$ is always
positive [31–33]. For a given set of solution conditions (set by the temperature $T$
in theories and simulations), the lower entropy globule phase is favored when $\omega_2$ is
negative (low $T$) and the higher entropy coil phase is favored when $\omega_2$ is positive.

We have reported results from atomistic simulations [21] based on the OPLS-
AA/L (Optimal Potentials for Liquid Simulations — All Atom/LMP2) charge set [34]
and the ABSINTH implicit solvation model [35] to study homopolymeric polyglutamime. The energy function used is of the form \( E = U_{\text{tor}} + W_s + W_{\text{el}} + U_{\text{EV}} + U_{\text{Disp}} \). Here, \( U_{\text{tor}} \) denotes torsional potentials and \( W_s \) is the direct mean-field interaction term that uses experimentally determined free energies of solvation for model compounds; \( W_{\text{el}} \) denotes the mean-field electrostatic term. \( W_s \) and \( W_{\text{el}} \) are modulated by the degree of solvation for each atom through a non-linear function of the solvent accessible volume. \( U_{\text{EV}} + U_{\text{Disp}} \) models van der Waals interactions using the Lennard-Jones model, where \( U_{\text{EV}} \) and \( U_{\text{Disp}} \) refer to short-range repulsive and attractive terms, respectively. When all terms except \( U_{\text{EV}} \) are turned off, the simulations sample conformations from self-avoiding random walk ensembles [18, 21] because \( \omega_2 > 0 \). Conversely, when the only nonzero terms in the Hamiltonian are \( U_{\text{EV}} \) and \( U_{\text{Disp}} \), conformations are sampled from a heterogeneous ensemble of globules [18, 21].

For simulations of polyglutamine with the full Hamiltonian, the ensemble of preferred conformations is predominantly globular for a range of physiologically relevant temperatures [21]. These results are congruent with experimental data and a speculative rationalization is provided below.

Model compound studies have been used to classify polar amino acids such as glutamine as being hydrophilic [36]. This is based on transfer free energies of model compounds [37, 38] from oil into water and the favorable free energies of solvation (free energy of transfer from gas phase into solvent) of these compounds in water. Naive extrapolation suggests that sequences such as polar polyglutamine should be miscible in water and form structures that maximize the solute-solvent interface. In reality, these expectations do not hold; instead, polyglutamine forms compact globules in water [24, 27–29]. In polyglutamine, each residue experiences a reduction in translational entropy when compared to free glutamines diffusing in solvent. Mean-
field models show that the entropy of mixing between solute and solvent molecules is reduced by a factor of $N$ if we compare the entropy of mixing for $N$ freely diffusing solute molecules to $N$ residues in a polymer [33]. Additionally, polymers have access to different phases such as globules and coils. In globules, the concentration of chain units around each other is independent of $N$, whereas in coils this concentration decreases as $N^{-0.77}$ [33]. As a result, within globules, residues have access to collective intrachain interactions that can compete against additive interactions of individual residues with the solvent. If the effects of these collective self interactions are stronger than the sum of individually favorable interactions of repeating units with solvent, these favorable self interactions combined with the diminished translational entropy per residue yield the globule as the preferred phase. Studies of model compounds do not account for the diminution in translational entropy nor do they account for the competition between collective intrapolymer and additive polymer-solvent interactions. Such effects may be quantified either by measuring free energies of solvation of appropriate polymeric constructs or by measuring these quantities for model compounds near their solubility limits to account for the competition between self- and cross-interactions.

One might be tempted to conclude that all IDPs form heterogeneous ensembles of collapsed structures in aqueous solvents. Uversky et al. [22] showed that many IDP sequences have high net charge per residue. In a two-dimensional space defined by mean hydrophobicity $\langle H \rangle$ and mean net charge $\langle q \rangle$, a single line $\langle q \rangle = 2.785 \langle H \rangle - 1.151$ separates IDPs from those with well-defined folds [22]. Consequently, we focus here on the effect of net charge on the phase behavior of archetypal IDPs in aqueous solvents. Specifically, we present results from molecular simulations using the ABSINTH model and fluorescence experiments for protamines and
protamine-like polypeptides. These are naturally occurring arginine-rich IDPs [39, 40] that are associated with condensation of chromatin during spermatogenesis [41, 42] and packaging of viral genomes [43].

2.2 Results and discussion

2.2.1 Characteristics of protamine sequences used in this study

Figure 2.1 summarizes the properties of all protamine sequences used in our study. In our simulations, we assumed that the protonation states of ionizable residues are those of isolated amino acids at pH 7.4. For each sequence, Figure 2.1 shows values of $f_+$ and $f_-$, which refer to the fraction of positively and negatively charged residues, respectively, and the values of the net charge per residue defined as $(f_+ - f_-)$. Protamines with nonzero values of both $f_+$ and $f_-$ are asymmetric polyampholytes, whereas sequences for which $f_+ > 0$ and $f_- = 0$ are polyelectrolytes. All of the protamines fall within the intrinsically disordered region of the net charge/hydropathy plane of Uversky et al. [22]. This is confirmed by the disorder scores computed using the VSL2B predictor [44]. For all protamines, the disorder score is close to unity.

Protamines 1 and 3 are predicted to exist based on an open reading frame near the tRNA1Tyr gene of *Escherichia coli* [45]. (MRGRMRSFDQGSTRAPARERCRRQRPEGRSAQR is the sequence of protamine 3; protamine 1 has an identical sequence with the four N-terminal residues removed). Their physiological function is unknown, but their discoverers deemed them “protamine-like” on the basis of size and amino acid sequence. Our results appear to challenge this designation: they have lower arginine content and higher acidic residue content than the true protamines, which cause them to adopt conformational ensembles that are distinct from those characteristic of true protamines, as shown later. Protamine 2 (MFDNASTRNKRERGKRQGKQTRTQRH
**Figure 2.1:** Inventory of protamine sequences. For each protamine, the columns show numeric and graphic identifiers, amino acid sequence, number of residues, UniProtKB accession code, $f_+$, $f_-$, mean Kyte-Doolittle hydropathy score, and the minimum VSL2B disorder prediction score over all residues. Sequences are sorted by their net charge per residue, $f_+ - f_-$. The symbols shown here are used throughout this chapter as protamine identifiers. Note that filled shapes (solid diamonds, circles, and squares) denote polyelectrolytes, whereas thin or hollow shapes denote polyampholytes.

ADRSQT is predicted to exist based on a genomic study of *Bordetella avium* [46]. Like protamines 1 and 3, it has been labeled a “protamine-like protein” despite its distinct sequence and conformational ensemble characteristics. Protamine 4 (MACYPVNIARGGLKNMGMKSRGKGRGK) is predicted to exist based on the genome of *Cricetulus griseus* [47]. Protamine 5 (AGSKRSRSRSRSRSKSPKASAPKSSSRASR) has been detected in vivo in *Ensis minor* [48]. Consistent with our assumption of intrinsic disorder, circular dichroism studies suggest a lack of canonical secondary structure [49].

Protamine 6 (PSPTRRSSRGSRRSRRSASAGKAAKRASKTAK) has been detected in vivo in *Mytilus californianus* [50]. Protamines 7 (MARGRSRSRSVRRRRGGSPRRRRRAGRRSQ RAGAGGLRRRRHRRRADQE) and 8 (MAYGRARSRSVRRRRGRSRRSPRRRRGRSRRDNDAPRRRRRR)
have been detected in vivo in *Cynops pyrrhogaster* [51]. Protamine 9 (ARRRHSMKKKR KSVRRRKTTRKQRKRKNSLGRSFKQHFKPQPPFRP) has been detected in vivo in *Hydrolagus colliei* [52]. Protamine 10 (MRQQASLPARRRRRVRTRVRVRRRRVRGRNH) is predicted to exist based on mRNA transcripts from *Oryzias latipes* [53]. Protamines 11 (PRRR REASRPVRRRRRYYRSTAARRRRRVVRRR) and 15 (PRRRQQASRPVRRRRYRRSTAARRRRRVVRRR RR, a natural variant that differs at only one amino acid) have been detected in vivo in *Thunnus thynnus* [54, 55]. Protamine 12 (PRRRQQASRPVRRRRRTTRSTAERRRRRVVRRR) has been detected in vivo in *Dicentrarchus labrax* [56]. Protamine 13 (PRRRREE TSRPIRRRAARRAPIRRRRRVVRRR) has been detected in vivo in *Mugil cephalus* [57]. Protamines 14 (RRRRRRARQGRRGRRRHRGGRRGGRRRSREQTGRRRRRRRRRMSF), 17 (RRRR RRRGRRGRRGRRRHRGGRRGRRRGRKERTRRRRRRRRRR), and 19 (RRRRRRRRGGGGRRRRRRRRRRRR HGRRRRRRSSREQTRRRRRRRRRR) share identical N-termini and have been detected in vivo in *Chrysemys picta* [58]. Protamine 16 (PRRRRSSRPVRRRRYRRSTAARRRRRVVRRR R) has been detected in vivo in *Sarda orientalis* [59]. Protamine 18 (ARRRRRSSRPQRRR RRRHGRRRRGRRR) has been detected in vivo in *Acipenser guldenshtadi* [60]. Protamine 20 (RRRRRRRRRRRRGRRGRSSSRR) has been detected in vivo in *Octopus vulgaris* [61].

### 2.2.2 Net charge per residue segregates protamine conformations along a globule-to-coil transition

Figure 2.2 shows the ensemble average radii of gyration (Rg) plotted against net charge per residue. Each (Rg) value was divided by that of the same protamine simulated as a self-avoiding random walk, effectively normalizing it by a swollen random coil (referring to self-avoiding random walks, not Flory random coils) reference state. Figure 2.2 suggests the existence of a globule-to-coil transition as the net charge per residue increases. In the range $0.4 < (f_+ - f_-) < 0.65$, polyelectrolytic protamines have slightly larger (Rg) values than their polyampholytic counterparts.
This is true even for polyelectrolyte-polyampholyte pairs with similar/identical values of \((f_+ - f_-)\).

**Figure 2.2:** Normalized \(\langle Rg \rangle\) plotted against net charge per residue. To enable comparisons between polypeptides of different length, the \(\langle Rg \rangle\) values obtained using the full ABSINTH Hamiltonian were divided by that of the same protamine simulated as a self-avoiding random walk. Error bars denote the standard deviation of the normalized \(\langle Rg \rangle\) obtained from the five independent replicate simulations. Since each individual replicate itself constitutes a large sample from a population, we interpret this standard deviation as the standard error of the mean (SEM).
2.2.3 \( \langle R_g \rangle \) exhibits minor sensitivity to sequence permutation

Our usage of sequences from protein databases raises the question of whether or not our observations of the correlation between net charge per residue and quantities such as \( \langle R_g \rangle \) are due to specific effects that arise from the particular sequence. For instance, examination of protamine 4 reveals that hydrophobic residues are concentrated near the N-terminus while basic residues occur more towards the C-terminus. Visualizing the trajectories revealed that, indeed, the N-terminus collapsed while the C-terminus remained extended. Is its apparent character of being an intermediate between globule and coil merely due to being a chimera of a globule and coil? To answer this question, probe the robustness of net charge per residue as an order parameter, and explore the sensitivity of \( \langle R_g \rangle \) to sequence rearrangement, we simulated polypeptides containing permutations of the amino acid sequence of protamine 4. Figure 2.3 shows these sequences and their motivations. By construction, all of them have identical composition, and thus the same net charge per residue and hydropathy, as protamine 4.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Sequence</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>◆</td>
<td>MACYPVINIRARGLGMKKNMGKSRGRGKG</td>
<td>Original sequence</td>
</tr>
<tr>
<td>▼</td>
<td>RRKRRKMACYPVINIAGLGMGMSGGG</td>
<td>Charges to N-terminus</td>
</tr>
<tr>
<td>●</td>
<td>MACYPVINIAGLGMGMSGGGRRRKRRK</td>
<td>Charges to C-terminus</td>
</tr>
<tr>
<td>○</td>
<td>MACYPVINIAGRKKRRKLNMGMSGGG</td>
<td>Charges centered</td>
</tr>
<tr>
<td>☐</td>
<td>GSGGKGMVRIGACRKKMNYPGRRALN</td>
<td>Random permutation</td>
</tr>
<tr>
<td>☠</td>
<td>GRKKNKMICVGYRSRANLRGGANGGP</td>
<td>Random permutation</td>
</tr>
<tr>
<td>+</td>
<td>GRKNCARKGSIGMAKMLRMYPNGGV</td>
<td>Random permutation</td>
</tr>
<tr>
<td>-</td>
<td>YGNKMRAGGLKVMIRAMKGC5RNRGSP</td>
<td>Random permutation</td>
</tr>
</tbody>
</table>

**Figure 2.3:** Inventory of permuted sequences using the composition of protamine 4.

The \( \langle R_g \rangle \) values for each permutant are shown in Figure 2.4. Some permutants
exhibit significant differences: having all the basic residues at either termini resulted in slightly increased $\langle Rg \rangle$ relative to the original sequence, while clustering the basic residues in the middle caused a decreased $\langle Rg \rangle$. However, none of the permutant $\langle Rg \rangle$ values approached that of either the globule or coil reference states, indicating that they remained near the midpoint of the globule-to-coil transition inferred from Figure 2.2. In fact, three of the four random permutant $\langle Rg \rangle$ values were within error of the $\langle Rg \rangle$ for the original protamine. Since the original value is in the transition region, one would expect maximal sensitivity to any effects due to sequence rearrangements. Therefore, our conclusions about the relationship between coarse-grain conformational properties such as $\langle Rg \rangle$ and net charge per residue are robust for this sensitive example. This suggests that it is possible to modify acid/base composition or rearrange residues while keeping the net charge per residue constant without altering polymeric phase behavior.

**Figure 2.4:** Comparison of ensemble average radii of gyration (non mass weighted) for the globule and coil reference states, protamine 4, and its permutants. Error bars convey the standard error of the mean. Symbols are taken from Figure 2.3.
2.2.4 Additional tests support the proposed globule-to-coil transition

Analysis of other properties confirms the globule-to-coil transition suggested in Figure 2.2. Asphericity ($\delta^*$) quantifies the extent of deviation from a perfect sphere ($\delta^* = 0$), attaining its maximum value of one for a perfect rod. For polymers in theta solvents, $\delta^* = 0.39$, whereas $\delta^* = 0.43$ [62] for self-avoiding random walks [63]. Figure 2.5 shows that protamine asphericities change from 0.15 to 0.53 as the net charge per residue increases from 0.24 to 0.4. This is consistent with a transition from nearly spherical globules to ellipsoidal coils. The globule-to-coil transition is also evident in Figure 2.6, which shows the scaling of ensemble average interresidue distances plotted against their linear sequence separations. Protamines 5-21, which are on the coil side of the transition, reveal their power law behavior, in that spatial distances between residues $i$ and $j$ increase with sequence separation $|j - i|$ as expected for statistical coils. They also show increased swelling vis-à-vis the self-avoiding reference chain as net charge per residue increases. Conversely, protamines 1-3 exhibit the characteristic flattening of their internal scaling curves, which is expected for globules [18, 21, 64].

2.2.5 Explicit solvent simulations do not exhibit salt dependence

Prior to the ABSINTH simulations of protamines described in this study, we carried out comparative simulations of Ac-(Arg)$_{25}$-Nme polypeptides using both implicit solvation in ABSINTH and explicit solvation in TIP3P water at three different salt concentrations. The resulting histograms for the mass-weighted radius of gyration are displayed in Figure 2.7. As salt concentration increases, the radius of gyration histograms calculated using the ABSINTH model exhibit an intuitive shift towards more compact conformations. In contrast, the histograms from explicit solvent sim-
Figure 2.5: Asphericity $\delta^*$ plotted against net charge per residue. Error bars denote the SEM.

Simulations do not show any systematic trend with salt concentration; instead, they are fairly similar to each other and to the 1000 mm ABSINTH histogram, and exhibit highly overlapping domains with nearly indistinguishable modes. Given the decrease in $\langle R_g \rangle$ with increasing salt anticipated by both theoretical and experimental investigations of polyelectrolytes [65, 66], the insensitivity of the explicit solvent results to salt concentration is puzzling. One possible explanation is that despite the umbrella sampling protocol (described below) that was designed to achieve thorough sampling of all $R_g$ values, the exploration of conformational space was insufficient to capture effects of different salt concentrations.
Figure 2.6: Scaling of the ensemble average internal distances, \( \langle R_{ij} \rangle \), between residues \( i \) and \( j \) plotted against chain separation, \( |j - i| \). Capping groups at the N and C termini were included in all calculations of internal distances. Gray squares and circles show data obtained from reference simulations for atomistic self-avoiding random walks and self-attracting versions of sequences 16 and 7, respectively. For the self-avoiding random walks, all interactions except the nonbonded steric repulsions are turned off; the self-attracting reference also includes the van der Waals dispersions. Gray diamonds denote the internal scaling profile for a reference rod-like chain. The latter data were obtained from a fully extended conformation for a 25-residue polyarginine chain with all backbone and side chain dihedral angles in trans.
Figure 2.7: Comparison of mass-weighted radius of gyration histograms between different salt concentrations and solvation models. The salt concentration is given in molal units for explicit solvent and molar units for ABSINTH. The bin width is 0.15 Å for all histograms. Histograms from explicit solvent simulations were reconstructed from eight biased histograms using the weighted histogram analysis method (WHAM) [67].

2.2.6 Fluorescence experiments corroborate implicit solvent simulations

To test the predictions made in Figures 2.2, 2.5, and 2.6, we studied a subset of the protamines (2, 4, 5, 6, 8, 11, 12, 16, 18, and 21) using fluorescence correlation spectroscopy (FCS) and fluorescence anisotropy. Both sets of measurements were performed using peptides labeled at their C termini with tetramethylrhodamine-5-maleimide (TMR). Quantitative comparison between simulation results and FCS was
achieved using HYDROPRO [68], which allows us to calculate the ensemble average translational diffusion coefficient from simulation results.

Figure 2.8 shows quantitative agreement between measured and calculated translational diffusion coefficients — the latter obtained from analysis of simulation results. No adjustment to simulation or HYDROPRO parameters was performed to optimize fitting. The fluorescent dye was not included in the simulated systems. To test the contributions from the fluorophore, we also performed FCS measurements using peptides labeled with an AlexaFluor-488-C5-maleimide (Alexa), which differs from TMR in terms of its charge, molecular weight, and flexibility of the linker. Results shown in Figure 2.9 demonstrate that the bulkier, charged, and flexible Alexa dye weakens the quantitative agreement between measured and calculated diffusion coefficients for sequences that form globules. However, both datasets show that simulations based on ABSINTH yield quantitatively accurate descriptions of features such as the scaling of overall sizes of IDPs as a function of the net charge per residue.

We also used steady-state fluorescence anisotropy measurements to assess how rotational diffusion varies with increasing net charge per residue. Fluorescence anisotropy ($r$, not to be confused with the Pearson correlation coefficient) is influenced by the average shape and size of a labeled polymer [69]. Data for $r$ plotted against the net charge per residue are shown in Figure 2.10 for protamines 2, 4, 5, 6, 8, 11, 12, 16, 18, and 21, respectively. Because both size and shape influence rotational diffusion, both $\langle R_g \rangle$ and $\delta^*$ should be correlated to $r$, and these correlations are plotted in Figures 2.11 and 2.12, respectively. Taken together, the FCS and fluorescence anisotropy measurements provide two independent tests of predictions that result from molecular simulations.

The level of detail provided by atomistic simulations opens the possibility of
Figure 2.8: Comparison of translational diffusion coefficients obtained from analysis of simulation results (ordinate) and FCS experiments (abscissa). The black line is the result of linear regression. The regression parameters are such that calculated \( D = 11.2 \, \mu m^2/s + 0.94 \times \) measured \( D \). The relevant quantities for comparing numbers from simulation and experimental data are the Pearson \( r \) value, which is 0.96, and the rmsd between calculated and measured values of \( D \), which is 7.05 \( \mu m^2/s \). Error bars denote the SEM.
Figure 2.9: Correlation between measured and calculated translational diffusion coefficients. This figure is analogous to Figure 2.8, except the data are from experiments with Alexa 488 instead of TMR. Symbols are consistent with Figure 2.1. Error bars denote the standard error of the mean. The black line is the result of linear regression. The regression parameters are such that calculated $D = -62.0 \mu m^2/s + 1.52 \times$ measured $D$. The Pearson $r$ is 0.90 and root mean square deviation is 16.3 $\mu m^2/s$. Error bars denote the SEM.
Figure 2.10: Steady-state fluorescence anisotropy $r$ plotted against the net charge per residue. $r$ is a dimensionless quantity that is defined in Equation 2.7. The data convolve contributions from size and shape, and this explains the increased anisotropy of protamine 8, which is a 44-mer, vis-à-vis those with higher net charge per residue.

A quantitative comparison with experimental measurements beyond the ones reported here. In particular, X-ray and neutron scattering are capable of probing solution-state conformational ensembles on a variety of length scales simultaneously. Since a scattering profile conveys more information than a single number, these experiments would provide a more detailed report on the average structural arrangements of the conformational ensemble compared to steady-state fluorescence measurements. Although we lack the ability to perform these experiments, we present the calculated
Figure 2.11: Correlation between fluorescence anisotropy and radius of gyration. The Pearson correlation coefficient is 0.87. Error bars denote the standard error of the mean.
Figure 2.12: Correlation between fluorescence anisotropy and asphericity. The Pearson correlation coefficient is 0.89. Error bars denote the standard error of the mean.
Kratky profiles of each protamine in Figure 2.13 as a computational prediction that may be tested by experimentalists in the future.

![Kratky profiles](image)

**Figure 2.13:** Calculated Kratky profiles for all protamines. Symbols are consistent with Figure 2.1. The wavenumbers range from $q = 0$ to $q = 0.9 \, \text{Å}^{-1}$ in steps of $0.015 \, \text{Å}^{-1}$.

### 2.2.7 Increased swelling of charged protamines is realized through local conformational preferences for P$_{II}$ helices

Krimm and Mark proposed that left-handed, three-residue-per-turn helices such as polyproline II (P$_{II}$) are optimal regular conformations for minimizing electrostatic repulsions between charged side chains in polyelectrolytic sequences [70]. We analyzed the preferences of each simulated protamine for these conformations. Figure 2.14 plots the P$_{II}$ population for each protamine; these were calculated as the average fraction of conformations with at least three contiguous residues in P$_{II}$ bins.
This population is smallest for globule-forming sequences and largest for the polyarginine sequence; in accord with the predictions of Krimm and Mark [70], the preference for PII increases with increasing net charge per residue. For sequences with similar net charge per residue, this preference is smaller for the asymmetric polyampholytes than for the polyelectrolytes.

![Polyproline II helix propensity](image)

**Figure 2.14:** Ensemble average values for PII propensity plotted against net charge per residue. Error bars denote the SEM.

In contrast with the polyproline II content, there was no clear correlation between net charge content and α-helical propensity, which is shown in Figure 2.15. Protamines 11 and 12 exhibited markedly increased mean α-helical propensity, a finding that is also evident from their contact maps. However, the large error bars
resulting from high variation between the means from each trajectory indicate that the equilibrium ensemble contains a diversity of states with different \( \alpha \)-helical content with substantial energy barriers between them, making adequate sampling of the distribution of alpha-helical content more difficult.

**Figure 2.15:** Plot of \( \alpha \)-helical propensity against net charge per residue for the protamines. Propensities are quantified as the fraction of conformations with three or more residues in the \( \alpha \) basin shown in Figure 2.31. Error bars denote the SEM.

Similar to polyproline II propensity, \( \beta \)-strand propensity (Figure 2.16) appears to increase with net charge, with the weak polyampholytes exhibiting diminished signal compared to polyelectrolytes with the same net charge. However, the fraction of residues in \( \beta \) strands does not increase significantly past a net charge per residue value
of 0.5, and the maximum value attained by polyarginine of about 8% is significantly below the 23% for polyproline II.

![Graph showing β-strand propensity against net charge per residue for the protamines.](image)

**Figure 2.16**: Plot of β-strand propensity against net charge per residue for the protamines. Propensities are quantified as the fraction of conformations with two or more residues in the β basin shown in Figure 2.31. Error bars convey the standard error of the mean. Note the greatly reduced β-strand propensities compared to polyproline II helix propensity, despite the inclusion of two-residue segments.

### 2.2.8 Conformational properties of protamines exhibit sequence specificity

Analysis of conformational characteristics in terms of \( \langle H \rangle \) and net charge per residue can mask sequence-specific conformational preferences that distinguish pro-
tamines with similar values for either parameter. Column 1 in Figure 2.17 compares contact maps for protamines 8 and 9, which have similar hydrophobicity and identical net charge per residue. Protamine 8 has a higher fraction of charged residues and is an asymmetric polyampholyte, whereas protamine 9 is a polyelectrolyte. The contact map for protamine 8 indicates a finite probability of forming a compact microdomain at the C terminus. This is realized through electrostatic interactions involving the C-terminal arginine residues and two aspartate residues that are directly N-terminal to the arginine cluster. In contrast, protamine 9 exhibits quantifiable preference for nonlocal contacts within its C-terminal region, which is consistent with depletion of charged residues in this region. Column 2 in Figure 2.17 compares the contact maps of protamines 11 and 12, respectively. They have identical lengths, net charge per residue, and values for $f_+$ and $f_-$. They differ only at positions 6, 17, and 23 and exhibit similar radii of gyration, asphericities, and $P_{II}$ populations. However, the contact maps reveal differences in the extent and location of segmental alpha-helical preference. These results suggest that atomistic simulations and experimental characterization are needed to discern the preferred local and nonlocal contact patterns that are specified by the details of amino acid sequence.

Figures 2.18 and 2.19 show all twenty-one ensemble average contact maps. The density of the contact maps for the three protamines with lowest net charge indicates an abundance of long-range contacts, which is consistent with our description of them as collapsed globules. Likewise, the sparseness of contact maps for the higher net charge protamines is anticipated from theoretical treatments of random coil ensembles [71]. The presence of gray cells in all contact maps, indicating contacts that are present in only a fraction of samples from the ensemble, emphasizes the diversity of conformations encountered for each of these disordered proteins.
Figure 2.17: Selected ensemble average contact maps. Protamines 8 and 9 are shown in the first column, and protamines 11 and 12 are shown in the second column. Each contact map is annotated with a structure taken from the simulated ensemble and is intended to assist in visual interpretation of the contact maps.
Figure 2.18: Ensemble average residue contact maps for protamines 1–12.
Figure 2.19: Ensemble average residue contact maps for protamines 12–21. The protamine symbol is in the lower-right corner of each contact map. Two residues were considered contacting if any of their atoms were within 3.5 Å. Self-contacts are omitted. The same linear grayscale shading is used for all contact maps.
2.2.9 Classification of different collapsed states

The increase of $\langle R_g \rangle$ with increasing net charge per residue can be rationalized as a consequence of intramolecular electrostatic repulsions between arginine side chains and the favorable solvation of these moieties. The observed preference for collapsed states in three of the 21 protamines merits further discussion. Results from previous studies and data in Figures 2.2-2.10 suggest the existence of at least three classes of globule-forming sequences for which $\langle H \rangle \approx 0$ or $\langle H \rangle$ is sufficiently small and $f_+ - f_- \leq 0.2$. These are as follows: (i) Polar tracts, where $f_+ = f_- = f_+ - f_- = 0$, which collapse as discussed earlier. (ii) Weak polyelectrolytes/polyampholytes, where $f_+ - f_- \leq 0.2$ and each of $f_+$ and $f_-$ are small. In weak polyelectrolytes, favorable free energies of solvation of the charged moieties in the side chains and intramolecular electrostatic repulsions cannot overcome the driving force that leads to collapse of polar tracts. In polyampholytic variants, electrostatic interactions between oppositely charged side chains provide an additional source of stabilization for globules. (iii) Strong polyampholytes, where $f_+$ and $f_-$ are large and approximately equal. In such sequences, electrostatic interactions between solvated side chains of opposite sign make important contributions to globule stability [72]. This effect is reminiscent of work by Makhatadze and coworkers who showed that through-solvent electrostatic attractions between opposite charges on protein surfaces play an important role in stabilizing globular structures of water-soluble proteins [73, 74]. To test the validity of the strong polyampholyte classification, which was not represented in the protamine inventory, we performed simulations for eight short polyampholytic sequences. As expected [72], these strong polyampholytes are strong globule formers. Analysis presented in the following section shows that electrostatic interactions between oppositely charged, surface-exposed side chains can contribute to the collapse of strong
polyampholytes.

2.2.10 Electrostatic attraction drives collapse of polyampholyte sequences

Since sequences with low net charge per residue are sparse within our set of protamines, we performed simulations of eight additional sequences labeled as intrinsically disordered regions within the DisProt database [75]. The sequences, which are listed in Figure 2.20, achieve low net charge by having more balanced and moderate fractions of acidic and basic residues compared to the protamines, which are polyelectrolytic or weakly polyampholytic. However, they range from twenty to thirty residues in length and are therefore shorter than the protamines on average.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Sequence</th>
<th>Length</th>
<th>( f_+ )</th>
<th>( f_- )</th>
<th>( f_+ - f_- )</th>
<th>( H )</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>QQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ</td>
<td>30</td>
<td>0.00</td>
<td>0.00</td>
<td>+0.00</td>
<td>0.11</td>
</tr>
<tr>
<td>23</td>
<td>QGNDKDYIDLAFSKKKADDRKE</td>
<td>22</td>
<td>0.27</td>
<td>0.27</td>
<td>+0.00</td>
<td>0.30</td>
</tr>
<tr>
<td>24</td>
<td>SDNREENRRNRQAQQQTAETRESRQAEV</td>
<td>30</td>
<td>0.23</td>
<td>0.20</td>
<td>+0.03</td>
<td>0.21</td>
</tr>
<tr>
<td>25</td>
<td>RDRDQRDLTGDPWGGRTLE</td>
<td>20</td>
<td>0.20</td>
<td>0.25</td>
<td>–0.05</td>
<td>0.28</td>
</tr>
<tr>
<td>26</td>
<td>EREAARQKKQRAIGSADTDRDAKREFHSKY</td>
<td>30</td>
<td>0.30</td>
<td>0.20</td>
<td>+0.10</td>
<td>0.29</td>
</tr>
<tr>
<td>27</td>
<td>GNVPNKSSTKGKGRKLVDDDD</td>
<td>23</td>
<td>0.30</td>
<td>0.17</td>
<td>+0.13</td>
<td>0.30</td>
</tr>
<tr>
<td>28</td>
<td>MTETDKKQEQQENHAECEDKPK</td>
<td>22</td>
<td>0.18</td>
<td>0.32</td>
<td>–0.14</td>
<td>0.24</td>
</tr>
<tr>
<td>29</td>
<td>DDKRQAQQEAKALNVEEQSVQETEQEER</td>
<td>28</td>
<td>0.14</td>
<td>0.32</td>
<td>–0.18</td>
<td>0.27</td>
</tr>
<tr>
<td>30</td>
<td>PEEKKEEEGSANRQPEDQELESLSA</td>
<td>25</td>
<td>0.16</td>
<td>0.36</td>
<td>–0.20</td>
<td>0.28</td>
</tr>
</tbody>
</table>

**Figure 2.20:** Inventory of additional simulated sequences with low net charge per residue and hydrophobicity.

Figures 2.21, 2.22, and 2.23 are analogous to Figures 2.2, 2.5, and 2.6 and respectively plot the normalized radii of gyration, asphericity, and internal scaling profiles, respectively, for the additional sequences plus a polyglutamine dataset from previous work [76]. Consistent with the protamines, these analyses support the existence of a globule-to-coil transition along the axis of net charge per residue.

Comparison between the histograms in Figure 2.24 and 2.25 reveals the im-
Figure 2.21: Normalized $\langle \text{Rg} \rangle$ plotted against net charge per residue. This figure is analogous to Figure 2.2, but includes the additional sequences listed in Figure 2.20. Error bars denote the SEM.

The importance of attractive electrostatic interactions between solvent-exposed sidechains of opposite charge in stabilizing the globular states of strong polyampholytes. Contacts between sidechains of opposite charge occur on the surfaces of these globules, as shown in Figure 2.26. These data reinforce the classification scheme for globules that was introduced in the preceding section. However, the relatively short lengths of the sequences studied prevents the existence of more complicated patterning effects from being ruled out, as simultaneous realization of solvent exposure and proximity to opposite charge is only possible for small globules.
2.2.11 Anion contact numbers are mildly enhanced around basic residues

Figure 2.27 shows the residue-resolved ensemble average number of chloride ion contacts for each protamine. Similar to the contact maps, a residue was deemed to be in contact with an ion if any of its atoms were within 3.5 Å of the ion; contacts were accumulated every 1000 steps. As expected, there tends to be an accumulation of chloride ions around stretches rich in basic residues; however, no residue has greater than 0.2 ion contacts on average. This is consistent with a liquid-like organization of
Scaling of internal distances

Figure 2.23: Scaling of the ensemble average internal distances, \( \langle R_{ij} \rangle \) plotted against chain separation, \( |j - i| \). This figure is analogous to Figure 2.6, but includes the additional sequences listed in Figure 2.20.

the counterions around the polypeptide macroion.

2.2.12 Simulation and experimental results are consistent with theories for polyelectrolytes and polyampholytes in poor solvents

Ha and Thirumalai [77] developed a self-consistent variational theory [78] to describe how sizes of individual polyelectrolyte chains vary in a poor solvent as a function of the net charge per monomeric unit. Their theory yields a compact expression for a quantity \( X_{\text{eff}} \) that measures an effective excluded volume. For a homopolymeric
Figure 2.24: Combined pairwise distance histograms for sidechain tips of polyampholytic sequences simulated with the full ABSINTH Hamiltonian. The histograms from all polyampholytic sequences (23 – 30) are pooled together. Each curve is independently normalized. For each residue, the sidechain tip is defined as the carbon atom farthest away from the backbone (glycine is omitted). The bin width is 0.1 Å.

Chain of length $N$, $X_{\text{eff}} = \left[ -k_1 |\omega_2| + k_2 \frac{Z^2 u}{(\kappa b)^2} \right] N^{1/2}$; here, $k_1$ and $k_2$ are constants and $\omega_2$ is the magnitude of the effective two-body interaction in the absence of electrostatics. This term is negative in a poor solvent. $Z$ is the net charge, $\kappa^{-1}$ is the Debye screening length, $b$ is the radius of each monomeric unit, and $u = l_B/b$, where $l_B$ is the Bjerrum length — the length scale at which intermonomer electrostatic interactions equal thermal energy. If the net charge per monomer is small and $X_{\text{eff}} < 0$, then the chain prefers collapsed states because the term $-k_1 |\omega_2| N^{1/2}$ dominates. Within a nar-
Figure 2.25: Combined pairwise distance histograms for sidechain tips of polyampholytic sequences simulated with a Lennard-Jones reference Hamiltonian.

row range of net charge values, $X_{\text{eff}}$ approaches zero. The value of the net charge per monomer for $X_{\text{eff}} = 0$ corresponds to the theta point for the polyelectrolyte in a poor solvent. For $X_{\text{eff}} > 0$, the chain expands and accesses the coil state that is congruent with the state accessed by the uncharged polymer in a good solvent. This state is distinct from the rod-like state because entropy opposes the effects of electrostatic repulsion, preventing the maximal extension necessary to attain a rod. Additionally, $X_{\text{eff}} > 0$ increases gradually with increased net charge per monomer, causing an increase in chain dimensions vis-à-vis the swollen random coil state. The theory is also applicable to the case of weak/strong polyampholytes because the balance between collapsed and swollen states is dictated by the interplay between the first and second
Figure 2.26: Illustrative structures from the ensembles of polyampholytic globules. The structures were selected to emphasize contacts between distant residues. Polypeptides are rendered in space filling mode using the following color scheme: positively charged residues are blue, negatively charged residues are red, uncharged polar “hydrophilic” residues are green and canonical hydrophobic residues are yellow. Histidine is deemed to be polar here since it is uncharged in our simulations. Hydrogen atoms were not included to generate the structural representations. These structures show the local and non-local clustering of residues of opposite charge.
Chloride ion contacts

Our results motivate a speculative generalization, which if valid would enable the prediction of polymeric phase behavior of proteins from simple sequence characteristics. Figure 2.28 depicts a schematic phase diagram that summarizes the findings of Xeff. All aspects of our results seem to be congruent with these theoretical predictions, thereby providing a coherent explanation for our observations.

2.2.13 Toward a sequence-space phase diagram for proteins

Rooney et al. [79] noted that, although protamine sequences vary rapidly through evolution, their overall arginine content is conserved. This feature should lead to similar values for the types of properties quantified in Figures 2.2–2.10, and might be relevant for maintenance of physiological function despite evolutionary pressures. Our results motivate a speculative generalization, which if valid would enable the prediction of polymeric phase behavior of proteins from simple sequence characteristics. Figure 2.28 depicts a schematic phase diagram that summarizes the findings.

**Figure 2.27:** Residue-resolved ensemble average chloride ion contact numbers. A residue was deemed to be in contact with an ion if any of its atoms were within 3.5 Å of the ion.
from previous studies [20, 22, 24–26] and those from the current investigation. The three axes denote \( \langle H \rangle \), \( f_+ \), and \( f_- \), respectively, and each of these parameters varies between zero and one. The line of Uversky et al. [22] is a plane separating folded proteins from IDPs. Contrary to previous assumptions [17], the phase diagram is not featureless below this plane. Sequences with low overall hydrophobicity can either be swollen coils or compact globules and the net charge per residue determines the preferred phase. An anonymous reviewer notes that the aggregation propensities of the protamines studied here decreases with increasing net charge per residue, suggesting a role for electrostatic interactions in promoting IDP solubility. These propensities may be calculated using the Zyggregator program [80].

The results shown in Figures 2.2, 2.6, and 2.14 illustrate small yet statistically significant differences between polyelectrolytes and polyampholytes, even for cases where the net charge per residue is identical. This suggests that the phase boundaries depicted in Figure 2.28 may vary with \( f_+ \) and \( f_- \) individually rather than being a function of the net charge per residue alone.

Testing predictions of the proposed phase diagram requires quantitative studies of conformational characteristics for a wide range of IDPs. This should be tractable in light of the computational efficiency of simulations based on implicit solvation models such as ABSINTH. Indeed, the protamine simulations would not have been feasible without ABSINTH — a point underscored by comparisons between simulations using explicit and implicit solvent models.

2.3 Materials and methods

2.3.1 Sequence selection

Protamine sequences were selected by searching for the term “protamine” on the National Center for Biotechnology Information (NCBI) Protein (http://www.
Figure 2.28: Proposed schematic phase diagram for the single chain phase behavior of unbound, single domain proteins. The three-dimensional sequence space is defined by $f_+$, $f_-$, and mean hydropathy. The space is a pyramid instead of a cube because high hydropathy and high fractions of charged residues are mutually exclusive. The boundary separating folded proteins from IDPs is a three-dimensional rendering of the results from Uversky et al. [22]. Within the intrinsically disordered region, the boundaries separating disordered globules from swollen coils are extrapolated from the results of the present study.
ncbi.nlm.nih.gov/sites/entrez) and UniProt Knowledgebase [81] databases on June 20, 2008. The search was performed on all fields in order to maximize the pool of candidate sequences. The results were exported as FASTA files and duplicate sequences were removed. Sequences containing unknown residues, more than one cysteine residue, fewer than 20, or greater than 50 total residues were excluded. The resulting list of 70 sequences was sorted according to net charge per residue, and the final 20 were selected from this list to represent a broad range of net charge per residue values while exhibiting a variety of sequence lengths and compositions. The 20 protamines in our list included pairs selected in order to compare sequences with equal net charge: 8 and 9 are totally dissimilar, while 11 and 12 differ at only three positions. Finally, an artificial polyarginine sequence of length 34 was included to provide a sequence with maximal net charge per residue.

2.3.2 Simulations

Markov chain Metropolis Monte Carlo simulations using the ABSINTH implicit solvation model and the OPLS-AA/L charge set were performed in the canonical ensemble \((T = 298K)\). Each capped sequence was enclosed in a spherical droplet of radius of at least 70 Å. We modeled explicitly represented \(\text{Na}^+\) and \(\text{Cl}^-\) ions sufficient to neutralize the net polypeptide charge and mimic a 125 mM salt solution. Details regarding the design of simulations, assessments of convergence, and the analysis of simulation results are described in the following sections.

The ABSINTH model for molecular simulations

All simulations were carried out using a molecular mechanics forcefield based on the OPLS-AA/L parameters [34] and the ABSINTH implicit solvent model [35]. Polypeptides and ions were modeled in atomic detail. The energy function has the
form shown in Equation 2.1:

\[ E_{\text{total}} = W_s + W_{el} + U_{\text{LJ}} + U_{\text{tor}} \]  

(2.1)

The terms \( W_s, \) \( W_{el}, \) and \( U_{\text{LJ}} \) apply for intrapeptide, ion-peptide, and ion-ion interactions, whereas the term \( U_{\text{tor}} \) applies for intrapeptide interactions alone and denotes torsional potentials taken from the OPLS-AA/L forcefield to maintain peptide dihedral angles in predominantly trans-configurations. \( W_s \) is the direct mean field interaction (DMFI) term that captures the transfer of solutes (polypeptide plus ions) in a specific conformation from the gas phase into the continuum solvent with dielectric constant of \( \epsilon = 78, \) and \( W_{el} \) denotes the mean field electrostatic term. \( U_{\text{LJ}} \) models van der Waals interactions using the Lennard-Jones (LJ) model. Parameters for LJ radii and well depths are based on heats of fusion data for model compounds and are different from the choices made in standard forcefields. In the ABSINTH model, polypeptide chains are decomposed into a set of distinct solvation groups. Each freely diffusing ion also forms a distinct solvation group. \( W_s \) is a sum of contributions from each solvation group and for each of these groups experimentally measured free energies of solvation of appropriate model compound analogs are used as references for fully solvated states. The degree of solvent accessibility modulates the DMFI and consequently \( W_s \) varies with chain conformation and arrangement of the ions. Solvent accessible volume fractions are used as the metric for solvent accessibility and this is used to evaluate the solvation states \( \nu_k^i \) for atom \( k \) in solvation group \( i. \) Intrapeptide, peptide-ion, and ion-ion electrostatic interactions are fully screened by the continuum dielectric if the atoms are fully exposed to solvent. However, the screening of electrostatic interactions varies with chain conformation and ion arrangements. The
solvation states of atoms also determine the extent to which screening of electrostatic interactions is modulated. In accord with the EEF1 model [82], the process of transferring solutes from the gas phase into the continuum solvent is treated in one step without parsing the distinct contributions from polar and non-polar components. However, ABSINTH differs from EEF1 in the handling of electrostatic interactions between solute atoms. There are no explicit distance dependencies for the dielectric response. Hence, ABSINTH captures the conceptual strengths of generalized Born models [83–85] while retaining the efficiency of the EEF1 paradigm. Although ABSINTH is designed for studying disordered systems, it does not inherently bias systems towards disorder and is capable of maintaining the stability of folded proteins [35].

Simulation design

The implicit solvation, electrostatic, Lennard-Jones, and Engh-Huber crystallographic geometry parameters described previously [35] were used with some changes. First, the free energies of solvation for model compounds comprising ionic sidechains were restored to $-70$ kcal/mol from their artificially lowered values that were intended to prevent salt bridging. The histidine sidechains were protonated at the epsilon position, which makes the sidechain uncharged. Second, the solvation state of each atom was replaced by the charge-weighted average solvation state of its charge group during computation of electrostatic screening factors. Finally, cutoffs for the Lennard-Jones and electrostatic interactions between neutral groups were set at 10 and 14 Å, respectively. Electrostatic interactions between ions in solution and sidechain moieties with an overall net charge were computed without employing cutoffs.

Each sequence was capped with acetyl and N-methylamide groups at the N and C-termini, respectively, and placed inside a spherical droplet along with explicitly represented Na$^+$ and Cl$^-$ ions sufficient to neutralize the net polypeptide charge and
mimic a 125 mM salt solution. This concentration corresponds to 108 excess ion pairs for 70 Å droplets and 315 excess ion pairs for 100 Å droplets. The simulation conditions mimic poor solvent conditions for the uncharged polymer without counter or coions [21]. Metropolis Markov chain Monte Carlo (MC) simulations were performed in the canonical ensemble at 298 K. The degrees of freedom for the MC simulations were the backbone $\phi$, $\psi$, $\omega$, and sidechain $\chi$ dihedral angles and the rigid-body coordinates for the polypeptide chains and ions. For each sequence, five independent replicates were simulated using randomly generated starting conformations taken from the excluded volume ensemble [71].

Eight protamines (2, 3, 4, 10, 15, 16, 18, and 20) with lengths up to 34 residues were simulated using a droplet radius of 70Å for a total of $2.6 \times 10^7$ steps, with the first $10^6$ omitted from analysis as equilibration. The remaining twelve sequences, with lengths up to 49 residues, were simulated later using a droplet radius of 100 Å for a total of $5.2 \times 10^7$ steps, with the first $2 \times 10^6$ omitted.

The permutant simulations were performed with a more recent version of our simulation software and slightly different parameters. This newer version implemented cluster rigid body moves, which proposed concerted translation of and rotation around the center of mass of several molecules. These moves were attempted with a probability of 0.1 once a rigid body move was selected, for an overall probability of 0.005. Sampling was enhanced by increasing the total number of Monte Carlo steps to $6.5 \times 10^7$, with the first $5 \times 10^6$ discarded as equilibration. The droplet radius was increased to 110 Å. Finally, the temperature was increased to 310.15 K, and the solvation state of each atom was no longer replaced by the charge-weighted average solvation state of its charge group during computation of electrostatic screening factors, restoring the behavior described in Vitalis and Pappu [35]. Note that the
original protamine 4 and reference states were all simulated again using this revised methodology, so the comparison is internally consistent. The change in the calculated \( \langle R_g \rangle \) values for protamine 4 was minor: 11.7 ± 0.2 Å for the original versus 11.2 ± 0.3 Å for the revised methodology. For the additional sequences with low net charge per residue, the simulation methods were similar to those employed for the permutant sequences, with the following differences: a total of \( 6.0 \cdot 10^7 \) Monte Carlo steps were attempted, with the first \( 10^7 \) discarded as equilibration, the droplet radius was restored to 100 Å, and KCl was used in place of NaCl.

**Move set details**

The Monte Carlo move sets used to sample rigid body and conformational degrees of freedom is presented in Figure 2.29. Rigid body moves consisted of simultaneous translations and rotations. Both types of movement allowed either full randomization or local sampling with a 2 Å maximum displacement for translation and a 10° maximum step for rotation. Each sidechain dihedral move attempted four consecutive cycles of simultaneous adjustments to \( \chi \) angles from the same residue. During each cycle, a \( \chi \) angle was randomly perturbed with probability equal to \( 2 / (\text{number of } \chi \text{ angles in the sidechain}) \), or 1 if the sidechain contained two or fewer \( \chi \) angles. Changes to \( \chi \) included full randomization or local perturbations with a maximum step of 30°. Proline constituted a special case, where sidechain moves consisted of choosing one of two pucker states (\( C_{\gamma} \) endo or \( C_{\gamma} \) exo). Backbone dihedral angles were sampled either individually or eight at a time using the algorithm of Favrin et al. [86] with \( a = 10 \) and \( b = 10 \). Individual backbone moves either pivoted the \( \omega \) dihedral angles alone or \( \phi \) and \( \psi \) simultaneously while keeping either the N or C-terminus fixed, using full randomization or local steps of 2.5° for \( \omega \) and 5° for \( \phi/\psi \). For all local moves, the actual step size attempted was chosen uniformly on the interval from zero.
to the maximum, with both directions being equally probable.

Figure 2.29: The decision-tree used by the simulation software to select a Monte Carlo move at each step. Each non-leaf node corresponds to a class of moves; each node is annotated with the overall probability of that move or class of moves being selected. Each edge is annotated with the probability of the decision process branching towards the child once the parent has been reached. The decision is complete once a leaf node is reached.

One weakness our simulation methodology shares with nearly all molecular mechanics computational studies of biomolecules is the use of fixed protonation state
of ionizable sidechains. Although the protonation and tautomeric state expected to
dominate at neutral pH was used to model each residue, the resulting ensemble is
distinct from that which would be rigorously explored at neutral pH. In particular,
conformations that are stabilized with non-dominant protonation and tautomeric
states are quenched, and the actual pH is undefined due to the lack of protonation
equilibrium. However, the intrinsically disordered nature of the protamines in this
work should enable ample access to solvent for all the charged sidechains, leading to
pKa shifts significantly less severe than the 4-5 log units [87, 88] observed for buried
residues. Given the high pKa of the charged arginine sidechain, the low numbers
of histidine residues in our selected protamines, and the expected stabilization of
deprotonated states of acidic residues in an arginine-rich context, we expect that
the approximation of fixed protonation and tautomeric states is reasonable. Going
forward, investigations of the conformational equilibria of IDPs should benefit from
the development of constant pH simulation methods, which adjust the protonation
and tautomeric states of titratable residues to attain equilibrium where the pH is a
true thermodynamic ensemble variable [89–91]. In particular, the implicit solvation
Monte Carlo approach described by Mongan et al. [92] should be ideally suited to
integration with our simulation software, with the ABSINTH implicit solvation model
replacing the generalized Born model.

**Similarity of initial conformations**

Distinct, randomly generated starting conformations drawn from the excluded
volume ensemble [71] were used to initiate the five independent simulations for each
protamine. The degree to which adequate mixing of Markov chains and convergence
of ensemble average estimates were attained is quantified by the size of our error bars,
provided that starting conformations are sufficiently different from one another. To
quantify this diversity, we calculated the least root-mean-square-deviation (RMSD) between each pair of starting conformations. Structures were aligned to minimize the RMSD between their backbone atoms using the Kabsch algorithm [93] as implemented in Visual Molecular Dynamics (VMD) [94]. For each protamine, Figure 2.30 shows the minimum, maximum, and average RMSD between backbone atoms over the ten possible pairs of five starting conformations for that protamine. The data show that starting conformations are never closer than 5 Å in RMSD space, and generally about 9 to 11 Å apart on average. This provides confidence that our error bars reflect the true uncertainty of ensemble average estimates from a set of well-sampled trajectories as opposed to a set of trajectories that started and remained trapped in the same basin of the energy landscape.

Simulation analysis

The non-mass-weighted radius of gyration of the polypeptide was accumulated once every 100 steps. Contact maps were accumulated every 1000 steps. Two residues were considered to be in contact if any of their atoms were closer than 3.5 Å; self and chain-neighbor contacts were explicitly excluded. Internal scaling curves were computed by averaging the distance between every pair of atoms belonging to residue pairs with a given chain separation, with the separation ranging from zero to one plus the total number of residues due to inclusion of the caps; curves were accumulated every 1000 steps.

Calculation of secondary structure propensities: We quantified secondary structure propensities in terms of population of various basins on the Ramachandran map. To calculate helical/strand propensities, the Ramachandran map was divided into \((10^\circ)^2\) bins and certain bins were labeled as alpha, beta, or proline II (see Figure 2.31). In a given conformation, a residue was classified as helical
<table>
<thead>
<tr>
<th>Protamine</th>
<th>Min</th>
<th>Avg</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>06.7</td>
<td>08.8</td>
<td>12.7</td>
</tr>
<tr>
<td>2</td>
<td>07.2</td>
<td>08.9</td>
<td>11.6</td>
</tr>
<tr>
<td>3</td>
<td>07.4</td>
<td>09.1</td>
<td>11.1</td>
</tr>
<tr>
<td>4</td>
<td>06.2</td>
<td>08.5</td>
<td>11.0</td>
</tr>
<tr>
<td>5</td>
<td>08.5</td>
<td>11.3</td>
<td>14.3</td>
</tr>
<tr>
<td>6</td>
<td>06.9</td>
<td>10.1</td>
<td>13.5</td>
</tr>
<tr>
<td>7</td>
<td>09.7</td>
<td>13.4</td>
<td>21.3</td>
</tr>
<tr>
<td>8</td>
<td>08.8</td>
<td>11.5</td>
<td>16.5</td>
</tr>
<tr>
<td>9</td>
<td>09.1</td>
<td>11.6</td>
<td>14.2</td>
</tr>
<tr>
<td>10</td>
<td>07.5</td>
<td>08.5</td>
<td>10.0</td>
</tr>
<tr>
<td>11</td>
<td>05.3</td>
<td>09.2</td>
<td>11.2</td>
</tr>
<tr>
<td>12</td>
<td>09.1</td>
<td>09.8</td>
<td>11.6</td>
</tr>
<tr>
<td>13</td>
<td>07.4</td>
<td>08.7</td>
<td>10.9</td>
</tr>
<tr>
<td>14</td>
<td>08.0</td>
<td>11.2</td>
<td>15.4</td>
</tr>
<tr>
<td>15</td>
<td>06.4</td>
<td>08.8</td>
<td>12.4</td>
</tr>
<tr>
<td>16</td>
<td>07.0</td>
<td>09.2</td>
<td>13.4</td>
</tr>
<tr>
<td>17</td>
<td>07.9</td>
<td>11.7</td>
<td>13.9</td>
</tr>
<tr>
<td>18</td>
<td>05.0</td>
<td>08.2</td>
<td>11.8</td>
</tr>
<tr>
<td>19</td>
<td>07.8</td>
<td>12.0</td>
<td>15.6</td>
</tr>
<tr>
<td>20</td>
<td>05.6</td>
<td>07.4</td>
<td>10.6</td>
</tr>
<tr>
<td>21</td>
<td>07.0</td>
<td>08.6</td>
<td>10.2</td>
</tr>
</tbody>
</table>

**Figure 2.30:** Listing of the aligned backbone atom RMSD between pairs of starting conformations. The minimum, average, and maximum over all ten possible pairs is shown for each protamine. RMSD values are in Angstroms.
if its backbone dihedral angles fell within a labeled bin. Propensities were defined as the average fraction of residues belonging to contiguous segments of length three or greater for alpha and P_{II} and two or greater for beta. These propensities were accumulated every 100 steps.

**Calculation of asphericities:** Average shapes of linear polymers take on specific values for coils, globules, and rods, respectively. For a specific conformation of a polymer, the gyration tensor is defined as:

\[
T = \frac{1}{Z_m} \sum_{i=1}^{Z_m} (\mathbf{r}_i - \bar{\mathbf{r}}) \otimes (\mathbf{r}_i - \bar{\mathbf{r}})
\]

(2.2)

Here, \(Z_m\) is the number of atoms in the molecule, \(\mathbf{r}_i\) are the position vectors of individual atoms, \(\bar{\mathbf{r}}\) is the position vector of the centroid, and the symbol \(\otimes\) refers to the dyadic product. If we use \(L_{1,2,3}^2\) to denote the eigenvalues of \(T\), the ensemble average value of asphericity, which measures the average shape of a polymer is given as [95]:

\[
\delta^* = 1 - 3 \left< \sum L_1^2 + L_2^2 + L_3^2 \right> \left( \sum L_1^2 + L_2^2 + L_3^2 \right)^2
\]

(2.3)

**Calculation of internal scaling profiles:** The ensemble average internal distances between residues \(i\) and \(j\) plotted as a function of sequence separation is also used as an order parameter in theories for globule-to-coil transitions of polymers [64]. This quantity is computed as follows:

\[
\left< R_{ij} \right> = \left< \frac{1}{Z_{ij}} \sum_{m \in i} \sum_{n \in j} |\mathbf{r}_m - \mathbf{r}_n| \right>
\]

(2.4)

Here, \(\mathbf{r}_m\) and \(\mathbf{r}_n\) denote the position vectors of atoms \(m\) and \(n\), which are part of residues \(i\) and \(j\), respectively, and \(Z_{ij}\) is the number of unique pairwise distances.
Figure 2.31: Operational definitions for local structure designations used in this work. A residue in a particular conformation was classified as having alpha, beta, polyproline II, or no local structure based on the location of its $\phi$ and $\psi$ dihedral angles within this Ramachandran map. The map is divided into $10^\circ \times 10^\circ$ squares.
between the two residues. For small separations in linear sequence, $|i - j| \leq 4$, $\langle R_{ij} \rangle$ is indistinguishable between different polypeptides. This reflects the constraints imposed by chain connectivity. Conversely, for longer separations in linear sequence, the scaling of $\langle R_{ij} \rangle$ with sequence spacing shows characteristic signatures that reveal the phase behavior of polypeptides. If the chain is in a globular state, then $\langle R_{ij} \rangle$ will achieve a plateau value that is consistent with the density of the globule. For non-globular phases such as random coils, self-avoiding random walks, and rod-like chains, $\langle R_{ij} \rangle$ will increase as $|j - i|^\nu$ with sequence spacing. Specifically, $\nu = 0.5$ for a Flory random coil [96], $\nu \approx 0.59$ for a self-avoiding random walk [31], and $\nu \approx 1$ for a rod-like polymer.

**Calculation of translational diffusion coefficients:** For each snapshot, the translational diffusion coefficient ($D$) was calculated using HYDROPRO version 7.C [68]. In these calculations, we used an atomic element radius of 1.2 Å, temperature of 298 K, solvent viscosity of 0.9 centipoise, and solvent density of 1.003 g/mL. Four values of the minibead radius in equal intervals from 0.75 Å to 1.2 Å were used for the extrapolation. 10,000 snapshots were analyzed to yield the ensemble average value for $D$ for each trajectory. The value of 1.2 Å for the atomic element radius was chosen because it causes the translational diffusion coefficient calculated by HYDROPRO for the free dye Rhodamine 6G to approximate the literature value of $4.14 \times 10^{-6}$ cm$^2$/s [97].

**Calculation of Kratky profiles:** The scattering form factor $P(q)$ for a single chain conformation as a function of scattering wave number $q$ is calculated using Equation 2.5:

$$
P(q) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j\neq i}^{N} \frac{\sin qR_{ij}}{qR_{ij}}
$$

(2.5)
Here, $N$ is the number of atoms, $q$ denotes wavenumbers, and $R_{ij}$ is the distance between atoms $i$ and $j$. For each protamine, the form factor was calculated for each of $10^4$ snapshots generated from the Monte Carlo simulations. The ensemble average form factor was used to compute the average Kratky profile $q^2P(q)$.

**Explicit solvent simulation design**

Simulations were performed using a version of GROMACS 3.3.1 [98] modified to include a harmonic restraining potential on the mass-weighted radius of gyration of the form $U_{\text{restrain}} = \frac{k}{2}(R_g - R_{g0})^2$. Eight umbrella sampling windows with the following values for $(R_{g0}$ in nm, $k$ in kJ/nm$^2$) were used at each salt concentration: (1.1, 1486.6), (1.2, 991.1), (1.4, 743.3), (1.6, 495.5), (1.8, 495.5), (2.0, 495.5), (2.2, 743.3), (2.3, 991.1). $R_{g0}$ values were chosen to cover the entire plausible range of $R_g$ for the (Arg)$_{25}$ polypeptide. Greater values for the force constant $k$ were used for windows near the ends of this range in order to overcome the expected natural tendency of the polypeptide to avoid adopting conformations with extreme $R_g$ values. This term of the Hamiltonian was in addition to the standard OPLS-AA/L force field with all bonds constrained to their equilibrium lengths. Lennard-Jones interactions were cut off at 1 nm. Particle mesh Ewald summation [99] with a real-space cutoff of 1 nm was used for computing long-range electrostatic interactions. Simulations were performed in the isothermal-isobaric ensemble using the Berendsen thermostat with protein and solvent separately coupled to a 298 K bath with a time constant of 0.2 ps and the Berendsen manostat [100] with ambient pressure of 1 atm, time constant of 1 ps, and compressibility of $4.5 \times 10^{-5}$ bar$^{-1}$.

One trajectory was produced for each umbrella sampling window at each salt concentration. For each trajectory, an initial conformation for the polypeptide was randomly drawn from the excluded volume ensemble, placed in a cubic box 10.5 nm
to a side and energy minimized in vacuum. Next, a pre-equilibrated cube of TIP3P water molecules was tiled over the simulation volume; we omitted water molecules that overlap with the protein, halting when exactly 38150 waters were successfully inserted. This solvated system was then energy minimized. The final preparation step was the transmutation of randomly selected water molecules into electrolyte ions: 25 waters were changed to neutralizing Cl\(^-\) counterions, then an additional 85, 337, or 663 pairs of waters were changed to NaCl to attain salt concentrations of 125 mm, 500 mm, or 1000 mm, respectively. Finally, molecular dynamics was performed for 55 ns using a 2 fs time step. The first five ns were discarded as equilibration, and snapshots were saved every picosecond. The weighted histogram analysis method (WHAM) [67] was used to reconstruct the Rg histograms from the umbrella-sampled trajectories. A single trajectory took about three months to produce on two cores of a 3.0 GHz dual-processor, dual-core Intel Xeon 5160 system with four megabytes of L2 cache and two GB of memory; eight windows were used for each salt concentration, and one 55 ns simulation was performed for each window. In contrast, each ABSINTH trajectory was generated on a single core of a 2.8 GHz dual-core AMD Opteron 254 system 1 MB of L2 cache and 1 GB of memory, and required between eight days for 125 mM and 24 days for 1000 mM salt concentration to complete; four trajectories were obtained for each salt concentration. The ABSINTH calculations for (Arg)\(_{25}\) were performed using methods similar to those for the shorter protamine studies; a total of 2.6 \times 10^7 Monte Carlo steps were performed, with the first 10^6 discarded as equilibration; the main difference being that the droplet radius was 60 Å (instead of 70 Å). Together with the increased sampling efficiency and the successful recapitulation of salt-dependent trends, the vastly reduced computational expense of ABSINTH compared to explicit solvent suggest that the use of ABSINTH is superior to simulations using explicit
representations of solvent for exploring the conformational ensembles of multiple, highly charged intrinsically disordered proteins.

### 2.3.3 Experiments

#### Peptide preparation

All peptides were purchased in crude form from Yale University’s Keck Biotechnology Center (New Haven, CT). The peptides were synthesized using solid-phase synthesis with a cysteine at the N-terminus to allow fluorescent labeling. The lone cysteine residue in protamine 4 was changed to a serine (which should be isosteric with cysteine) in the synthesized peptide to prevent improper fluorophore attachment. The crude peptides were purified by RP-HPLC on a semi-preparative C18 reverse-phase column (Agilent, Palo Alto, CA).

Upon purification, two sets of labeled peptides were generated for each sequence by chemical modification via a through-cysteine covalent attachment of either tetramethylrhodamine-5-maleimide (TMR) or AlexaFluor-488-C5-maleimide (Alexa) (Molecular Probes, Portland, OR). Freshly disaggregated peptides were reacted overnight with four-fold excess of dye at room temperature in 20 mM Hepes buffer [N-(2-hydroxy ethyl)piperazine-N-(2-ethanesulfonic acid)], pH 8.0 with 10 mM Tris-(2-carboxyethyl)phosphine. After the labeling reaction, the reaction mixture was loaded onto a size exclusion column with a cutoff of 1,400 Da (Sigma, St. Louis, MO) where they were eluted into pure water and stored at 4°C until use.

#### Fluorescence correlation spectroscopy (FCS)

Before carrying out the FCS measurements, peptide samples were diluted to a concentration of 50 nM in 125 mM NaCl dissolved in pure H$_2$O and adjusted to pH 2 using HCl. Data were collected at pH 2 to match the simulated systems, in which arginine sidechains are always protonated, as closely as possible. The sample pH was
checked before and after each experiment. Four hundred microliters of this solution was placed in a single well of an eight-chamber Nunc/Lab-Tek (Rochester, NY) 1.0 Borosilicate Coverglass System. Initial experiments showed a very high tendency for these peptides to stick nonspecifically to the cover glass. To alleviate this, all Nunc wells were treated for 15 minutes with concentrated solutions of 70 kDa polyarginine (Sigma, St. Louis) which was subsequently removed before the addition of the sample.

All measurements were performed on a Confocor II LSM system (Carl Zeiss-Evotec, Jena, Germany) with a 40X water-immersion objective. Data for fluorescence intensity autocorrelation functions were analyzed with Zeiss Confocor II FCS software. FCS measurements were performed using TMR- and Alexa-labeled peptides. The TMR-labeled samples were excited at 514 nm with an argon laser and emissions were collected in the 530 to 600 nm range. The Alexa-labeled samples were excited at 488 nm with an argon laser and emissions were collected in the 505 to 550 nm range. In all experiments, the laser power was allowed to stabilize for at least 30 min before beginning data collection to minimize any nonlinearity during startups.

For a given peptide sample, an independent measurement refers to a single $15 \times 15$ scan, which corresponds to the collection of FCS data 15 times where the duration for each data collection run was 15 seconds. Each scan yielded a distinct estimate for the diffusion time $\tau_D$ wherein the autocorrelation curves from all 15 experiments were averaged and the resultant curve was fit by using the model shown in Equation 2.6. We carried out five different $15 \times 15$ scans to obtain five independent estimates of $\tau_D$ for each of the peptide samples.

$$G(t) = 1 + \frac{1}{n} \left(1 - f_T + f_T \exp(-t/\tau_T)\right) \left[\frac{1}{(1 + \frac{t}{\tau_D})\sqrt{1 + \frac{t}{S^2\tau_D}}}\right] \quad (2.6)$$
In Equation 2.6, \( n \) is the average number of fluorescent molecules in the beam volume, \( f_T \) is the fraction of the triplet state formed per dye molecule, \( \tau_T \) is the decay constant of the triplet, \( S \) is a structure factor that describes the shape of the beam volume, and \( \tau_D \) is the translational diffusion time [101, 102]. \( S \) is a fixed parameter for an independent experiment, i.e., for a 15 × 15 scan. All other parameters were estimated by using a Levenberg-Marquardt nonlinear least-squares fit of the model to observed data. The parameters \( f_T \) and \( \tau_T \) are determined primarily by the photophysics of the fluorescent dye. As a result, the fitting procedure is deemed to be robust if \( f_T \) and \( \tau_T \) are essentially invariant with chain length. The measured values of \( \tau_D \) were used to estimate the translational diffusion coefficient \( D_{\text{protamine}} \) using the relation \( D_{\text{protamine}} = D_{R6G} \tau_{R6G} / \tau_D \). Here, \( D_{R6G} \) is the known translational diffusion coefficient for rhodamine 6G [97] and \( \tau_{R6G} \) is our measured translational diffusion time of rhodamine 6G.

**Fluorescence anisotropy**

These measurements were performed using TMR-labeled peptides. Before carrying out the anisotropy measurements, peptide samples were diluted to a concentration of 500 nM in 125 mM NaCl dissolved in pure H\(_2\)O and adjusted to pH 2 using HCl. The sample pH was checked before and after each experiment. Three milliliters of each sample was placed in a 1 cm path length quartz cuvette (Hellma, Plainview, NY) and the anisotropy measurements were not started until the total fluorescence had stabilized after non-specific adsorption to the glass. Fluorescence anisotropy measurements were performed with a PTI Quantamaster 40 L-format scanning spectrofluorometer (Birmingham, NJ) at a temperature of 25°C. The samples were excited at 515 nm and emission was monitored at 550 nm. Anisotropy (\( r \)
was calculated automatically by the instrument’s software (Felix) using Equation 2.7:

\[ r = \frac{I_{VV} - G I_{VH}}{I_{VV} + 2GI_{VH}} \]  

(2.7)

In Equation 2.7, \( I_{VV} \) is the light intensity with excitation and emission polarizers mounted vertically, \( I_{VH} \) is the light intensity with excitation polarizer mounted vertically and the emission polarizer mounted horizontally, and the G-Factor (\( G \)) is a system dependent correction factor which measures the emission channel’s efficiency at detecting horizontally polarized light in reference to vertically polarized light. The G-factor was determined before measuring each sample and was essentially invariant. The average anisotropies reported correspond to the average of 15 independent measurements and the error bars shown in Figure 2.10 are standard errors of the mean.

### 2.4 References


95


96
Chapter 3

Counterion-mediated collapse of acid-rich intrinsically disordered proteins in simulations depends on salt ion models

The speculative phase diagram of Figure 2.28 described in Section 2.2.13 is a mathematical mapping from the set of protein sequences to their presumed polymeric phase behavior. Its parametrization of the combinatorially immense space of protein sequences with only three descriptors ($f_+, f_-$, and hydropathy) constitutes a significant reduction in complexity. The globule-to-coil transition that defines the mapping itself is parametrized by a single quantity, the net charge per residue, and provides a coarse answer to the overarching question of how primary structure determines attributes of IDP conformational ensembles. Although this simple picture is consistent with polyelectrolyte theories [1], its accuracy needs to be tested outside the domain that guided its construction. The ease of locating an arbitrary sequence within the phase diagram enables rapid prediction of whether it will adopt a globule or coil-like conformational ensemble. In this study, we use molecular simulations in ABSINTH implicit solvent to explore the phase diagram and test its predictions across a larger and broader variety of IDP sequences.

We employed two complementary approaches for selecting IDP sequences to serve as test cases. The first approach expands upon the polyampholyte exploration
described in Section 2.2.10 by sampling from DisProt [2], a curated database of intrinsically disordered proteins. It is intended to cover a broad range of IDPs that exist in vivo and generate insight across the disordered proteome. The second approach uses synthetic polyglutamate and polyaspartate IDPs to probe a corner of the phase diagram that is distant from DisProt sequences and explore the impact of solution conditions on conformational ensembles. Polyelectrolyte theories that predicted expansion of protamines make identical predictions when all charges are replaced by ones with opposite sign [1]. These acid-rich IDPs test whether this charge conjugation symmetry holds at the level of amino acids when negatively charged acidic residues are substituted for positively charged basic ones.

3.1 Intrinsically disordered regions modeled in isolation

Although some IDPs are completely disordered [3], most consist of one or more intrinsically disordered regions (IDRs) connected to and interspersed amongst folded regions [4, 5]. In these hybrid arrangements, the proportion of disordered regions spans a broad spectrum of possibilities. These take the form of mostly folded proteins with frayed ends or internal loops missing from the crystal structure, multiple folded domains connected by a flexible linker, long disordered tails attached to a folded body, and mostly disordered proteins with short segments that exhibit transient secondary structure. The question of how folded regions of IDPs modulate the intrinsic behavior of their attached disordered regions is an ongoing subject of inquiry. Conversely, it is also possible that removing disordered regions would influence the structure and stability of adjacent folded regions in certain situations. To focus this investigation on the intrinsic polymeric characteristics of disordered proteins and keep the simulations tractable, we excise the IDRs from their folded context and model them as isolated
entities in solution.

We select intrinsically disordered regions from DisProt using a procedure designed to achieve uniform coverage of the phase diagram. Version 4.9 of the database, which contains 523 proteins, was downloaded in its entirety as a XML file and parsed to extract the 1195 disordered regions. Sequences with fewer than 20 residues, greater than 55 residues, more than two cysteine residues, or any non-standard amino acids are filtered out, leaving 246 candidate IDRs. These candidates are processed in order of increasing length by adding the candidate to the final set if its Euclidean distance to every other previously added member of the set in \((f_+, f_-, \text{ hydropathy})\)-space is at least 0.04 units. This algorithm eliminates redundancy in sampling locations on the phase diagram while favoring shorter sequences over longer ones. We simulate the first 110 IDRs yielded by this procedure.

### 3.1.1 Simulation methods

A protocol similar to the one described in Section 2.3.2 is used for simulating all intrinsically disordered regions. Each sequence is capped with acetyl and N-methylamide groups at the N and C-termini, respectively, and placed inside a 100 Å radius spherical droplet along with explicitly represented K\(^+\) and Cl\(^-\) ions sufficient to neutralize the net polypeptide charge and mimic a 125 mM salt solution. Five independent replicas of each sequence are simulated using randomly generated starting conformations taken from the excluded volume ensemble [6]. To provide reference states, one full-length simulation of each sequence without any ions is performed using both the excluded volume Hamiltonian and a Hamiltonian with only the Lennard-Jones terms enabled. Histidine sidechains are singly protonated at their \(\epsilon\) nitrogen, which makes the sidechain uncharged. The ABSINTH implicit solvation, electrostatic, Lennard-Jones, and Engh-Huber crystallographic geometry parameters described in
Vitalis and Pappu [7] are used with one adjustment: the free energies of solvation for model compounds comprising ionic sidechains are restored to $-70$ kcal/mol from their artificially lowered values that were intended to prevent salt bridging. Cutoffs for the Lennard-Jones and electrostatic interactions between neutral groups are set at 10 and 14 Å, respectively. Electrostatic interactions between ions in solution and sidechain moieties with an overall net charge are computed without employing cutoffs. Metropolis Markov chain Monte Carlo (MC) simulations are performed in the canonical ensemble at 298 K for a total of $6 \times 10^7$ steps, with the first $10^7$ steps treated as equilibration and omitted from analysis. The move set is identical to the one described in Section 2.3.2 and includes the cluster rigid body moves.

### 3.1.2 Disordered regions are typically short and have moderate net charge per residue

The length distribution of IDRs from the DisProt database is shown in Figure 3.1 and reveals that most disordered regions are short, with few having lengths in excess of 100 residues. If the entries in DisProt are a representative sample, this would imply that IDRs generally contain fewer residues than folded protein domains [8]. Therefore, our restriction to IDRs with 55 or fewer residues does not constitute a severe distortion of the population; as long as disordered regions are modeled in isolation, simulations of the majority of IDRs are computationally tractable. However, only 18% of the 523 IDPs are fully disordered, and 25% have at least half of their residues within a disordered region. Therefore, it is possible that modulation of IDR conformational ensembles by their folded context is a common phenomenon throughout the disordered proteome. Since the simulations model IDRs in isolation, these context-dependent effects are not captured by the simulations.

Figure 3.2 shows the position of all 110 simulated IDRs within the phase
Figure 3.1: Length histogram of intrinsically disordered regions from DisProt. Less than 0.5% of the IDRs have length greater than 1000 residues. The bin width is 5 residues.

diagram. Most IDRs exhibit moderate values of $f_+$ and $f_-$, with the majority falling within the region predicted to be disordered globules. Relative to this set, the base-rich protamines studied in Chapter 2 are outliers, as few IDRs approach the corners of the phase diagram that correspond to high values of the net charge per residue.

3.1.3 Most IDRs adopt ensembles of collapsed, globular conformations

Simulations of all 110 IDRs as isolated entities suggests that the disordered proteome is largely characterized by collapsed, globular ensembles. This “medium-throughput” approach is practical because of the efficiency offered by the ABSINTH implicit solvation model; a comparable level of conformational sampling for this num-
**Figure 3.2:** Phase diagram annotated with IDRs from DisProt. The boundaries and regions are identical to those of Figure 2.28.
ber of polypeptides in explicit solvent would have been unattainable using available computational resources. Following the precedent of Chapter 2, we present the ensemble average radius of gyration $\langle R_g \rangle$ normalized by its excluded volume value as a readout of polymeric phase behavior plotted against the net charge per residue $f_+ - f_-$. Figure 3.3 shows that most of these IDRs exhibit a normalized $\langle R_g \rangle$ consistent with a collapsed, globular conformational ensemble. For IDRs with low absolute net charge per residue such that $|f_+ - f_-| < 0.25$, this agrees with the results from Chapter 2 and matches the behavior predicted by their position on the phase diagram. The tendency towards increased average size for IDRs with $f_+ - f_- > 0.25$ is also consistent with predictions, though the lack of sequences with net charge per residue significantly exceeding 0.4 prevents this region of the phase diagram from being thoroughly tested. However, this trend does not hold when the net charge per residue is titrated in the opposite direction: IDRs with $f_+ - f_- < -0.25$ exhibit no swelling and appear to contradict predictions from the phase diagram. Therefore, the results of this broad set of simulations imply that negatively charged acidic residues and positively charged basic residues are asymmetric in their ability to induce electrostatic swelling of a polypeptide chain.

The expanded set of sequences provides additional examples of the findings described in Section 2.2.3 where $\langle R_g \rangle$ is relatively insensitive to sequence details. The variation in $\langle R_g \rangle$ for the multiple sequences with $f_+ - f_- \text{near any one value between} -0.25 \text{and} 0.25 \text{demonstrates the perturbative effect of varying sequence details while keeping net charge per residue constant. As before, perturbations generally do not increase the normalized $\langle R_g \rangle$ above 0.6 and therefore do not change the polymeric classification of any IDRs from globule to swollen coil. Interestingly, the actual ordering of $\langle R_g \rangle$ at each $f_+ - f_-$ is uncorrelated to the specific values of $f_+$ and $f_-$ as}
Figure 3.3: Normalized radius of gyration versus net charge per residue for IDRs from DisProt. Each point represents one IDR and is colored according to its charge composition. The inset contains a legend that shows how color varies as a function of $f_+$ and $f_-$. indicated by plot symbol color. This suggests that the magnitude of perturbations around a disordered globular state is actually sensitive to sequence details and not merely a function of $f_+$ or $f_-$. The observed asymmetry between moderately charged IDRs of opposite polarity is intriguing because it violates expectations based on both experiments and polymer physics theories. Polymer theories are based on abstract polyelectrolyte chains and do not distinguish between positive and negative charge [1, 9]; they yield identical predictions if the signs of all charges are reversed. Our speculation that the
boundary between disordered globules and swollen coils for acidic sequences is the mirror image of the boundary for basic sequences is motivated by this symmetry. In contrast to their theoretical counterparts, the charged sidechains of real polypeptides need not exhibit charge conjugation symmetry because aspartate and glutamate are structurally and chemically distinct from charge-negated versions of lysine and arginine. In addition, co-ions and counterions in solution also differ in characteristics besides having opposite charges. These differences should be captured by the atomistic representation of polypeptides and ions used in our simulations, and could explain the finding of asymmetry and deviation from phase diagram predictions. This higher degree of physical realism provided by atomistically detailed models is an advantage of molecular simulations compared to analytically solvable theories. However, single-molecule Förster resonance energy transfer efficiency histograms for highly acidic IDPs are consistent with swollen coil ensembles [10], a finding that supports theories and phase diagram predictions while contradicting the simulation results. This motivates an investigation into the reason that simulated acid-rich IDRs remain collapsed in ABSINTH implicit water combined with explicit ions.

3.1.4 Acidic and basic sidechains exhibit distinct levels of counterion accumulation

Visual inspections of simulation trajectories revealed a qualitatively greater and more persistent accumulation of counterions around acid-rich IDRs compared to base-rich IDRs. Figure 3.4 is an arbitrarily selected sample from the ensemble for the IDR with sequence PEEKKEEGSANRRPDEDQELESLSA that illustrates this tendency. To quantify this tendency, we calculate distance histograms between sidechain tips and counterions. When compared to an appropriate reference, such histograms reveal the degree of counterion condensation around charged sidechain moieties. Similar to the
analysis described in Section 2.2.10, we define the sidechain tip to be the carbon atom farthest away from the polypeptide backbone. This atom is the carboxylate carbon in aspartate and glutamate, the ε-carbon in lysine, and the ζ-carbon in arginine, all of which occupy positions which are representative of their sidechain’s charge distribution. Figure 3.5 shows the pooled distance histograms across all 110 IDRs, and quantitatively support the observation that potassium cations exhibit increased accumulation around acidic sidechain tips compared to chloride anions around basic sidechain tips.

Pooling the data in this way provides a broad overview of general tendencies; we also examine IDRs individually to gain insight into the diversity of their ion accumulation behavior. Figure 3.6 plots the ion-sidechain pair correlation function \( g(r) \) for two representative sequences, PEEKKEEGSANRPERPQELSLSA and AARKEVIRN KIRAIGKMARVFSVLR. In these plots, the ion-to-neutral sidechain tip distributions have been used as a prior distribution in order to produce normalized pair correlation functions \( g(r) \) for each sidechain-ion combination. The differing vertical scales of the plots show that while chloride anion concentrations do exhibit relative enhancement around basic sidechain tips, the degree of accumulation is much less pronounced than that of potassium cations around acidic sidechain tips. These examples also illustrate a sequence-specific effect, where basic residues in close sequence proximity to acidic residues can exhibit remote accumulation of cations because they are constrained to be near the cation-attracting acidic sidechains.

Differential accumulation of counterions around acid versus base-rich IDRs would have implications for understanding their interactions with other biological macromolecules. The milder association of anions around basic residues is consistent with the salt concentration dependence of binding affinities between nucleic acids and
Figure 3.4: Example conformation showing collapse of PEEKEEEGSANRRPEDQELESLSA with associated K\(^+\) ions. The rendering is produced using Visual Molecular Dynamics (VMD) [11] and includes all ions within 5 Å of any protein atom. K\(^+\) ions are colored green and Cl\(^-\) ions are colored cyan; the absence of any cyan spheres implies no Cl\(^-\) ions are within 5 Å of any protein atom in this sample from the ensemble. As an indicator of scale, VMD renders K\(^+\) ions as 3 Å diameter spheres. In rendering the protein, carbon, oxygen, and nitrogen atoms are respectively colored cyan, red, and blue, while hydrogen is hidden. In addition, a smoothly curved spline highlights the course of the protein backbone and changes from red at the N-terminus to blue at the C-terminus.
Figure 3.5: Pooled ion-sidechain distance histograms for all IDRs from DisProt. Each histogram is individually normalized to enable meaningful comparisons despite unequal numbers of K$^+$ and Cl$^-$ ions as well as acidic and basic sidechains. The raw probabilities are not divided by any prior distribution and therefore exhibit the bell-shaped curve characteristic of two points randomly selected from the interior of a sphere. The bin width is 0.2 Å.
Figure 3.6: Individual ion-sidechain pair correlation functions for two example IDRs. To calculate \( g(r) \), ion-charged sidechain histogram probabilities are divided by raw ion-neutral sidechain histogram probabilities. Before division, each histogram is individually normalized to account for unequal numbers of charged and neutral sidechains. The bin width is 0.2 Å.
base-rich polypeptides, which reveals that anion release from polypeptide is negligible compared to cation release by polymnucleotide [12]. This is pertinent to the large variety of base-rich IDRs that bind nucleic acids [13], as sparse inclusion and positioning of acidic residues could modulate their affinity both directly and indirectly via counterion release.

3.2 Synthetic polyelectrolyte IDPs

We investigate the asymmetry between acid-rich and base-rich IDRs by asking if it persists in the limit of strong polyelectrolytes. As described in Section 3.1.2, the simulated IDRs from DisProt mostly lie near the center of the phase diagram and do not exhibit extreme values for the net charge per residue. Therefore, we use synthetic polypeptides that are polymers of either glutamate or aspartate to attain the polyelectrolyte limit. These proteins, which are uniformly predicted to be intrinsically disordered, have $f_+ - f_- = -1$ and are predicted to have the greatest degree of electrostatically-induced expansion by polyelectrolyte theories.

3.2.1 Polyglutamate and polyaspartate collapse while entraining counterions in simulations

Simulations of polyglutamate and polyaspartate 34 residues in length reveal that the asymmetry persists even in the polyelectrolyte limit. We choose a length of 34 to maintain consistency with the polyarginine sequence simulated in Chapter 2. The extent of collapse and counterion accumulation appear to be even more severe compared to the acid-rich IDRs. As shown in Figure 3.7, an arbitrarily selected conformation from the simulated ensemble for $(\text{Glu})_{34}$, potassium counterions are clustered so tightly around glutamate sidechains that they lie in close proximity to each other. In contrast, the chloride anions are completely excluded from the vicinity of the polypeptide.
**Figure 3.7:** Example conformation showing collapse of (Glu)$_{34}$ with entrained counterions. The rendering method is identical to the one used for Figure 3.4. The slightly reduced zoom level is evident from the apparent size of $K^+$ ions, which are still drawn as 3 Å diameter spheres.
This finding is reproducible across independent trajectories initiated from different random starting conformations. However, numerous close contacts between counterions and polypeptide hinder convergence of polyglutamate and polyaspartate simulations because the Monte Carlo sampling moves are generally unable to alter the system’s degrees of freedom in concert to avoid steric clashes. Consequently, the system fails to equilibrate even after \( 6 \times 10^7 \) Monte Carlo steps, preventing quantitative analysis of the conformational ensemble. For comparison, Figure 3.8 shows an example conformation from the simulated ensemble for \((\text{Arg})_{34}\) in KCl, which remains extended. While co-ions are excluded in favor of a loose shell of Cl\(^-\) counterions, the enhancement in the local counterion concentration is much less severe, and the sampling moves remain effective.

The extent to which like-charged counterions are packed together in Figure 3.7 challenges physical intuition and warrants scrutiny of the model and parameters used to represent ions in the simulation. In addition, the polyglutamate and polyaspartate sequences studied here resemble human prothymosin-\(\alpha\), an IDP that Müller-Späh et al. found [10] to be extended and distinct from collapsed IDPs with low net charge per residue. Therefore, a detailed investigation of the phenomenon’s sensitivity to the details of the ion representation is warranted.

3.2.2 Development of a simplified symmetric model for salt ions in ABSINTH

To probe the mechanism underlying asymmetric behavior of acid-rich and base-rich IDPs in the presence of counterions, we remove one aspect of asymmetry from the system. As described in Section 3.1.3, salt ions and titratable sidechains are distinct from their charge conjugated versions. We replace KCl with a simplified, symmetric electrolyte to determine if the asymmetry in counterion-mediated collapse
Figure 3.8: Example conformation showing \((\text{Arg})_{34}\) with associated counterions. The rendering method is identical to the one used for Figure 3.4, although the zoom level is significantly reduced to accommodate the enhanced size of the conformation. In this conformation, it is the \(\text{K}^+\) cations that are excluded from the vicinity of the protein. \(\text{Cl}^-\) ions are drawn as cyan colored spheres 3 Å in diameter.
between polyaspartate/polyglutamate and polyarginine/polylysine persists. This isolates the effects of sidechain structure and chemistry by eliminating any physical distinction other than opposite charge between the salt ions. Since truly symmetric electrolytes do not exist in nature, this is a controlled mechanistic experiment that can only be performed in silico.

To create the simplest possible symmetric electrolyte within the context of ABSINTH, we modify the simulation model for ions to remove attractive dispersion interactions from the Lennard-Jones potential. This modification only applies to atomic ions; polypeptide atoms still interact with each other through the standard ABSINTH energy function. The Lennard-Jones $\epsilon$ parameter, which is now redundant with the diameter $\sigma$ since only the inverse power potential term is preserved, is set to 1 kcal/mol for all atomic ions. We also eliminate atomic radius overrides, causing solvation shell overlap calculations to use the same diameter $\sigma$ that determines an ion’s excluded volume interaction. Finally, we constrain each ion’s solvation free energy to the value that results from adding the cavitation penalty of Pierotti’s scaled particle theory [14] to the favorable Born energy of charging a sphere embedded in a medium with dielectric constant 78.2. This solvation free energy, which is dominated by the Born energy, is shown as a function of ion diameter in Figure 3.9. In this simplified model, an ion’s diameter alone completely specifies its interactions, and its only short range interaction is a stiff repulsive steric exclusion. We therefore refer to it as the ABSINTH restricted primitive model (ARPM).

3.2.3 Polyelectrolyte IDPs simulated with simplified salt ions do not collapse

Using ARPM ions with diameters from 1 to 10 Å at concentrations ranging from 50 to 250 mM, we simulate 34-residue sequences of polyarginine, polylysine,
Figure 3.9: Reference hydration free energy as a function of diameter for ABSINTH restricted primitive model ions. For comparison, the original values used for Na\(^+\), K\(^+\), and Cl\(^-\) in previous ABSINTH studies [7] are shown using their Lennard-Jones \(\sigma\) as the diameter.

polyaspartate, and polyglutamte as a broad search for solution conditions where asymmetric counterion-mediated collapse persists. Figures 3.10 and 3.11 plot \(\langle R_g \rangle\) as a function of both ion diameter and concentration for each of these synthetic polyelectrolyte IDPs. \(\langle R_g \rangle\) is left unnormalized because all polypeptides are the same length, enabling direct comparison of their numerical values. Compared with the reference ensemble \(\langle R_g \rangle\) values of \(\sim 8\ \text{Å}\) for globules and \(\sim 19\ \text{Å}\) for coils, it is evident that polyelectrolyte IPDs simulated with ARPM ions are extended coils in every solution condition tested. Although the degree of extension is least for ion diameters around

115
3 Å and shows the expected trend of decreasing with increasing salt concentration, all of the \( \langle R_g \rangle \) values fall within a narrow range that is well above even the swollen random coil value. The lack of asymmetric counterion-mediated collapse in these symmetric ARPM simulations suggests that the difference between acidic and basic sidechains is insufficient to explain the phenomenon of asymmetry and points towards the salt ions as its mechanistic origin.

Analysis of ion-sidechain distance histograms corroborate this conclusion. We perform simulations of shorter 17-residue polyaspartate and polyglutamate sequences with 150 mM symmetric 5 Å ARPM ions in a 70 Å radius droplet, and compare them to control simulations with KCl using original ABSINTH parameters. Similar to the analysis presented in Section 3.1.4, we analyze the accumulation of ions around sidechains by calculating ion-sidechain distance histograms. Figures 3.12 and 3.13 show these histograms for polyaspartate and polyglutamate, respectively. The ARPM histograms once again exhibit accumulation of counterions around aspartate acidic sidechain tips, but the accumulation is diffuse compared to that of the control simulations.

3.3 Discussion

3.3.1 Solution conditions modulate the conformational ensembles of IDPs

These results show that solution conditions can strongly influence the conformational ensembles of intrinsically disordered proteins. Although it is apparent that the simulated entrainment of K\(^+\) ions by acid-rich IDRs described in Section 3.2.1 is actually an artifact of the original simulation model, the different behavior between the original model and ABSINTH RPM may also be interpreted as evidence that some 1:1 electrolyte solutions induce collapse of acidic IDRs while others do not.
Figure 3.10: Radius of gyration for acidic synthetic polyelectrolytes in ARPM solution. For comparison, \( \langle R_g \rangle \) for globule/coil reference states is \( 7.8 / 18.2 \) Å for polyaspartate and \( 8.1 / 18.9 \) Å for polyglutamate. Therefore, the entire range of values lies above swollen coil limit. Error bars denote the standard error of the mean (SEM).
Figure 3.11: Radius of gyration for basic synthetic polyelectrolytes in ARPM solution. Data for polyarginine gathered at 50 and 100 mM are qualitatively similar, and are not shown. For comparison, \( <R_g> \) for polylysine globule and coil reference states are 9.1 Å and 19.7 Å, respectively. Error bars denote the standard deviation of the normalized \( <R_g> \) obtained from the five independent replicate simulations. Since each individual replicate itself constitutes a large sample from a population, we interpret this standard deviation as the standard error of the mean (SEM).
Figure 3.12: Ion-sidechain distance histograms for \((\text{Asp})_{17}\). As with Figure 3.5, each histogram is individually normalized. The bin width is 0.2 Å.
Figure 3.13: Ion-sidechain distance histograms for (Glu)$_{17}$. Each histogram is individually normalized. The bin width is 0.2 Å.
Therefore, the primary structure of an IDP is insufficient to fully determine even a coarse polymeric classification of its conformational ensemble; the same IDP can exhibit distinct ensembles under different solution conditions. This finding has implications for understanding IDP function on a molecular level, as local intracellular solution conditions can fluctuate in response to external stimuli [15]. It also suggests the possibility of engineering IDPs as molecular sensors: by attaching fluorescent probes and quenchers to opposite ends of the polypeptide, a change in solution conditions could be transduced as an observable change in fluorescence signal via the collapse transition. Such engineering efforts will require a quantitative understanding of the interplay between solution conditions and primary structure in determining conformational ensembles that goes beyond the qualitative picture provided here.

Anecdotal experimental evidence suggests that base-rich IDPs can also exhibit anion-dependent precipitation. During early stages of sample preparation for the fluorescence experiments of Chapter 2, Crick observed that addition of phosphate buffer to a polyarginine solution resulted in formation of a cloudy white precipitate. In contrast, no precipitation occurred upon addition of 125 mM NaCl even though the pH was titrated to match that of the phosphate solution using HCl and NaOH. This points toward an effect for IDPs analogous to the preferential salting out of folded proteins with ions according to their position within the Hofmeister series. It is also consistent with the role of protamines in condensing phosphate-rich DNA during spermatogenesis [16, 17]. The entrainment of counterions and collapse of acid-rich IDRsof the original ABSINTH salt ion representation may also constitute an ion-dependent salting out effect. However, simulations of more than one molecule at a time would be necessary to justify and quantify the connection between microscopic and macroscopic observations of counterion-mediated precipitation.
pH is another solution condition that is likely to influence conformational ensembles of IDPs. As explained in Section 2.3.2, the fixed protonation state of ionizable sidechains within simulations is an approximation that may detract from their accuracy. This is especially pertinent for IDPs whose sequences are enriched in ionizable sidechains. α-helix propensity data from simulations and experiments gathered by Das et al. [18] for the intrinsically disordered basic regions (bR) of bZIP transcription factors suggest that this approximation is valid for many, but not all, IDPs. Circular dichroism (CD) measurements supported simulation results for eleven out of thirteen bZIP-bRs, which included both wild-type and chimeric sequences. The pH dependence of CD spectra suggested that modulation of average protonation states induced by close proximity of acidic and basic sidechains accounted for the discrepancy between simulation and experiment. In addition, simulation accuracy for the eleven sequences was contingent upon using lowered values for sidechain reference hydration free energies of $-107.3$ kcal/mol for aspartate and glutamate and $-100.9$ kcal/mol for lysine and arginine; the simulations exhibited negligible helix propensity if the experimentally indicated value of $-70$ kcal/mol was used instead. Going forward, this motivates the artificially lowered reference hydration free energies for ionizable sidechains as a mitigating factor until a true constant pH simulation engine that implements ABSINTH is available.

3.3.2 Can the original salt ion parameters used in ABSINTH be improved?

Each alkali and halide ion requires the specification of five parameters within ABSINTH. One of these parameters is the electrostatic charge, which is exactly $\pm e$. Lennard-Jones parameters $\sigma$ and $\epsilon$, which respectively describe the length and energy scales of an ion’s short-range van der Waals interactions, are general in that they are
also required for other simulation models, including explicit solvent. The other two parameters, which are the reference solvation free energy and atomic radius, are specific to ABSINTH and intended to be directly determined by experimental data. The simulations presented in this chapter, Chapter 2, and other studies using salt ions with ABSINTH implicit solvent [7, 19–21] have adopted Lennard-Jones parameters for alkali and halide ions from the Optimal Potentials for Liquid Simulations — All Atom (OPLS-AA) force field [22–24]. In this force field, cations including Na\(^+\) and K\(^+\) have Lennard-Jones parameters adapted from a classic study by Åqvist, who fit them to reproduce hydration free energies and ion-oxygen radial distribution function peak distances with the single point charge (SPC) water model [25]. In contrast, anions such as Cl\(^-\) were parametrized to reproduce gas phase ion-water binding energies and geometries [26] from Hartree-Fock ab initio calculations.

The physical justification for using these legacy Lennard-Jones ion parameters in the context of ABSINTH implicit solvent is weak. Since calculated hydration free energies are sensitive to the water model [27], parameters fit to hydration free energies are best suited for simulations that use the same water model as their calibration. The SPC water model is a simple approximation with known limitations that are entirely distinct from the tradeoffs made by ABSINTH, making it unlikely that ion parameters developed in conjunction with SPC water are well-suited for ABSINTH. In addition, measurements of hydration free energies have improved [28] since the date of Åqvist’s calibration. More significantly, OPLS-AA and its parameters are intended for use with geometric mixing rules, while ABSINTH consistently employs Lorentz-Berthelot mixing rules. The mismatch between intended and employed mixing rules for a set of parameters has been shown to be responsible for significant artifacts [29], including the formation of salt crystals at concentrations well below
the solubility limit. Despite these concerns, ions have not been a source of error in previous ABSINTH simulations, possibly due to the low concentrations employed and avoidance of highly acidic polypeptides in those studies.

The phenomenon of asymmetric precipitation encountered in this chapter and its sensitivity to the simulated ion model motivate the development of an improved set of Lennard-Jones parameters for alkali and halide ions. Ideally, these new parameters should be calibrated in a way that makes them transferable and robust rather than being limited to the specific situation used in their calibration. In addition to addressing the theoretical concerns regarding applicability of parameters from other force fields in ABSINTH, a transferable set of ion parameters would be generally useful to the practice of molecular simulation. The prevalence of acid-rich IDPs, the known sensitivity of nucleic acids and their binding properties to solution conditions, and the ubiquity of electrolyte solutions make this effort an important and necessary next step.

3.4 References


Crystal lattice properties fully determine short-range interaction parameters for alkali and halide ions

Accurate models of alkali and halide ions in aqueous solution are necessary for computer simulations of a broad variety of systems. Previous efforts to develop ion force fields have generally focused on reproducing experimental measurements of aqueous solution properties such as hydration free energies and ion-water distribution functions. This dependency limits transferability of the resulting parameters because of the variety and known limitations of water models. We present a solvent-independent approach to calibrating ion parameters based exclusively on crystal lattice properties. Our procedure relies on minimization of lattice sums to calculate lattice energies and interionic distances instead of equilibrium ensemble simulations of dense fluids. The gain in computational efficiency enables simultaneous optimization of all parameters for Li\(^+\), Na\(^+\), K\(^+\), Rb\(^+\), Cs\(^+\), F\(^-\), Cl\(^-\), Br\(^-\), and I\(^-\) subject to constraints that enforce consistency with periodic table trends. We demonstrate the method by presenting lattice-derived parameters for the primitive model and the Lennard-Jones model with Lorentz-Berthelot mixing rules. The resulting parameters successfully reproduce the lattice properties used to derive them and are free from the influence of any water model. To assess the transferability of the Lennard-Jones
parameters to aqueous systems, we used them to estimate hydration free energies and found that the results were in quantitative agreement with experimentally measured values. These lattice-derived parameters are applicable in simulations where coupling of ion parameters to a particular solvent model is undesirable. The simplicity and low computational demands of the calibration procedure make it suitable for parametrization of crystallizable ions in a variety of force fields.

This chapter is adapted from an article [1] published in the Journal of Chemical Physics. Rohit V. Pappu suggested the approach of using lattice energies in combination with lattice constants to fix both energy and length scales. The candidate, Albert H. Mao, designed and implemented the calibration procedure, used it to obtain the presented parameters, and wrote the paper with assistance from Rohit V. Pappu. The article may be accessed via its digital object identifier (DOI) name 10.1063/1.4742068. This work was supported by National Science Foundation MCB 0718924 and MCB 1121867.

4.1 Introduction

Alkali and halide ions play important roles in biological and physico-chemical systems that include protein [2–5], nucleic acid [6–11], lipid [12], and carbohydrate [13] solutions [14], salt crystals [15], molten salts [16], electrolytes [17], and liquid-vapor interfaces [18]. Computer simulations are useful for developing a molecular scale description and understanding of electrolyte dependencies and ion-mediated interactions in these systems. Most classical simulation approaches to modeling alkali and halide ions employ the Born-Oppenheimer approximation where the ions are hard spheres or van der Waals spheres with a charge of ±e. van der Waals interactions are commonly modeled using the empirical Lennard-Jones 12-6 potential. These models for ions are used with either explicit [19] or continuum [20–23] (implicit) descriptions.
of the surrounding solvent. In order to achieve accuracy in simulation results, one needs reliable hard sphere or van der Waals parameters.

Experiments that measure structural and excess thermodynamic properties of electrolyte solutions can provide constraints for the calibration of these parameters. Numerous collections of parameters have been developed for ions in aqueous solutions where the relevant constraints come from gas phase ion-water binding energies and geometries [24–27], hydration free energies [25–29] and entropies [29], structural properties such as water-ion pair distribution functions [26, 28–30], and transport properties regarding the degree of hydration and ion association [31]. The resulting parameters are intricately dependent on the water model used in the parameterization procedure. This dependence is awkward because water models themselves are more complicated and require more parameters than any one alkali or halide ion. The need for transferable and generally applicable parameters is prominent in applications such as biomolecular simulation, where matter besides water and ions is present. Given the sheer number, diversity, and known limitations [32] of available water models, it is difficult to be confident that ion parameters derived using a particular water model reflect the intrinsic properties of the ions that are also transferable for use in a specific simulation system.

The interionic distance and lattice energy of alkali halide salt crystals are properties that do not require any consideration of the specific model used for solvent molecules. They constitute a set of measurable observables that constrain the length and energy scales for van der Waals interactions of non-polarizable alkali and halide ions. In this work, we simultaneously obtain values for the sphere diameter ($\sigma$) and well depth ($\epsilon$) parameters of five alkali cations (Li$^+$, Na$^+$, K$^+$, Rb$^+$, Cs$^+$) and four halide anions (F$^-$, Cl$^-$, Br$^-$, I$^-$). These parameters, which are designed for use with a
hard sphere or Lennard-Jones 12-6 potential based on Lorentz-Berthelot mixing rules, were obtained using lattice properties as the only calibration targets and are therefore independent of any specific water or solvent model. Our calibration procedure relies on minimization of lattice sums to compute the lattice energies and interionic distances. It requires modest computational resources compared to approaches involving explicit construction and simulation of periodic crystals or dense fluids. We assess the transferability of the derived parameters by determining minimum energy lattice configurations and testing the accuracy of single-ion hydration free energies across three water models estimated using bicubic surfaces constructed by Joung and Cheatham [27].

4.2 Methods

The fitting is accomplished through minimization of a calibration objective function that maps any candidate parameter set to one real number quantifying deviation from experimental measurements of lattice observables. Each evaluation of the objective function itself involves minimization of every salt’s parameter-dependent potential energy to calculate its lattice energy and interionic distance. Since this minimization does not account for thermal fluctuations, it is a ground state calculation suitable for comparison with experimental lattice energies and interionic distances measured at absolute zero. The following sections describe these steps in detail.

4.2.1 Calibration targets

Experimental data for the twenty alkali halide salts arising from combinations of Li\(^+\), Na\(^+\), K\(^+\), Rb\(^+\), or Cs\(^+\) with F\(^-\), Cl\(^-\), Br\(^-\), or I\(^-\) form the basis of our calibration. All of these salts form cubic crystals whose structures are described by one interionic distance (ID), which is defined as the distance between centers of two nearest-neighbor ions of opposite charge. The cations and anions are arranged in inter-
penetrating simple cubic lattices (denoted BCC, since the unit cell is body-centered) for CsCl, CsBr, and CsI and interpenetrating face centered cubic (FCC) lattices for the other seventeen salts. Sirdeshmukh et al. have compiled X-ray diffraction measurements of interionic distances [33], while Jenkins and Roobottom have gathered lattice energy (LE) measurements [34] derived through the Born-Fajans-Haber thermochemical correlation [35]. Ghate [36] extrapolated interionic distances to 0 K, a reduction of $\sim 1\%$ relative to room temperature values. The lattice energies have even lower temperature sensitivity, changing by less than 0.1% from room temperature to 0 K [37]. Therefore, we treat all parameters as temperature-independent quantities and adopt the extrapolated interionic distances and room temperature lattice energies, listed in Table 4.1, as our calibration targets because they closely approximate the 0 K values that result from minimizing the potential energy of a crystal lattice. Since there are forty independent measurements and nine ions, models with four or fewer parameters per ion are overdetermined, which is a desirable characteristic in discouraging overfitting and promoting transferability.

### 4.2.2 Ion interaction models

We focus on two common models for ions in molecular simulations. Both models are pairwise additive potentials where a short-range interaction is superimposed upon the Coulomb electrostatic interaction. For a pair of particles denoted as $i$ and $j$,

$$U_{i,j}^{\text{elec}}(r_{ij}) \equiv \frac{k z_i z_j}{r_{ij}}$$

(4.1)

where $k \equiv e^2/4\pi\varepsilon_0 \approx 332.06$ kcal/mol, the valences $z$ are 1 for alkali cations and $-1$ for halide anions, and $r_{ij}$ is the distance between their centers. The dielectric constant is uniformly 1 because the salt crystals are modeled in the absence of other
Table 4.1: Lattice energy (LE) and interionic distance (ID) measurements used as calibration targets. For each salt, the top number is the negative lattice energy and the bottom number is the interionic distance.

<table>
<thead>
<tr>
<th></th>
<th>F</th>
<th>Cl</th>
<th>Br</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>Li</td>
<td>250.7</td>
<td>206.5</td>
<td>196.0</td>
<td>182.6</td>
</tr>
<tr>
<td></td>
<td>1.996</td>
<td>2.539</td>
<td>2.713</td>
<td>2.951</td>
</tr>
<tr>
<td>Na</td>
<td>222.3</td>
<td>188.8</td>
<td>180.2</td>
<td>168.5</td>
</tr>
<tr>
<td></td>
<td>2.295</td>
<td>2.789</td>
<td>2.954</td>
<td>3.194</td>
</tr>
<tr>
<td>K</td>
<td>198.1</td>
<td>172.1</td>
<td>165.2</td>
<td>155.4</td>
</tr>
<tr>
<td></td>
<td>2.648</td>
<td>3.116</td>
<td>3.262</td>
<td>3.489</td>
</tr>
<tr>
<td>Rb</td>
<td>190.0</td>
<td>166.1</td>
<td>159.7</td>
<td>151.1</td>
</tr>
<tr>
<td></td>
<td>2.789</td>
<td>3.259</td>
<td>3.410</td>
<td>3.628</td>
</tr>
<tr>
<td>Cs</td>
<td>181.4</td>
<td>160.1</td>
<td>154.6</td>
<td>146.5</td>
</tr>
<tr>
<td></td>
<td>2.982</td>
<td>3.523</td>
<td>3.668</td>
<td>3.898</td>
</tr>
</tbody>
</table>

In the primitive model (PM), ions are hard spheres which cannot overlap. Each ion has one parameter, its diameter $\sigma$:

$$U_{ij}^{PM}(r_{ij}) \equiv U_{ij}^{elec}(r_{ij}) + \begin{cases} \infty & \text{if } r_{ij} < \frac{1}{2}(\sigma_i + \sigma_j) \\ 0 & \text{otherwise} \end{cases} \quad (4.2)$$

In the Lennard-Jones (LJ) model, ions exhibit short-range attractive van der Waals interactions that compete against a repulsive barrier. Each ion has two parameters, $\sigma$ and $\epsilon$, which respectively describe the length and energy scales of its interactions. We adopt Lorentz-Berthelot mixing rules, which specify arithmetic means for $\sigma$ and geometric means for $\epsilon$, to combine the parameters of two ions into one pairwise interaction.
\[
\sigma_{ij} \equiv \frac{1}{2}(\sigma_i + \sigma_j) \quad \epsilon_{ij} \equiv \sqrt{\epsilon_i \epsilon_j} \quad (4.3)
\]

\[
U_{ij}^{\text{LJ}}(r_{ij}) \equiv U_{ij}^{\text{elec}}(r_{ij}) + 4\epsilon_{ij} \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] \quad (4.4)
\]

### 4.2.3 Lattice energy and interionic distance calculations

Due to the symmetry of a periodic alkali halide crystal lattice, the contribution from one ion to the total potential energy is the same for every cation and for every anion. This contribution is equal to the potential energy of one ion in the field generated by all other ions in the lattice, divided by two to correct for double counting. The sum of the contributions from one cation and one anion, respectively denoted as \( c \) and \( a \), gives the potential energy per salt pair as a function of the tentative interionic distance \( d \):

\[
U_{ca}(d) \equiv \frac{1}{2} \sum_i \sum_{j \neq i} U_{ij}(r_{ij}(d)) \quad (4.5)
\]

Proceeding in a manner analogous to the derivation of the classic Born-Landé equation, we compute the calibration observables by minimizing this intensive potential energy with respect to \( d \). The minimum energy is the lattice energy (\( \text{LE} \)) and optimal value of \( d \) is the interionic distance (\( \text{ID} \)). Since the shape of \( U_{ca}(d) \) depends on the parameters \( \mathcal{P} \), the calibration observables are functions of \( \mathcal{P} \). In Equations 4.6, 4.7, and 4.12, we make this functional relation explicit. \( \mathcal{P}_{\text{ca}} \) denotes parameters pertaining to the cation \( c \) and anion \( a \) and consists of the pair \((\sigma_c, \sigma_a)\) for the primitive model and the quadruple \((\sigma_c, \epsilon_c, \sigma_a, \epsilon_a)\) for the Lennard-Jones model.

\[
\text{ID}_{ca}(\mathcal{P}_{\text{ca}}) \equiv \arg \min_{d>0} U_{ca}(d)
\]

\[
\text{LE}_{ca}(\mathcal{P}_{\text{ca}}) \equiv \min_{d>0} U_{ca}(d) \quad (4.6)
\]
For the primitive model, since the Coulomb interaction draws the lattice together as tightly as possible, the optimal value of $d$ is the one at which lattice ions come into contact with each other:

$$U_{ca}^{PM}(d) = \begin{cases} \infty & \text{if } d < \max \left\{ s\sigma_c, s\sigma_a, \frac{1}{2}(\sigma_c + \sigma_a) \right\} \\
-\frac{b_1k}{d} & \text{otherwise} \end{cases}$$

$$ID_{ca}^{PM}(\sigma_c, \sigma_a) = \max \left\{ s\sigma_c, s\sigma_a, \frac{1}{2}(\sigma_c + \sigma_a) \right\}$$

$$LE_{ca}^{PM}(\sigma_c, \sigma_a) = -\frac{b_1k}{ID_{ca}^{PM}(\sigma_c, \sigma_a)}$$

(4.7)

where $s$ is $1/\sqrt{2}$ for FCC lattices and $\sqrt{3}/2$ for BCC lattices and $b_1$, the Madelung constant, is approximately 1.7476 for FCC lattices and 1.7627 for BCC lattices.

For the Lennard-Jones model, the intensive potential energy can be written as a rational function of $d$ because the lattice sums $b_{12}$ and $b_6$ of the Lennard-Jones potential are numerical constants and can be precomputed:

$$U_{ca}^{LJ}(d) = -\frac{b_1k}{d} + \sum_{i,j}^{\{c,a\}} 2\epsilon_{ij} \left[ b_{12,ij} \left( \frac{\sigma_{ij}}{d} \right)^{12} - b_6,ij \left( \frac{\sigma_{ij}}{d} \right)^6 \right]$$

(4.8)

Unlike the classic Madelung constant, these lattice sums are absolutely convergent [38]. They can be approximated by summing over a finite cube centered at and omitting the origin, but care must be taken to prevent inexact floating point arithmetic from making the summation converge to an inaccurate result [39]. We use exact rational arithmetic for accumulating these sums and convert the final totals to IEEE 754 double-precision floating point numbers. Double-precision floating point arithmetic is used for all other numerical operations in this study. We sum over a cube with $601^3$ ions, keeping separate totals for “even” and “odd” lattice sites, to
obtain the numerical values for FCC lattices in Equation 4.9:

\begin{align*}
b_{6,cc} &= b_{6,aa} \approx 1.8067 \quad b_{6,ca} = b_{6,ac} \approx 6.5952 \\
b_{12,cc} &= b_{12,aa} \approx 0.1896 \quad b_{12,ca} = b_{12,ac} \approx 6.0126
\end{align*}

Likewise, we perform separate sums over cubes with $601^3$ and $600^3$ ions for “like” and “unlike” lattice sites, respectively, to obtain the corresponding BCC lattice sums. Equation 4.10 presents these sums multiplied by $\sqrt{3}/2$, the ratio of interionic distance to lattice constant, raised to the sixth power for $b_6$ or twelfth power for $b_{12}$:

\begin{align*}
b_{6,cc} &= b_{6,aa} \approx 3.5446 \quad b_{6,ca} = b_{6,ac} \approx 8.7091 \\
b_{12,cc} &= b_{12,aa} \approx 1.1038 \quad b_{12,ca} = b_{12,ac} \approx 8.0103
\end{align*}

Decreasing cube side lengths by a factor of three changes the resulting sums by an amount less than $1.2 \times 10^{-6}$.

Minimizing $U_{ca}^{LJ}(d)$ requires the real positive roots of the polynomial $Q$ in Equation 4.11, which was obtained by implementing the mixing rules in Equation 4.3 and factoring the derivative $dU_{ca}^{LJ}/dd$:

\begin{equation}
Q_{ca}(d) = -256b_1kd^{11} - 96 \left[ b_{6,ca} \sqrt{\epsilon_c \epsilon_a} (\sigma_c + \sigma_a)^6 + 32 \left( b_{6,cc} \epsilon_c \sigma_c^6 + b_{6,aa} \epsilon_a \sigma_a^6 \right) \right] d^6 + 3 \left[ b_{12,ca} \sqrt{\epsilon_c \epsilon_a} (\sigma_c + \sigma_a)^{12} + 2048 \left( b_{12,cc} \epsilon_c \sigma_c^{12} + b_{12,aa} \epsilon_a \sigma_a^{12} \right) \right]
\end{equation}

Given particular numerical values for the parameters $\sigma_c$, $\epsilon_c$, $\sigma_a$, and $\epsilon_a$, lattice sums $b$, and electrostatic constant $k$, $Q$ becomes an eleventh-degree polynomial in $d$ whose roots can be obtained via standard methods such as the Jenkins-Traub algorithm [40].
In cases where the derivative has multiple positive roots, the smallest one is taken to define the interionic distance:

\[
ID_{ca}^{LJ}(\sigma, \epsilon, \sigma_a, \epsilon_a) = \min \{ d \in \mathbb{R}_{>0} : Q_{ca}(d) = 0 \}
\]

\[
LE_{ca}^{LJ}(\sigma, \epsilon, \sigma_a, \epsilon_a) = U_{ca}^{LJ}(ID_{ca}^{LJ}(\sigma, \epsilon, \sigma_a, \epsilon_a)) \tag{4.12}
\]

If no positive roots exist, the situation corresponds to an unstable crystal with undefined calibration observables.

### 4.2.4 Calibration objective function and parameter constraints

The root mean square relative deviation from \(O_{\text{measured}}\), the measured values of calibration targets given in Table 4.1, is used as the objective function whose minimization yields the optimized parameters.

\[
F(P) \equiv \sqrt{\frac{1}{40 \ \{\text{salts}\} \{ID,LE\}} \sum_{ca} \sum_{O} \left( \frac{O_{ca}(P_{ca})}{O_{ca}^{\text{measured}}} - 1 \right)^2} \tag{4.13}
\]

In Equation 4.13, the outer sum is over all twenty cation-anion salt pairs, and the inner sum is over the two observables \(ID\) and \(LE\). We used relative instead of absolute deviations because they put all the observables on an equal footing; biases due to differing magnitudes and units are naturally eliminated. As a special case, parameters that cause any salt crystal to be unstable are defined to have an infinite calibration objective function value. The domain \(\{P\}\) of the function consists of the nine-dimensional space \(\{\sigma\}^9\) for the primitive model and the eighteen-dimensional space \(\{\sigma, \epsilon\}^9\) for the Lennard-Jones model.

Following the reasoning of Peng et al. [41], we constrain the domain to main-
tain consistency with periodic table trends. Ions increase in size going down their respective groups of the periodic table, and cations are smaller than their isoelectronic anions. For both the primitive and Lennard-Jones models, Equations 4.14 and 4.15 show the constraints applied to $\sigma$, which correspond to the ion diameters:

$$0 < \sigma_{\text{Li}^+} < \sigma_{\text{Na}^+} < \sigma_{\text{K}^+} < \sigma_{\text{Rb}^+} < \sigma_{\text{Cs}^+}$$

$$0 < \sigma_{\text{F}^-} < \sigma_{\text{Cl}^-} < \sigma_{\text{Br}^-} < \sigma_{\text{I}^-}$$

(4.14)

$$\sigma_{\text{Na}^+} < \sigma_{\text{F}^-}$$

$$\sigma_{\text{K}^+} < \sigma_{\text{Cl}^-}$$

$$\sigma_{\text{Rb}^+} < \sigma_{\text{Br}^-}$$

$$\sigma_{\text{Cs}^+} < \sigma_{\text{I}^-}$$

(4.15)

In accord with the ions’ isoelectronic noble gases, the Lennard-Jones well depths $\epsilon$ also increase going down each group (Equation 4.16):

$$0 < \epsilon_{\text{Li}^+} < \epsilon_{\text{Na}^+} < \epsilon_{\text{K}^+} < \epsilon_{\text{Rb}^+} < \epsilon_{\text{Cs}^+}$$

$$0 < \epsilon_{\text{F}^-} < \epsilon_{\text{Cl}^-} < \epsilon_{\text{Br}^-} < \epsilon_{\text{I}^-}$$

(4.16)

The scale of the London dispersion interaction, which corresponds to the coefficient of the $r^{-6}$ term in the overall potential, is smaller for cations compared to isoelectronic anions because of their lower polarizabilities. It follows that this coefficient, which is equal to $4\epsilon\sigma^6$ in the Lennard-Jones model, must obey the inequalities in
Equation 4.17:

\[
\begin{align*}
\epsilon_{\text{Na}^+}(\sigma_{\text{Na}^+})^6 &< \epsilon_{\text{F}^-}(\sigma_{\text{F}^-})^6 \\
\epsilon_{\text{K}^+}(\sigma_{\text{K}^+})^6 &< \epsilon_{\text{Cl}^-}(\sigma_{\text{Cl}^-})^6 \\
\epsilon_{\text{Rb}^+}(\sigma_{\text{Rb}^+})^6 &< \epsilon_{\text{Br}^-}(\sigma_{\text{Br}^-})^6 \\
\epsilon_{\text{Cs}^+}(\sigma_{\text{Cs}^+})^6 &< \epsilon_{\text{I}^-}(\sigma_{\text{I}^-})^6
\end{align*}
\]

(4.17)

Imposition of these constraints focuses the search on the subset of parameter space that is physically reasonable and promotes transferability of the resulting parameters.

4.2.5 Constrained nonlinear numerical optimization

We implemented the interaction models and calibration objective function with Mathematica 7 (Wolfram Research) and used its constrained nonlinear numerical optimization routines to simultaneously determine all parameters through minimization of the objective function. Initial trials showed that for this global optimization problem, the differential evolution method [42] achieved better performance than simulated annealing [43], the Nelder-Mead simplex method [44], and local minimization from random initial points. Differential evolution is an iterative general-purpose function minimizer that evolves a population of solutions to search for the global minimum. In our application, each solution is a set of parameters for all nine ions. During a single iteration, every member of the population competes against a perturbed version of itself for survival. Perturbations consist of crossing a mutant parameter set with the original such that each parameter randomly inherits its value from either the mutant or the original. Mutant sets are generated by randomly selecting three distinct members of the population and vectorially adding the first to a scaled difference of the other two. If the perturbed solution improves the objective function score, it replaces
the original in the population for the next iteration. The crossover and mutation processes make differential evolution robust in the presence of many local minima and do not require evaluation of objective function gradients.

In employing differential evolution, we used a population size of 100, preserved the default options of 0.5 for the cross probability and 0.6 for the scaling factor, and enabled solution post-processing by the interior point local minimization algorithm [45]. In both differential evolution and local minimization, the convergence criterion was an absolute or proportional change of less than $10^{-8}$ in the appropriate units for all parameters and the objective function value. To facilitate convergence of the search procedure, we supplied randomly generated initial guesses that satisfied most or all of the constraints given in Equations 4.14–4.17. For each set of initial $\sigma$ guesses, nine uniformly and independently distributed random diameters between 1 Å and 5 Å were generated, sorted in ascending order, and assigned to the nine ions according to a random permutation uniformly selected from the fourteen that are consistent with Equations 4.14 and 4.15. For each set of initial Lennard-Jones $\epsilon$ guesses, five log-uniformly and independently distributed random interaction energies between 0.001 and 0.75 kcal/mol were generated, sorted, and assigned to the five cations, with a separate four to the anions, such that the constraints in Equation 4.16 were satisfied but those from Equation 4.17 were sometimes violated. These bounds on the initial guesses did not limit the evolution of the parameters during the minimization process. Initial guesses with constraint violations evolved towards compliance and still contributed during the early iterations via mutant generation.

The entire differential evolution process was executed 100 times with distinct random seeds. Therefore, $10^4$ distinct initial guesses were evolved in independent groups of 100 to produce 100 population champions that we compared to select the
optimal parameter set. In total, each model’s objective function was evaluated on the order of \(5 \times 10^6\) times. Despite the robustness of differential evolution and the large number of parameter sets evaluated, the parameter space is high dimensional and global optimality cannot be guaranteed. For the primitive model, all 100 populations converged toward one of two solutions. We selected the one with the better objective function score as our final recommended primitive model parameter set. In contrast, for the Lennard-Jones model, the 100 populations each produced a distinct champion parameter set and objective function score, with many pushing tightly against the constraints. This suggests that the function landscape is rugged with many local minima, and that deeper minima corresponding to unphysical parameters exist outside the region allowed by the constraints. We selected the champion with the best objective function score that satisfied all constraints to a tolerance of at least \(10^{-5}\) in their respective units as our final recommended Lennard-Jones parameter set.

### 4.3 Results

#### 4.3.1 Calibrated ion parameters

Final optimized parameters for both the primitive model and Lennard-Jones model are presented in Table 4.2. As expected, all parameters satisfy the periodic table trends expressed in Equations 4.14–4.17. If satisfaction of periodic table trends is desired, it is necessary to impose the constraints while simultaneously optimizing all parameters; calibration protocols that do not include these constraints are likely to produce parameters that violate them. In some cases [28, 30], the attained Lennard-Jones \(\epsilon\) actually reverse the expected trend by decreasing down each ion group. In addition to having correct relative magnitudes, the ranges of 1.7 to 5.2 Å for \(\sigma\) and 0.006 to 0.5 kcal/mol for \(\epsilon\) are comparable to those of Lennard-Jones parameters for noble gases [46], even though no absolute bounds other than positivity were enforced
during their optimization.

<table>
<thead>
<tr>
<th>Ion</th>
<th>Primitive model $\sigma$ (Å)</th>
<th>Lennard-Jones model $\sigma$ (Å)</th>
<th>$\epsilon$ (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Li$^+$</td>
<td>1.716</td>
<td>1.715</td>
<td>0.05766</td>
</tr>
<tr>
<td>Na$^+$</td>
<td>2.271</td>
<td>2.497</td>
<td>0.07826</td>
</tr>
<tr>
<td>K$^+$</td>
<td>2.902</td>
<td>3.184</td>
<td>0.1183</td>
</tr>
<tr>
<td>Rb$^+$</td>
<td>3.165</td>
<td>3.302</td>
<td>0.2405</td>
</tr>
<tr>
<td>Cs$^+$</td>
<td>3.559</td>
<td>3.440</td>
<td>0.5013</td>
</tr>
<tr>
<td>F$^-$</td>
<td>2.626</td>
<td>3.954</td>
<td>0.006465</td>
</tr>
<tr>
<td>Cl$^-$</td>
<td>3.600</td>
<td>4.612</td>
<td>0.02502</td>
</tr>
<tr>
<td>Br$^-$</td>
<td>3.903</td>
<td>4.812</td>
<td>0.03596</td>
</tr>
<tr>
<td>I$^-$</td>
<td>4.331</td>
<td>5.197</td>
<td>0.04220</td>
</tr>
</tbody>
</table>

Table 4.2: Short-range interaction parameters derived from crystal lattice properties.

4.3.2 Attained values of calibration observables

The optimized hard sphere diameters for the primitive model attain a relative root mean square deviation (rRMSD) from calibration targets of 4.3%, with RMSDs for interionic distances and lattice energies being 0.12 Å and 8.5 kcal/mol, respectively. While the anions are diminished relative to their crystallographically derived ionic diameters [47], the cations are swollen by a greater amount. As shown in Table 4.3, the attained magnitudes for all lattice energies and interionic distances are greater than their measured values. This indicates that lattice energies naively calculated using ionic radii as the hard sphere radii would be too negative, necessitating an overall swelling that trades off accuracy in interionic distances in exchange for better accuracy in lattice energies. The resulting compromise can be considered a best fit of the primitive model to lattice data, and highlights the deficiency of hard sphere exclusion as a model for short-range repulsion between ions. The need to introduce an overall swelling of ion diameters to match experimental measurements has also been encountered in studies of primitive model activity coefficients in the mean spherical
approximation [20].

<table>
<thead>
<tr>
<th></th>
<th>LE (kcal/mol)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ID (˚A)</td>
<td>F</td>
<td>Cl</td>
<td>Br</td>
</tr>
<tr>
<td>Li</td>
<td></td>
<td>267.3</td>
<td>218.3</td>
<td>206.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.171</td>
<td>2.658</td>
<td>2.810</td>
</tr>
<tr>
<td>Na</td>
<td></td>
<td>237.0</td>
<td>197.7</td>
<td>188.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.449</td>
<td>2.936</td>
<td>3.087</td>
</tr>
<tr>
<td>K</td>
<td></td>
<td>210.0</td>
<td>178.5</td>
<td>170.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.764</td>
<td>3.251</td>
<td>3.402</td>
</tr>
<tr>
<td>Rb</td>
<td></td>
<td>200.4</td>
<td>171.6</td>
<td>164.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.895</td>
<td>3.383</td>
<td>3.534</td>
</tr>
<tr>
<td>Cs</td>
<td></td>
<td>187.7</td>
<td>163.5</td>
<td>156.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.092</td>
<td>3.580</td>
<td>3.731</td>
</tr>
</tbody>
</table>

**Table 4.3:** Negative lattice energies and interionic distances attained using lattice-derived primitive model parameters. Digits colored blue indicate positive deviations from calibration targets. In each number, the most significant colored digit is shaded to indicate the magnitude of the deviation, with darker shades indicating smaller deviations.

The optimized Lennard-Jones parameters attain a rRMSD from calibration targets of 1.4%, with RMSDs for interionic distances and lattice energies being 0.031 Å and 3.0 kcal/mol, respectively. Table 4.4 shows that these deviations are smaller and more balanced than those of the primitive model, reflecting an improved ability of the Lennard-Jones model to capture the essential physics of van der Waals interactions. These deviations are also smaller than those obtained by Peng et al. [41], which is expected since they did not optimize all parameters simultaneously. Compared to Joung and Cheatham’s [27] recommended parameters for use with the TIP3P water model, our deviations are lower for lattice energies and higher for interionic distances. However, the significance of these comparisons is limited due to the distinct calibration targets used in each study and the violation of periodic table trend constraints by the Joung and Cheatham parameters. We rejected several candidate parameter sets.
with better objective function scores because they saturated one or more constraints by having multiple ions with nearly identical $\sigma$, $\epsilon$, and/or $r^{-6}$ coefficients. This suggests that unconstrained minimization of the objective function would lead to even lower objective function scores, underscoring the importance of the constraints in maintaining physical realism.

<table>
<thead>
<tr>
<th>$-LE$ (kcal/mol)</th>
<th>F</th>
<th>Cl</th>
<th>Br</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID (Å)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Li</td>
<td>258.9</td>
<td>210.7</td>
<td>198.4</td>
<td>182.5</td>
</tr>
<tr>
<td>2.073</td>
<td>2.565</td>
<td>2.732</td>
<td>2.976</td>
<td></td>
</tr>
<tr>
<td>Na</td>
<td>226.5</td>
<td>191.7</td>
<td>182.9</td>
<td>170.8</td>
</tr>
<tr>
<td>2.373</td>
<td>2.822</td>
<td>2.968</td>
<td>3.185</td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>200.7</td>
<td>172.9</td>
<td>166.0</td>
<td>156.6</td>
</tr>
<tr>
<td>2.686</td>
<td>3.137</td>
<td>3.278</td>
<td>3.481</td>
<td></td>
</tr>
<tr>
<td>Rb</td>
<td>192.7</td>
<td>166.4</td>
<td>160.0</td>
<td>151.2</td>
</tr>
<tr>
<td>2.814</td>
<td>3.275</td>
<td>3.417</td>
<td>3.622</td>
<td></td>
</tr>
<tr>
<td>Cs</td>
<td>185.4</td>
<td>158.8</td>
<td>152.5</td>
<td>143.8</td>
</tr>
<tr>
<td>2.953</td>
<td>3.521</td>
<td>3.675</td>
<td>3.903</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.4: Negative lattice energies and interionic distances attained using lattice-derived Lennard-Jones parameters with Lorentz-Berthelot mixing rules. Digits colored blue/red indicate where attained values exceed/fall short of calibration targets. In each number, the most significant colored digit is shaded to indicate the magnitude of the deviation, with darker shades indicating smaller deviations.

### 4.3.3 Lattice structure prediction

Using the optimized Lennard-Jones parameters, we compute lattice energies for all twenty salts in both FCC and BCC lattice arrangements to check whether the experimentally determined crystal structure is correctly favored. We find that FCC lattice energies are more negative for all twenty salts; this is incorrect for the three BCC salts and correct for the other salts. However, the gaps between FCC and BCC energies are only 1.6, 1.7, and 2.2 kcal/mol for CsCl, CsBr, and CsI, respectively.
These gaps are narrower than those of all seventeen FCC salts, where the correct lattice was favored by between 2.9 and 22. kcal/mol, and smaller than the deviations of attained lattice energies from their experimental values. Lingering deviations from calibration targets for the optimized parameters suggests that a limit on the accuracy of the Lennard-Jones model has been reached; a more realistic model that possibly includes additional parameters is necessary to achieve better agreement with experiments. This is especially relevant for ions with large electron clouds and high polarizability, such as Cs\(^+\) and the larger anions, and may be necessary to correct the lattice structure predictions.

4.3.4 Hydration free energies

To test the transferability of parameters derived using lattice properties, we estimate hydration free energies \(\Delta G_{\text{hyd}}\), an observable that is not used anywhere in their derivation. Joung and Cheatham [27] constructed bicubic surfaces \(\Delta G_{\text{hyd}}^{\text{calc}}(\sigma, \epsilon)\) that provide single-ion hydration free energies as a function of the ion’s \(\sigma\) and \(\epsilon\) by fitting the results of several hundred thermodynamic integration calculations (note that they used \(R_{\min} = 2^{(1/6)}\sigma\) instead of \(\sigma\)). They then combined these surfaces with experimental hydration free energies \(\Delta G_{\text{hyd}}^{\text{expt}}\) to obtain mappings between \(\sigma\) and \(\epsilon\) such that \(\Delta G_{\text{hyd}}^{\text{calc}}(\sigma, \epsilon) = \Delta G_{\text{hyd}}^{\text{expt}}\). However, the bicubic surfaces themselves, which are not influenced by the experimental values, independently constitute a useful and valuable tool because they enable one to estimate the results of hydration free energy calculations for arbitrary Lennard-Jones spheres with monovalent charge to within \(\sim 0.3\) kcal/mol without performing any simulations. We employ these surfaces to estimate the calculated hydration free energy for each alkali and halide ion using our Lennard-Jones parameters in TIP3P [48], TIP4P-Ew [49], and SPC/E [50] water at 298.15 K using a standard state of 1 mol/liter for both gas and solution phases. As
with the original calculations, these estimates incorporate Lorentz-Berthelot mixing rules for ion-water interactions and omit corrections for the liquid-vacuum surface potential [51–53]. Table 4.5 compares the estimated hydration free energies to experimental measurements compiled by Schmid et al. [54], which also do not account for a liquid-vacuum surface potential.

<table>
<thead>
<tr>
<th>Ion</th>
<th>TIP3P</th>
<th>TIP4P-Ew</th>
<th>SPC/E</th>
<th>Measured $-\Delta G_{\text{hyd}}$ (kcal/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Li$^+$</td>
<td>113.4</td>
<td>106.1</td>
<td>112.1</td>
<td>113.8</td>
</tr>
<tr>
<td>Na$^+$</td>
<td>87.6</td>
<td>82.5</td>
<td>85.6</td>
<td>88.7</td>
</tr>
<tr>
<td>K$^+$</td>
<td>69.7</td>
<td>65.8</td>
<td>67.5</td>
<td>71.2</td>
</tr>
<tr>
<td>Rb$^+$</td>
<td>65.4</td>
<td>62.1</td>
<td>63.4</td>
<td>66.0</td>
</tr>
<tr>
<td>Cs$^+$</td>
<td>61.5</td>
<td>58.9</td>
<td>59.9</td>
<td>60.5</td>
</tr>
<tr>
<td>F$^-$</td>
<td>118.7</td>
<td>122.4</td>
<td>123.8</td>
<td>119.7</td>
</tr>
<tr>
<td>Cl$^-$</td>
<td>87.9</td>
<td>90.0</td>
<td>90.2</td>
<td>89.1</td>
</tr>
<tr>
<td>Br$^-$</td>
<td>81.5</td>
<td>83.2</td>
<td>83.2</td>
<td>82.7</td>
</tr>
<tr>
<td>I$^-$</td>
<td>73.4</td>
<td>74.5</td>
<td>74.4</td>
<td>74.3</td>
</tr>
</tbody>
</table>

Table 4.5: Negative hydration free energies attained using lattice-derived Lennard-Jones parameters and three water models. Experimental measurements from Schmid et al. [54] are included for comparison. A temperature of 298.15 K and standard state of 1 mol/liter in both the gas and solution phase applies to all values. Digits colored blue/red indicate where calculated values exceed/fall short of measured values. In each number, the most significant colored digit is shaded to indicate the magnitude of the deviation, with darker shades indicating smaller deviations.

The attained RMSDs are 1.0, 4.1, and 2.4 kcal/mol for TIP3P, TIP4P-Ew, and SPC/E, respectively. Gradients of the bicubic surfaces evaluated at optimized parameter values indicate that hydration free energy is especially sensitive to Lennard-Jones parameters for small ions: across the three water models, the partial derivatives of $-\Delta G_{\text{hyd}}^\text{calc}$ with respect to $\sigma$ and $\epsilon$ are at least [31.9, 77.2] kcal/mol for Li$^+$ and
Therefore, a deviation of the these parameters on the order of 0.1 Å for \( \sigma \) or 0.01 kcal/mol for \( \epsilon \) away from their lattice-calibrated values could significantly degrade the agreement with experimental hydration free energies. The steepness of these bicubic surfaces is intrinsic to the water models and implies that hydration free energies are a stringent test of the accuracy of Lennard-Jones parameters. Greater discrepancy between simulation and experiment for TIP4P-Ew compared to the three-site models can be explained by displacement of its negative charge site relative to the oxygen Lennard-Jones center. This increases the distance from the negative charge of an ion approaching from the oxygen side and accounts for the observed trend of increasing severity of hydration favorability underestimation with decreasing size for the cations [55]. Although caveats regarding assumptions about proton solvation thermodynamics implicit to the experimental values apply [56], the general agreement across water models between simulation and experiment on an aqueous system that is completely unlike the crystalline phase used to derive these parameters is evidence of their transferability and validity.

4.4 Discussion

Previous studies have demonstrated the utility of lattice properties as calibration targets. Peng et al. [41] modified noble gas parameters by adjusting cation sizes and dispersion well depths separately to fit lattice properties and periodic table trends within the framework of a 9-6 van der Waals model with Waldman-Hagler mixing rules. Gee et al. [30] used lattice dimensions to determine dispersion well depths as part of an effort to reproduce experimentally derived Kirkwood-Buff integrals using simulations with SPC/E water. Joung and Cheatham’s thorough study [27] fit lattice properties and ion-water binding measurements subject to constraints that maintained accurate hydration free energies, ultimately recommending three com-
pletely different sets of parameters for three water models. They also speculated that it would be possible to use lattice properties as the sole calibration observables in order to achieve water-independent ion parametrization. The present study confirms this speculation.

The proliferation of alkali and halide ion parameters for nonpolarizable, pairwise additive force fields illustrates the inherent disadvantages of solution properties as calibration targets. Since solvent model parameters are fixed during these calibrations, any shortcomings and idiosyncrasies of the solvent model become embedded in the resulting parameters. Therefore, every combination of solvent model and calibration targets engenders its own ion parameters. In fact, since the calculated solution properties also depend on simulation design choices such as system size, cutoff distance, boundary condition, free energy calculation method, electrostatic approximation method, etc., careful sensitivity analyses must be performed during calibration to prevent the resulting parameters from becoming specific to those choices as well. For small systems, *ab initio* molecular dynamics and density functional theory [57, 58] alleviate this problem by explicitly modeling electronic degrees of freedom, providing a transferable way to simulate ions without deriving or choosing a force field. These calculations may serve as calibration targets for observables that lack definitive experimental measurements. However, the results are still sensitive to methodological details such as the choice of basis set, pseudopotential, exchange-correlation functional, etc., and simulations of larger systems such as solutions of biopolymers at this level of detail are currently intractable.

The ion parameters presented in this study provide an alternative to the paradigm of solvent-specific customization for use with nonpolarizable, pairwise additive force fields. Since no choices regarding a solvent model or simulation design were
made during their derivation, they are potentially suitable for a broader variety of systems compared to ion parameters built on top of a solvent model. Lattice-calibrated parameters are immediately applicable to salt crystal and molten salt simulations. Their independence makes them a reasonable first guess in simulations with solvents besides water that lack customized ion parameters. As with water, solvent-ion pair potentials determined by the mixing rule may be individually overridden [29, 59] to attain enhanced accuracy for solvent-specific properties such as solvation structure [60, 61] and gas phase cluster properties [62]. Retaining the mixing rule defaults for other pairs preserves the intrinsic character of the ions in their interactions with each other and with other matter. Encouragingly, the agreement with measured hydration free energies shows that true transferability, where no such tuning is necessary, is not impossible. In fact, these parameters even make it possible to reverse the direction of dependence by calibrating solvent models while keeping the ion parameters fixed. In some situations, aqueous simulations may benefit from adopting these parameters despite the abundance of ion parameters tuned for particular water models. Lattice-calibrated ions are well suited for the ABSINTH implicit solvation model [23], where ion-water geometries are unavailable due to the continuum solvent description and experimental hydration free energies are direct inputs to the model. They are also justified whenever additional solutes are present and the balance between ion-water and ion-solute interactions is a subject of inquiry; the absence of solvent in their derivation makes them inherently unbiased compared to ion parameters co-derived with a water model.

The general approach of using crystal lattice data to derive ion parameters should remain effective for more demanding problems such as multivalent ions and polarizable force fields. A wealth of experimental data and techniques from solid
state physics are already available, and the combinatorics that allow formation of many different salts from few ions readily lead to overdetermined systems that are desirable in calibrations. In addition, the periodic nature of crystal lattices makes their calibration observables inherently easier to calculate compared to ones involving dense fluids. As demonstrated in this study, the use of analytical differentiation and polynomial root solvers make calculation of calibration observables possible with minimal computational effort compared to full equilibrium ensemble simulations, enabling the calculation to be wrapped in a function that is called numerous times by a global minimization routine. The absence of solvent and symmetry of crystal lattices make this calibration process as simple as possible.

4.5 Conclusion

Crystal lattice properties are sufficient to determine a simultaneous fit of all alkali and halide ion parameters for both the primitive model and Lennard-Jones model with Lorentz-Berthelot mixing rules. The resulting parameters presented here are transferable, consistent with periodic table trends, and free from the influence of any solvent model. The success and generality of the method used to derive them suggests that it may be used as a template in future parametrization studies.

4.6 References


Chapter 5
Sequence patterning modulates composition in determining conformational ensembles of polyampholytic intrinsically disordered proteins

This chapter briefly presents results from simulations of IDPs obtained using the refined ion parameters developed in Chapter 4 and other enhancements intended to address the issues raised in Section 3.3. Three sets of IDP sequences are studied to investigate the impact of these methodological improvements on the phenomenon of asymmetric counterion-mediated collapse of acid- but not base-rich IDPs described in Chapter 3. The first set consists of the four synthetic polyelectrolyte IDPs that arise from constructing homopolymers of arginine, lysine, aspartate, and glutamate residues. The second set is a series of synthetic IDPs with glutamate and glutamine residues, with the fraction of glutamate varying from zero to one to systematically titrate the net charge per residue. Finally, the third set contains eight naturally occurring IDPs and maintains our focus on biologically relevant sequences.

Although this chapter serves as an epilogue for the dissertation, our broader effort to elucidate relationships between primary structure, solution conditions, and IDP conformational ensembles is only beginning and continues in earnest. The candidate, Albert H. Mao, designed the sequence sets and performed the simulations.
Anuradha Mittal contributed the initial analysis and interpretation of the results, which guided the candidate’s production of the figures. Rahul K. Das is investigating the effects of sequence patterning using synthetic glutamate/lysine polypeptides and has presented results at the 2012 Intrinsically Disordered Proteins Gordon Research Conference in Mt. Snow, Vermont. As in Chapter 2, Scott L. Crick is performing fluorescence experiments that will test the quantitative predictions of our simulations. Therefore, this chapter should be viewed as the initial step of a larger collaborative study which is still in progress.

5.1 Simulation methods

An enhanced version of the protocol described in Section 2.3.2 is used to simulate all IDP sequences in this study. To enable direct comparisons with fluorescence experiments, each IDP sequence is flanked by the residues WPP at the N-terminus and C at the C-terminus. Each construct is capped with acetyl and N-methylamide groups at the N and C-termini, respectively, and placed inside a 110 Å radius spherical droplet along with explicitly represented salt ions sufficient to neutralize the net polypeptide charge and mimic a 125 mM salt solution. Ten independent replicas of each sequence in both NaCl and KCl salt are simulated using randomly generated starting conformations taken from the excluded volume ensemble [1]. To provide reference states, one full-length simulation of each sequence without any ions is performed using both the excluded volume Hamiltonian and a Hamiltonian with only the Lennard-Jones terms enabled. Histidine sidechains are singly protonated at their $\epsilon$ nitrogen, which makes the sidechain uncharged. The ABSINTH implicit solvation, electrostatic, Lennard-Jones, and Engh-Huber crystallographic geometry parameters described in Vitalis and Pappu [2], including artificially lowered reference sidechain hydration free energies of $-107.3$ kcal/mol for aspartate and glutamate and $-100.9$ kcal/mol for lysine
and arginine, are used with several modifications. For salt ions, the Lennard-Jones parameters are calibrated using crystal lattice properties [3] as described in Chapter 4, ionic radii used in solvation shell overlap calculations are updated to values of 1.16 Å for Na\(^+\), 1.52 Å for K\(^+\), and 1.67 Å for Cl\(^-\) suggested by Shannon [4], and hydration free energies are updated to values compiled by Schmid et al. [5] and listed in Table 4.5. In addition, refined torsional potentials and Lennard-Jones parameters developed by Radhakrishnan et al. [6] are enabled for proline. Cutoffs for the Lennard-Jones and electrostatic interactions between neutral groups are set at 10 and 14 Å, respectively. Electrostatic interactions between ions in solution and sidechain moieties with an overall net charge are computed without employing cutoffs.

The CAMPARI software package developed by Andreas Vitalis with contributions from other members of the Pappu lab has been released and is available at http://campari.sourceforge.net. It includes an updated version of the simulation engine used in previous studies employing the ABSINTH implicit solvation model. Among other enhancements, it implements the exact concerted rotation moves developed Dinner [7]. We replace the inexact version of Favrin et al. [8] with this exact version using a pre-rotation bias strength of 8.0, distribution width of 1.0 degrees, minimum segment length of one residue, and maximum length of three residues. The chain closure techniques used in concerted rotation are also responsible for achieving sampling of proline ring degrees of freedom. For moves targeted at the proline sidechain, we use a combination that mixes 25% reflection moves that invert the puckering state of the ring with 75% local random perturbations that alter bond angles by up to 1 degree or dihedral angles by up to 2 degrees. Metropolis Markov chain Monte Carlo (MC) simulations are performed in the canonical ensemble at 298 K for a total of \(6 \times 10^7\) steps, with the first \(10^7\) steps treated as equilibration and omitted from
analysis. Besides these modifications, the move set is similar to the one described in Section 2.3.2 and includes cluster rigid body moves. As a control condition, we simulate protamines 2 and 4 from Figure 2.1 find that they exhibit qualitatively similar behavior between the original and revised simulation protocols.

5.2 Results

5.2.1 Both acid-rich and base-rich polyelectrolytes adopt swollen coil ensembles

By comparison with the results presented in Sections 3.2.1 and 3.2.3, simulations of polyelectrolyte IDPs using the updated simulation protocol reveal the potential impact of solution conditions on conformational ensembles of IDPs. Figure 5.1 shows the normalized radius of gyration for (Asp)$_{50}$, (Glu)$_{50}$, (Lys)$_{50}$, and (Arg)$_{50}$. Similar to their behavior with ABSINTH restricted primitive model ions and unlike their behavior with original ABSINTH ions, all four of these synthetic polyelectrolyte IDPs adopt swollen random coil (referring to self-avoiding random walks, not Flory random coils) conformational ensembles. This is consistent with biophysical data gathered for prothymosin α, a protein rich in aspartate and glutamate, using a variety of methods [9, 10]. However, the asymmetry between positively and negatively charged sidechains persists in a more subtle form: the degree of electrostatic expansion for acid-rich polyelectrolytes is reduced compared to that of base-rich polyelectrolytes. While the phenomenon of counterion-mediated collapse and its concomitant obliteration of Monte Carlo sampling quality appears to have disappeared upon introduction of the updated parameters, the increased size of the error bars for polyaspartate and polyglutamate relative to polylysine and polyarginine indicate that sampling quality is still degraded for acid-rich IDPs. In addition, the increased uncertainty is more prominent for polyaspartate than for polyglutamate. Although this uncertainty pre-
vents a definitive interpretation of the different values, a reduction in expansion for aspartate compared to glutamate could be rationalized by its higher spatial charge density that results from its shorter sidechain length. Similar reasoning would also explain an increased compaction for polyelectrolytes in NaCl compared to KCl.

Figure 5.1: Normalized \( \langle R_g \rangle / \langle \text{Random coil} \ R_g \rangle \) for \((\text{Asp})_{50}, (\text{Glu})_{50}, (\text{Lys})_{50}, \) and \((\text{Arg})_{50} \). Error bars denote the standard deviation of the normalized \( \langle R_g \rangle \) obtained from the ten independent replicate simulations. Since each individual replicate itself constitutes a large sample from a population, we interpret this standard deviation as the standard error of the mean (SEM).
5.2.2 Synthetic glutamine/glutamate IDPs exhibit a globule-
to-coil transition with net charge per residue

Next, we explore the dependence of conformational ensemble properties for
acid-rich IDPs analogously to the investigation of Chapter 2 by simulating synthetic
glutamine/glutamate IDPs. The fraction of glutamate residues is varied from zero to
one while keeping the total number at 50 residues to achieve a systematic titration
through sequence space reminiscent of the base-rich protamine set. This approach
circumvents the dearth of naturally occurring IDPs with \( f_+ - f_- < -0.4 \). The pat-
terning of glutamate residues amongst glutamine residues is intentionally made as
even as possible using an algorithm that distributes the less common amino acid at
uniform intervals between the more common one and then distributing the remainder
in a way that minimizes clumping. Figure 5.2 provides an inventory of these se-
quences, which include polyglutamate and polyglutamine at its limits. Since \( f_+ = 0 \)
for all of these sequences, the net charge per residue \( f_+ - f_- \) is equal to \( -f_- \).

<table>
<thead>
<tr>
<th>Sequence</th>
<th>( f_- )</th>
</tr>
</thead>
</table>
| WPPQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ

**Figure 5.2:** Inventory of synthetic Q/E IDP sequences. Glutamate residues are colored
red. The values given for \( f_- \) reflect the additional WPP and C residues that make the total
sequence length equal to 54 residues.
As with the protamines, the net charge per residue positions these synthetic Q/E IDPs along a globule-to-coil transition. Figure 5.3 plots the normalized \( \langle R_g \rangle \) as a function of \( f_+ - f_- \). For values of the net charge per residue between 0 and approximately \(-0.6\), the plot resembles a horizontally reflected version of Figure 2.21, with normalized \( \langle R_g \rangle \) values increasing sharply as \( f_+ - f_- \) is decreased from about \(-0.2\) to \(-0.4\). The extent of expansion continues to increase past the swollen random coil limit, ultimately attaining a normalized \( \langle R_g \rangle \) of approximately 1.35. No significant or systematic difference between NaCl and KCl is evident in these results. Therefore, the asymmetry in polymeric classification encountered in Chapter 3 appears to be artifactual and qualitatively mitigated by the enhanced simulation protocol, making the simulated behavior of these synthetic acid-rich IDPs consistent with predictions of the phase diagram from Chapter 2.

Uversky et al. found that prothymosin \( \alpha \) undergoes a collapse transition as the pH is lowered using a combination of nuclear magnetic resonance spectroscopy, small angle x-ray scattering, circular dichroism, and ANS fluorescence [11]. Lowering the pH would have the effect of driving the equilibrium of the carboxylate groups in glutamate and aspartate sidechains towards their neutral, protonated tautomers, thereby decreasing the magnitude of their net charge per residue. If one views glutamine sidechains as a proxy for protonated glutamate sidechains, the results shown in Figure 5.3 are entirely consistent with these findings. This suggests that the substitution of glutamate for glutamine in both simulations and experiments may be an effective way of imitating or modulating the influence of pH on conformational ensembles.
Figure 5.3: Normalized $\langle R_g \rangle$ as a function of net charge per residue for synthetic Q/E IDPs. Error bars denote the standard error of the mean.

5.2.3 Conformational ensembles of proline-rich and strongly polyampholytic IDRs disagree with phase diagram predictions

To complement the physical insight provided by synthetic polypeptides, we investigate a set of eight naturally occurring acid-rich IDPs and IDRs to reassess the predictions of our phase diagram. The set includes three fully disordered proteins: thymosin $\beta 4$ [12] (MSDKPDMAIEKFDKSKLKTETQEKNLPSKETIEQEKQGES), Osteocalcin [13] (YLDSGLGAPVPYPDPLEPKREVSELNPSDELAHDIGFQEAYQRFYGPV), and $\alpha$-aminoadipate carrier protein lysW [14] (MVGTSPESGAELRLENPELGELVSEDGAELEVV)
GLDPLRLEPAPEEAEDWGE). The remaining five sequences are disordered regions from polyglutamine-binding protein 1 [15] (DRGHDKSDRERGDKVDERERERDREDRDGRYDKAD REEQRKRRHHRREE), canavalin [16] (MAFSARFPLWLLLLGVLASVSASFAHSAGSGAEDESES RAQ), protein farnesyltransferase/geranylgeranyltransferase type-1 subunit \( \alpha \) [17] (MAA TEGVGEAAQGEPQAPQQPPPQHPPPQHHQKKEEMAEAGAVEASPMDG), B-cell antigen receptor complex-associated protein \( \beta \) chain [18] (LDKDSKAGMEEHTYEGLDQATATYEDIVL RTGEVKWSVGEHPQGE), and tubulin \( \beta 4 \) chain [19] (FLHWYTGEEMDEFTAEESNMNDLVSEYQ QYQDATAEQGEFFEEAAEVEVA). As in Chapter 3, we excise them from their sequence contexts and model them as isolated entities. All sequences are chosen to have lengths in the vicinity of 50 residues and have been observed to exist in vivo. To facilitate comparisons with experiments that require fluorescent labeling, all internal cysteines in the sequence are replaced with serines to prevent dyes from attaching at the wrong location. As in previous sections, the resulting sequence is then flanked by WPP and C. The final constructs are shown in Figure 5.4. Since they have \( \left| f_+ - f_- \right| < 0.4 \), all of them lie within or near the boundary of the phase diagram region predicted to exhibit disordered globular ensembles.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Sequence</th>
<th>Length</th>
<th>UniProtKB</th>
<th>( f_+ )</th>
<th>( f_- )</th>
<th>( f_+ - f_- )</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>WPP</td>
<td>NGRKGGSDRGERDGTHVDEREREDRGERDREDRDGRYDKAD REEGKRRHHRREE</td>
<td>56</td>
<td>P64350</td>
<td>0.36</td>
<td>0.36</td>
<td>-0.00</td>
<td>0.17</td>
</tr>
<tr>
<td>WPPM</td>
<td>SASHAKGEEFSLKETTEKHKPLRRITTEKQKEKAGECSC</td>
<td>48</td>
<td>P58282</td>
<td>0.35</td>
<td>0.19</td>
<td>-0.16</td>
<td>-0.33</td>
</tr>
<tr>
<td>WPPMA</td>
<td>SASHAKGEEFSLKETTEKHKPLRRITTEKQKEKAGECSC</td>
<td>49</td>
<td>P40259</td>
<td>0.40</td>
<td>0.20</td>
<td>-0.20</td>
<td>-0.51</td>
</tr>
<tr>
<td>WPPMS</td>
<td>SASHAKGEEFSLKETTEKHKPLRRITTEKQKEKAGECSC</td>
<td>53</td>
<td>P49354</td>
<td>0.36</td>
<td>0.17</td>
<td>-0.19</td>
<td>-0.42</td>
</tr>
<tr>
<td>WPPMAA</td>
<td>SASHAKGEEFSLKETTEKHKPLRRITTEKQKEKAGECSC</td>
<td>58</td>
<td>P81455</td>
<td>0.02</td>
<td>0.16</td>
<td>-0.14</td>
<td>0.39</td>
</tr>
<tr>
<td>WPPMD</td>
<td>SASHAKGEEFSLKETTEKHKPLRRITTEKQKEKAGECSC</td>
<td>58</td>
<td>P50477</td>
<td>0.08</td>
<td>0.26</td>
<td>-0.19</td>
<td>0.39</td>
</tr>
<tr>
<td>WPPMV</td>
<td>SASHAKGEEFSLKETTEKHKPLRRITTEKQKEKAGECSC</td>
<td>58</td>
<td>P62328</td>
<td>0.03</td>
<td>0.28</td>
<td>-0.25</td>
<td>0.45</td>
</tr>
<tr>
<td>WPPMT</td>
<td>NGRKGGSDRGERDGTHVDEREREDRGERDREDRDGRYDKAD REEGKRRHHRREE</td>
<td>55</td>
<td>O98289</td>
<td>0.00</td>
<td>0.31</td>
<td>-0.31</td>
<td>0.39</td>
</tr>
</tbody>
</table>

**Figure 5.4:** Inventory of naturally occurring acid-rich IDPs and IDRs. Acidic residues are colored red and basic residues are colored blue. UniProtKB refers to the UniProt protein knowledge base [20]. The column labeled H gives the normalized Kyte-Doolittle hydropathy score [21] averaged over each sequence.

As shown in Figure 5.5, three of the eight sequences exhibit normalized \( \langle R_g \rangle \)
values greater than 0.75, a degree of expansion that is inconsistent with the disordered globular ensembles predicted by their positions on the phase diagram. This discrepancy is opposite to the one encountered in Chapter 3, where sequences predicted to be swollen coils actually collapsed in simulations. The other five sequences behave in accordance with phase diagram predictions, with the normalized $\langle R_g \rangle$ mildly increasing as $f_+ - f_-$ is decreased towards the transition region. Once again, no significant difference is observed between NaCl and KCl at this concentration.

Figure 5.5: Normalized $\langle R_g \rangle$ as a function of net charge per residue for naturally occurring acid-rich IDPs and IDRs. Colors are consistent with the symbols of Figure 5.4. Filled and hollow circles denote values obtained with 125 mM NaCl and KCl, respectively. Error bars denote the standard error of the mean.
5.3 Discussion

Examination of the sequences for these three exceptions suggests several hypotheses that account for their deviation from predictions. The IDR from protein farnesyltransferase/geranylgeranyltransferase type-1 subunit α is rich in proline residues, while the IDRs from thymosin β4 and polyglutamine-binding protein 1 have a high overall charge content and are therefore strong polyampholytes despite having a low net charge per residue. In the former case, the intrinsic stiffness of the proline-rich segments may prevent the chain from collapsing tightly enough to realize ⟨Rg⟩ values that are consistent with the globular reference state. In contrast, the two strong polyampholytes with significant quantities of both acidic and basic residues may expand because this allows all of their charged sidechains, which have highly favorable hydration free energies, to maintain solvent exposure. As shown in Figure 5.6, favorable electrostatic interactions of sidechains with ions can substitute for close proximity between oppositely charged sidechains that would tend to collapse the chain. This mechanism would only come into play as the total sequence length exceeds a certain threshold, as shorter polyampholytes such as the ones illustrated in Figure 2.26 can form a hydrophobic “core” from backbone atoms while keeping sidechain groups exposed to solvent. It also implies rich dependencies on both sequence patterning and salt concentration, with simple polymeric classifications based on ⟨Rg⟩ possibly becoming inadequate at describing conformational ensembles. Studies that explore these dependencies are the next logical step in refining the phase diagram to create a general predictor for conformational ensemble properties of IDPs that handles these strong polyampholytes.
Figure 5.6: Example expanded conformation of WPPDRGHDKSDRDREGEYDKVDREERDREERDRD
RGYDKADREEKERRHRREEC with associated KCl salt ions. The rendering is produced using
Visual Molecular Dynamics (VMD) [22] and includes all ions within 10 Å of any protein
atom. K⁺ ions are colored green and Cl⁻ ions are colored cyan, and both are drawn as
3 Å diameter spheres. In rendering the protein, carbon, oxygen, and nitrogen atoms are
respectively colored cyan, red, and blue, while hydrogen is hidden. In addition, a smoothly
curved spline highlights the course of the protein backbone and changes from red at the
N-terminus to blue at the C-terminus.
5.4 References


Chapter 6

Summary of contributions and future directions

This dissertation joins a broad, interdisciplinary, and ongoing scientific effort to understand intrinsically disordered proteins. Its main contribution is the elucidation of general principles that enable prediction of an IDP’s conformational ensemble using its primary structure. Chapter 2 presents results from molecular simulations and fluorescence experiments on a set of base-rich IDPs showing that net charge per residue segregates conformational ensembles along a globule-to-coil transition. It synthesizes these findings with other results to produce a phase diagram that predicts polymeric classifications of IDP conformational ensembles based on simple compositional sequence characteristics. Chapter 3 shows that some IDPs predicted to have swollen coil conformational ensembles exhibit counterion-mediated collapse under certain solution conditions, a finding that motivated the systematic refinement of alkali and halide ion parameters described in Chapter 4. The predictions of the phase diagram are found to hold for synthetic polyelectrolyte and polyglutamine/polyglutamate IDPs in Chapter 5, but several naturally occurring polyampholytic IDPs and IDRs are revealed to be counterexamples that exhibit swollen coil ensembles despite lying within the disordered globule region of the phase diagram. These exceptions suggest that for strongly polyampholytic IDPs of sufficient length, sequence patterning modulates the electrostatic attraction that would tend to collapse the chain. Going forward, detailed investigations into the role of sequence patterning and folded context will provide a
more complete understanding of the relationship between primary structure and conformational ensembles for IDPs. These investigations are already underway.

Improvements to the physical models employed by molecular simulations constitute a separate category of contributions described in this dissertation. Although they were motivated by specific challenges encountered during computational and biophysical investigation of IDPs, these advances are of general utility in the broad and expanding field of molecular simulation. In Chapter 4, we design and implement a strategy for calibrating interaction parameters of alkali and halide ions that relies exclusively on crystal lattice properties. Unlike the ion parameters used in chemical and biomolecular force fields, the resulting parameters have validity independent of any solvent model, and are potentially applicable to simulations of proteins, nucleic acids, lipids, carbohydrates, liquid-vapor interfaces, and electrolyte solutions in a variety of solvents. Before the parameters are turned loose on those systems, detailed tests of their ability to reproduce gas phase cluster energies and geometries, solution phase pair correlation functions, diffusion constants, and other experimental measurements are necessary. The calibration procedure itself is general and can be adapted to more challenging parametrization tasks such as multivalent ions and polarizable force fields. Attacking these problems is a promising direction for improving force fields to the next level of maturity and reliability.

Although they are not emphasized in the preceding chapters, methodological innovations are another aspect of this dissertation. In Appendix A, we describe and demonstrate a recording method that enables exact recording of Metropolis-Hastings-class Monte Carlo simulations using one bit per sample. The operational advantages of this method facilitate decoupling of simulation design from data analysis, enabling simulation data for systems of arbitrary size to be recorded with minimal demands
on storage or bandwidth resources. These benefits become more significant with increasing scientific ambition as simulations are applied to larger and more numerous systems. One intriguing possibility created by this method is an approach to multi-scale modeling that simply retains all degrees of freedom at their highest level of detail. Potential energy functions would still be computed over a coarse-grain representation for efficiency, but the fine-grain details would be preserved and modified at every step of the simulation without pressure to compromise for the sake of reducing the volume of generated data.

Additional advances in simulation techniques and analysis will be necessary in the endeavor to understand intrinsically disordered proteins and the relation between conformational ensembles, primary structure, and solution conditions. Several efforts to develop these improvements are already underway, and this dissertation contains descriptions of initial progress along several fronts that are likely sources of new results in the near future. First, the apparent relevance of sequence patterning in modulating IDP conformational ensembles discussed in Chapter 5 underscores the need for reliable quantifiers of patterning. Although the combinatorially immense number of possible sequences containing even moderate numbers of residues admits endless possibilities for patterning, we have devised a general and simple measure called the Mixing Extent that quantifies the degree of segregation of amino acids within a sequence. Appendix B describes this measure, an efficient algorithm that enables its calculation, and its distribution across both folded and disordered proteomes. Second, the importance of solution conditions demonstrated in Chapter 3 motivates the development of simulation techniques that accurately mimic the effects of titrating the bulk concentration of a solute in an experiment. Appendix C presents an overview of grand canonical Monte Carlo simulations that provide this
capability, and focuses on various practical aspects of computing activity coefficients in ABSINTH solvent. Finally, as suggested in Section 2.3.2, grand canonical Monte Carlo will become the core of a constant pH simulation engine that enables modeling of protonation equilibria. The combination of ABSINTH with the approach suggested by Mongan et al. [1] will yield an engine that is well suited for studying the pH dependence of IDP conformational ensembles.

Although it is difficult to predict the distant future, we speculate that the combination of computational simulations with theory and experimentation will become a dominant mode of inquiry across all areas of natural science. Implicit in this statement is our belief that simulation is distinct from both theory and experiment. This is a controversial position because the relationship between those three activities is a subject of active debate in epistemology and the philosophy of science [2–5]. In one pattern, simulation may be regarded as an extension of theory, as numerical simulations are used to obtain predictions from an analytically intractable theory. In a second pattern, simulations operate as a detail enhancement device for experiments, as simulation model parameters are fit to reproduce experimental measurements and then interpreted to gain information about quantities that are inaccessible to direct measurement. In a final and desirable pattern, models are accurate and trusted to the extent that simulations are regarded as in silico experiments. Scientific inquiry using simulations regularly follows each of these three patterns at various times. Practitioners of simulations must recognize which pattern is actually intended, and should strive to improve simulation methods until the third pattern is attainable. They must also understand both physical theories and experimental techniques in order to make quantitative and testable predictions. Acquiring these abilities may be challenging, but promises to be tremendously rewarding because they enable the synergy between
simulation, theory, and experiment that ultimately drives progress in computational and molecular biophysics.

6.1 References


Appendix A

Exact recording of Metropolis-Hastings-class Monte Carlo simulations using one bit per sample

The Metropolis-Hastings (MH) algorithm is the prototype for a class of Markov chain Monte Carlo methods that propose transitions between states and then accept or reject the proposal. These methods generate a correlated sequence of random samples that convey information about the desired probability distribution. Deciding how this information gets recorded is an important step in the practical design of MH-class algorithm implementations. Many implementations discard most of this information in order to reduce demands on storage capacity and disk writing throughput. Here, we describe how recording a bit string containing 1’s for acceptance and 0’s for rejection allows the full sample sequence to be recorded with no information loss, facilitating decoupling of simulation design from the constraints of data analysis. The recording uses only one bit per sample, which is an upper bound on the rate at which information about the desired distribution is acquired. We also demonstrate the method and quantify its benefits on a nontrivial colloidal system of charged particles in the canonical ensemble. The method imposes no restrictions on the system or simulation design and is compatible with descendants of the MH algorithm.

This appendix is adapted from an article [1] published in Computer Physics...
Communications. Albert H. Mao, the candidate, invented and implemented the recording method and wrote the paper. The article may be accessed via its digital object identifier (DOI) name 10.1016/j.cpc.2011.03.013. An open access version is also available at arXiv:1105.2266v1. A digital file named RecordingDemonstration.jar is provided along with the article. It is an executable Java archive (JAR) file that implements a demonstration of the recording method and includes full source code, unit tests, and documentation for the implementation. This work was supported by National Science Foundation MCB 0718924 and National Institutes of Health - National Institute of General Medical Sciences 5T32GM008802.

A.1 Introduction

Markov chain Monte Carlo (MCMC) methods enable importance sampling from complicated, concentrated probability distributions. The Metropolis-Hastings (MH) algorithm [2, 3] is the prototype for a class of MCMC algorithms where transition proposals are accepted or rejected to generate each sample. Its applicability to distributions where probability ratios, but not absolute probabilities, are easily computed motivates its broad usage. Since the complexity and scale of these applications routinely push against computational resource limits, practical techniques for minimizing storage requirements and execution time are important aspects of implementations. Ideally, the full sequence of samples is recorded in a compact format that facilitates analysis and interpretation.

Current implementations of Metropolis-Hastings-class algorithms typically reduce the information content of the full sample sequence in two ways before it is recorded. In the first, a small fraction of samples are recorded in complete detail, but the rest are discarded. This enables arbitrary post-analysis because all degrees of freedom are available for the retained samples. However, when sampling systems
with many degrees of freedom, retaining even a small fraction of samples requires considerable storage capacity. Even if capacity is abundant, disk throughput would limit the fraction of samples that could be recorded without having disk write operations dominate the execution time. In the second pattern, averages, moments, histograms, or other quantities are accumulated during simulation and recorded upon completion. This is parsimonious with respect to demands on storage bandwidth and capacity, but decisions about which quantities to accumulate, their frequencies of accumulation, and the number of samples to discard due to starting configuration bias must be made in advance. Altering these choices requires repeating the entire simulation.

These tradeoffs and losses of information complicate the usage of MH-class algorithms in real applications. A typical experience from our lab provides an illustrative example in the area of biomolecular simulation. A study of conformational and dimerization equilibria [4] involved a set of \( \sim 10 \) proteins, each consisting of \( \sim 10^3 \) interacting atoms and modelled at \( \sim 10 \) different temperatures either individually or in pairs. At least three independent replicate simulations were performed for each condition, with each replicate generating \( \sim 10^8 \) samples using MH. While some quantities of interest had their expectation values accumulated during the initial simulations, only 1 in \( \sim 10^4 \) samples had their full set of atomic coordinates recorded. Despite using GROMACS XTC compression with limited precision, these sparse recordings consumed a total of \( \sim 10^{11} \) bytes of storage. Subsequent analyses could only be performed by reanalyzing these recorded samples or repeating the entire set of simulations while accumulating expectations for the new quantities of interest. Given that the simulations consumed \( \sim 10^5 \) total hours of cpu time, both options were suboptimal in terms of efficient usage of computational resources.

These problems can be completely avoided. Here, we describe a simple method
for recording the full sample sequence from a MH-class simulation that stores one bit per sample. The method operates independently from any details of the system and the transition proposals, and is therefore generally applicable.

A.2 Description of the method

Recording a bit string of 1’s for acceptance and 0’s for rejection, with one bit per transition, suffices to preserve the complete information content of all samples generated during one run of a MH-class algorithm. This method follows naturally from an information theoretic perspective: since all influence from the underlying distribution is reduced to a binary accept-or-reject decision, MH-class algorithms can be viewed as communication channels with capacity of one bit per sample [5, chapter 30.5]. An existing implementation of a MH-class algorithm can be easily modified to perform this recording operation during each iteration. Buffered output is necessary because file systems do not allow writing of individual bits, and also helps to reduce the frequency of disk writes.

Regenerating the full sequence of samples for subsequent reanalysis requires the recorded bit string, the starting state, and any pseudorandom number generator seed(s) from the original simulation. The original simulation code should be reused to guarantee that the original sequence of transition proposals is recapitulated, but modified to use the recorded bit string for deciding whether to accept or reject proposals. Since the acceptance criterion no longer needs to be evaluated, all calculations involved in computing ratios of sample weights or proposal probabilities can be skipped during reanalysis. However, as with the original simulation run, all degrees of freedom for every sample are available for computing and accumulating distributions of any quantity of interest.

Care must be taken to avoid corrupting the sequence of samples during re-
analysis. The stream of pseudorandom numbers generated during reanalysis must be identical to the original simulation’s stream. In particular, if a single pseudorandom number stream is used for generating transition proposals as well as evaluating the acceptance criterion, the pseudorandom number that would have been used for the acceptance criterion must be generated and discarded during each iteration. If analysis routines themselves make use of pseudorandom numbers, they must generate their own independent streams. Insidious platform dependencies are another source of potential corruption when recordings generated on one computer are reanalyzed on another. For example, the implementation of floating point arithmetic differs between processor architectures and compilers. This source of error should be eliminated by adhering to best practices in the coding of floating point operations [6] or writing unit tests that assert platform-specific assumptions are valid during both original simulation and reanalysis runs.

A.3 Demonstration of the method

To create a nontrivial demonstration of this recording method, we implemented a MH simulation of a colloidal suspension of charged spherical particles. This three-dimensional off-lattice system consists of $10^3$ particles confined within a spherical droplet. Each particle has two charge sites that freely diffuse on the particle’s surface. The potential energy is the sum of a pairwise Lennard-Jones potential between particle centers and a pairwise Debye-Hückel screened electrostatic potential between charge sites. Intra-particle and inter-particle electrostatic interactions are screened using different dielectric constants. Transition proposals consist of local or full randomization of one particle’s center coordinates or charge site positions. An executable Java Archive file containing source code and compiled class files for this demonstration is available as supplementary material. To run it
on any computer where Java Runtime Environment version 6 is installed, execute `java -jar RecordingDemonstration.jar` at a command line.

A comparison with typical alternatives highlights the efficiency of the proposed recording method. A straightforward format would be a recording of all seven degrees of freedom \((x, y, z, \theta_1, \phi_1, \theta_2, \phi_2)\) for all particles. While this is obviously inefficient, it is simple and easy to parse, facilitating interoperability with other software. Storing the updated values, if any, for only the changed degrees of freedom during a transition would be significantly more efficient. However, this method would tie the recording format to the choice of transition proposals; modifying the simulation to use more sophisticated transitions that simultaneously perturb multiple particles would require redesigning the format. An even more efficient method would record the potential energy (or more generally, the weight) of each sample. As with the proposed method, this would require preserving the simulation code, but would enable full reconstruction of the original samples without calculating any weights. Table A.1 compares the storage efficiency of these recording schemes to the proposed one. `RecordingDemonstration.jar` can derive any of the other recordings starting from the bit string recording, proving that the bit string (along with simulation code and starting state) retains all information about the sample sequence.

Since the bit string recording obviates all sample weight and proposal probability ratio calculations, iterating over the sample sequence takes significantly less time than generating it. To enable benchmarking, `RecordingDemonstration.jar` provides two analyses for the demonstration system. The first is a histogram of the central angle between intra-particle charge sites and the second is a pairwise distance histogram between particle centers. The histograms are accumulated once every 100 and 1000 steps, respectively. Table A.2 compares the execution time for performing
Table A.1: Comparison of efficiencies for different recording methods. Calculated sizes assume that real numbers are represented and recorded using an eight-byte format. Method 2 uses one four-byte integer per sample to encode which particle, if any, changed, and assumes an acceptance rate of exactly 50%. Note that preserving the simulation code is necessary for methods 3 and 4. This would consume an additional amount of storage that is independent of the number of samples, and is not reflected in the table.

Both analyses during the original simulation versus using the bit string recording. Note that energy evaluations are efficiently implemented such that only changing terms are computed; each iteration after the first of the original simulation performs a number of computations that is linear, rather than quadratic, in the number of particles. However, once the bit string is recorded, energy evaluations are skipped; each iteration only needs to update the degrees of freedom for a single particle and therefore executes in constant time.

Table A.2: Comparison of total run times for generating $10^6$ samples of the demonstration system. Benchmarks were performed on a 2.6 GHz Intel Core 2 Duo system with 4 GB of 667 MHz DDR2 SDRAM and 6 MB L2 cache running version 1.6.0_22 of the Java Runtime Environment.
A.4 Discussion

This method enables efficient reanalysis as long as the time spent generating and executing transition proposals is small compared to the time spent computing ratios of sample weights and proposal probabilities. Fortunately, this condition is naturally satisfied because sample weights are generally determined by an energy function or other interaction between degrees of freedom that is more expensive to update than the degrees of freedom themselves. By recording the results of the most computationally expensive component of MH-class algorithm implementations, the method proposed here approaches the minimum achievable limits on both storage consumption and execution time.

A closely related alternative method would be to record the full sequence of sample weights in addition to the accept/reject decision. In some applications, the sample weights themselves are important subjects of analysis, and rederiving them from the samples during reanalysis would be computationally expensive. For instance, in a simulation of a physical system in the canonical ensemble, one may wish to calculate the heat capacity and other properties of the energy distribution. Assuming that weights are represented as IEEE 754 double precision floating point numbers, as in Table A.1, each sample would require an additional 64 bits of storage. While much less compact than the bit string recording alone, it would still be far more efficient than most alternatives.

Note that the size of the recording might be further reduced by compressing it using a lossless algorithm. One bit per sample is an upper bound on the entropy rate [5, chapter 30.5]; if the acceptance rate of the simulation is not exactly 1/2, the recorded bit string will contain biases that a compression algorithm can exploit.

The compactness of this recording method makes it susceptible to corruption:
during reanalysis, a single incorrectly recorded bit contaminates all subsequent samples. An effective solution would be to encode the raw bit string using error correcting codes (ECC) that provide robustness at the cost of increased storage consumption [5, chapter 11]. In fact, employment of Reed-Solomon [7] and low-density parity-check (LDPC) [8] codes at the hardware level is already widespread among designers of random access memory, persistent storage devices, and networking interfaces. Therefore, ECC at the software level would constitute a redundant layer of protection against error. Non-redundant detection of error can be achieved by recording all degrees of freedom for a small number of checkpoint samples, and verifying equality for the corresponding samples generated during reanalysis. Alternatively, the common practice of comparing cryptographic hash function outputs can be applied to verifying the integrity of recorded bit strings.

This method is compatible with descendants of the MH algorithm that retain its general accept-or-reject architecture. Examples include expanded ensemble techniques [9], replica exchange Monte Carlo [10], multiple-try Metropolis [11], simulated annealing [12], and simulated tempering [13]. For some algorithms, such as Wang-Landau sampling [14], efficient post-analysis would require recording the system energy at every iteration as described above.

A.5 Conclusion

The method described here is a simple and effective approach to data storage and representation in the design of MH-class algorithm implementations. It facilitates decoupling of simulation design from the constraints of data analysis. The information theoretic perspective through which this method was conceived deserves broader appreciation in the development of Monte Carlo algorithms.
A.6 References


Appendix B

Mixing Extent: a quantifier of sequence patterning

B.1 Introduction

The results presented in Chapters 2, 3 and 5 generally characterize protein sequences according to their charge composition. Sequence descriptors based on composition are ultimately functions of residue frequency histograms, which are independent of the ordering or patterning of residues within a sequence. However, as demonstrated in Section 2.2.8 and Chapter 5, patterning is significant in that two IDPs with identical sequence composition but different patterning can exhibit distinct conformational ensembles. Therefore, in our search for systematic relationships, it is desirable to quantify patterning in a way that enables meaningful comparisons between sequences regardless of whether they have similar composition. Intuitively, this would make patterning “orthogonal” to composition as an axis of sequence space. This appendix describes the Mixing Extent, a quantity that reflects the extent to which distinct symbols within a string are mixed together. We formulate this quantifier in terms of general mathematical strings, making it applicable to any abstract sequence of symbols from a finite alphabet.

B.2 Definition of the Mixing Extent

The Mixing Extent of a string $S$ is defined as the fraction of permutations of $S$ with fewer runs of identical contiguous symbols than $S$. As an example, the string $\text{CABBAB}$ has five runs: $\text{C, A, BB, A, and B}$. Its Mixing Extent is exactly equal to $2/5$, as
Table B.1: Permutations of the string CABBAB organized by number of runs. There are a total of 60 permutations, and $6 + 18 = 24$ have fewer than five runs. The Mixing Extent of CABBAB is therefore $24/60 = 2/5$.

The number of runs within a string is one plus the number of adjacent pairs of distinct symbols (with the exception of the empty string, which has zero runs and zero adjacent pairs). Intuitively, a “well-mixed” string has many adjacent distinct symbols.

<table>
<thead>
<tr>
<th>3 runs</th>
<th>4 runs</th>
<th>5 runs</th>
<th>6 runs</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 permuations</td>
<td>18 permutations</td>
<td>26 permutations</td>
<td>10 permutations</td>
</tr>
<tr>
<td>AABBBC</td>
<td>AABBCC</td>
<td>ABABBC</td>
<td>ABABCB</td>
</tr>
<tr>
<td>AACBBB</td>
<td>AABCBB</td>
<td>ABACBB</td>
<td>ABCBBA</td>
</tr>
<tr>
<td>BBBAAC</td>
<td>Abbas</td>
<td>ABBAAC</td>
<td>BABABC</td>
</tr>
<tr>
<td>BBCCAA</td>
<td>ABCBBA</td>
<td>ABACBC</td>
<td>BABCBC</td>
</tr>
<tr>
<td>CAABBB</td>
<td>ACABBB</td>
<td>ABBABC</td>
<td>BABABC</td>
</tr>
<tr>
<td>CBBBAA</td>
<td>ACBBBA</td>
<td>ABBBBA</td>
<td>BABCBA</td>
</tr>
<tr>
<td>BAABBC</td>
<td>ABCABB</td>
<td>BACABB</td>
<td></td>
</tr>
<tr>
<td>BAACBB</td>
<td>ABCBBA</td>
<td>BCABAB</td>
<td></td>
</tr>
<tr>
<td>BBAABC</td>
<td>ACBABB</td>
<td>BCBABA</td>
<td></td>
</tr>
<tr>
<td>BAAACB</td>
<td>ACBBBA</td>
<td>CBABAB</td>
<td></td>
</tr>
<tr>
<td>BBBACA</td>
<td>BAAACB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BBCCA</td>
<td>BABBCA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BBCBAA</td>
<td>BABBCA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BCAABB</td>
<td>BACABB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BCBBA</td>
<td>BACBBA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CABBA</td>
<td>BBABAC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAAABB</td>
<td>BBABCA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CBBAAB</td>
<td>BBACAB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BBACBA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BACABA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BCABBA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BCBAAB</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CABABB</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CABBAB</td>
<td>CBABBA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CBBABA</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
and therefore many runs, giving it a high Mixing Extent. The Mixing Extent always
lies within the half-bounded interval $[0, 1)$, and is effectively a “mixedness percentile”
of the string amongst its permutations. This facilitates interpretation and meaningful
comparison between strings with different lengths or compositions. Formulation in
terms of permutations accomplishes the desired separation between composition and
patterning: composition is invariant with respect to permutation, and all strings
with a given composition can be permuted into one another. It also adjusts for the
inherent tendency of strings with lopsided compositions, where some symbols are
significantly more frequent than others, to have fewer runs than strings with more
balanced compositions.

Despite these favorable attributes, the Mixing Extent is only a single num-
ber that barely scrapes the surface of the information present in string patterns. In
particular, it is entirely blind to correlations on a length scale greater than direct
adjacency. A satisfactory description of patterning may require additional measures
that quantify correlations at longer, and possibly all, length scales. We present the
Mixing Extent as a first step towards these generalizations, and recommend its use in
conjunction with other available sequence characteristics including length and com-
position in the search for systematic relationships between IDP sequences and their
conformational ensembles.

B.3 Algorithms for calculating the Mixing Extent

The illustrative approach in Table B.1 of enumerating permutations to cal-
culate the Mixing Extent is infeasible for all but the shortest strings. Consider a
polypeptide containing 380 total amino acids, a number close to the median length
across the human proteome [1]. If the composition is uniform such that each of
the twenty standard amino acids occurs nineteen times, the protein sequence has
380!/\binom{19}{20} \approx 1.9 \cdot 10^{475}$ distinct permutations, an intractable number for any enumerative approach. Since any quantifier of sequence patterning is useless if it cannot be computed for actual strings with relevant lengths, it is necessary to devise an efficient algorithm for calculating Mixing Extents. In this section, we describe several combinatorial techniques that lead to a dynamic programming algorithm with running time that is polynomial in the string length.

Let $S$ be a string with $R$ runs of symbols from an alphabet of size $m$. Using $i$ to denote an index over the alphabet, $n_i$ is the number of symbols of type $i$ within $S$ such that $\sum_{i=1}^{m} n_i = n$, the string length. Our general approach is to iterate over this symbol frequency histogram, counting the number of ways $n_i$ symbols of type $i$ can be simultaneously inserted amongst all possible strings comprised of the previous $i-1$ symbol types that yields a new string with exactly $r$ runs, for all $r < R$. The key insight is that only the length and number of runs in a string, and not the particular sequence of symbols, are necessary for this counting. As long as the new symbols to be inserted are identical to each other and distinct from all symbols already within a string, a run of new symbols will always increase the number of runs by two if inserted within an existing run or by one if placed between existing runs or at either end of the string. This is the basis of a recurrence relation, which we describe in detail later, for computing an array whose elements $C_{i,r}$ count the number of strings comprised of the first $i$ symbol types with exactly $r$ runs. Thus, a potentially immense number of permutations is organized into run number “buckets”, with each bucket only needing to count, rather than enumerate, its contents. The total number of permutations of $S$ with fewer runs than $S$ is equal to $\sum_{r=0}^{R-1} C_{m,r}$. Dividing this quantity by the total number of permutations, which is given by the multinomial coefficient $n! / \prod_{i=1}^{m} n_i!$, yields the Mixing Extent of $S$. 

191
As a brief digression, we relate a series of improvements in the computation of $C_{i,r}$ as an illustration of mutually beneficial exchange between the academic and competitive programming communities. Our initial approach used memoized recursion to compute $C_{i,r}$ from $C_{i-1,r}$ using $O(n^5)$ arithmetic operations. In June of 2011, organizers of the Google Code Jam, an international algorithmic programming contest administered by Google Inc., were searching for difficult and interesting algorithmic problems for use in the competition’s final round. We proposed the calculation of Mixing Extents as a suitable challenge, and the organizers decided to include it with minor modifications. Two of the organizers, Bartholomew Furrow and David Arthur, created a reference solution that uses dynamic programming to compute $C_{i,r}$ in a manner similar to our memoization approach; the bottom-up nature of dynamic programming enabled a more efficient implementation that uses $O(n^4)$ arithmetic operations. The 25 contestants in the World Finals were survivors of four prior rounds that eliminated over 21,900 of their competitors. During the round, they had four hours to solve five problems, of which ours was intended as the easiest. In the end, eleven of the 25 contestants successfully solved this problem, with ten using approaches similar to the one by Arthur and Furrow. The remaining solver, contestant Natalia Bondarenko, surpassed this common approach by devising an algorithm that uses only $O(n^3)$ arithmetic operations.

Here, we describe a version of Bondarenko’s algorithm adapted to the original task of calculating Mixing Extents. The algorithm is expressed using pseudocode in Figure B.1. Initially, before any symbols have been inserted, the base case is an empty string which has only one permutation. Therefore, $C_{0,0} = 1$ and $C_{0,r} = 0$ for all $r > 0$ because there is a single permutation that contains zero runs and zero permutations that contain one or more runs.
In the first nested loop, the outer loop is over each element $i$ of the alphabet, the middle loop is over each possible number of runs $r$, and the inner loop is over each possible number of additional runs $q$ contributed by simultaneously inserting $n_i$ symbols of type $i$. Since $n_i$ is assumed to be greater than zero, and since a run must consist of at least one symbol, $1 \leq q \leq \min(n_i, r)$. Within the inner loop, the count $C_{i,r}$ is increased by the number of ways $n_i$ symbols can be split into $q$ runs and then spliced amongst $r - q$ runs built from the previous $i - 1$ symbol types to yield exactly $r$ total runs. According to the classic “stars-and-bars” explanation [2], the number of ways of splitting is the binomial coefficient $\binom{n_i - 1}{q - 1}$ and the number of ways of splicing is $\binom{r}{q}$. Since each combination of these ways with each permutation built from the first $i - 1$ symbols yields a distinct result, the total increase is equal to the product of $C_{i-1,r-q}$ with both of these binomial coefficients. Upon completion of these loops, $C_{m,r}$ contains the number of permutations of exactly $r$ run objects built from all symbols of $S$. However, it does not yet satisfy our starting definition of $C$ because adjacent runs can be composed of the same symbol. For instance, both "AA AA B BB AA" and "A BB B A AAAA" might be counted within the $C_{2,5}$ bucket, even though both actually contain only three runs.

In the second nested loop, this discrepancy is corrected by inductively subtracting the number of extraneous permutations from $C_{m,r}$ for each $r < R$. If the true number of permutations $C_{m,t}$ with $t$ runs for some $t < r$ is known, it can be used to compute the number of extraneous permutations counted within $C_{m,r}$ that contain exactly $r - t$ “false” boundaries between adjacent runs of identical symbols. A string with $n$ symbols and $t$ true runs has $n - 1$ internal boundaries between symbols, of which $t - 1$ are true internal boundaries between runs. Therefore, $(n-1)-(t-1) = n-t$ internal boundaries separate identical symbols within runs, and the number of ways
to choose \( r - t \) of these false boundaries is the binomial coefficient \( \binom{n-1}{r-t} \). Each way can be combined with each permutation, so the subtraction from \( C_{m,r} \) due to permutations with \( t \) true runs is equal to \( C_{m,t} \binom{n-1}{r-t} \). The inner loop performs this subtraction for each \( t < r \), yielding the correct final value of \( C_{m,r} \) that can be used in the next iteration of the outer loop to correct \( C_{m,r+1} \).

By inspection, the first nested loop performs \( O(mR^2) \) arithmetic operations, while the second nested loop does only \( O(R^2) \) operations. \( m \) and \( R \) are bounded by \( n \), so the entire algorithm uses \( O(n^3) \) arithmetic operations. This analysis implicitly assumes that efficient schemes are used to compute the binomial coefficients: they can be updated using the identity \( \binom{a}{b} = \binom{a-1}{b-1} \frac{a-b+1}{b} \) in the first loop and \( \binom{a}{b} = \binom{a-1}{b-1} + \frac{a}{b} \) in the second loop. Also note that the values \( C_{i-1,r} \) can be discarded once \( C_{i,r} \) has been computed for all \( r \), so the algorithm requires only enough memory to store \( O(n) \) integers. However, since large integers up to \( n! \) may be encountered during execution and in the final results of the algorithm, proper analysis of memory usage and running time must account for storage of and arithmetic on integers that are \( \lg n! = O(n \lg n) \) bits long. The fastest known algorithm for multiplying integers has an asymptotic runtime of \( O(b \lg b \ 2^{\log^* b}) \) \cite{3}, where \( b \) is the total number of bits necessary to represent the two integers. We simplify this to \( O(b \lg b) \) since the iterated logarithm \( \log^* b \) grows so slowly with \( b \) that it can be effectively regarded as constant. Therefore, Bondarenko’s algorithm actually runs in \( O(n^4 \lg^2 n) \) time and uses \( O(n^2 \lg n) \) memory. We have implemented the algorithm and successfully tested it on a string with 3000 symbols, demonstrating that it is efficient both in theory and in practice.
function mixingExtent(n[1] ... n[m], R)

    C[0][0] = 1
    C[0][1 ... (R-1)] = 0
    for i = 1 to m
        for r = 0 to R-1
            C[i][r] = 0
            for q = 1 to min(n[i], r)
                C[i][r] += C[i-1][r-q] * binCoef(n[i]-1, q-1) * binCoef(r, q)
            end
        end
    end

    total = 0
    for r = 0 to R-1
        for t = 0 to r-1
            C[m][r] -= C[m][t] * binCoef(n-t, r-t)
        end
        total += C[m][r]
    end

    return total / multiCoef(n[1] ... n[m])
end

Figure B.1: Bondarenko’s algorithm for calculating Mixing Extents. As in the text, n[i] denotes the number of symbols of type i, m the alphabet size, n the total number of symbols, and R the number of runs in the input string S. binCoef(a,b) denotes the binomial coefficient \( \binom{a}{b} \) and multiCoef denotes a multinomial coefficient. The operators += and -= respectively add and subtract the value of the expression on the right side to the variable on the left side.
B.4 Both folded and disordered proteins exhibit the full range of Mixing Extents

As a preliminary and exploratory test of this sequence patterning measure, we quantify the Mixing Extent distribution across both folded and disordered proteomes. We obtain a representative, non-redundant sample of the folded proteome by extracting the protein sequences from PDB SELECT 25 [4] in January of 2011. Similarly, for the disordered proteome, we extract intrinsically disordered region sequences from version 5.6 of the DisProt database [5]. Both proteomes are filtered to select sequences at least 20 and at most 100 amino acids in length, resulting in 1636 folded sequences and 434 disordered regions. Normalized histograms of the calculated Mixing Extents for both proteomes is shown in Figure B.2. Surprisingly, the histograms are relatively similar to each other and contain counts across the full range of possible values. This implies that the full spectrum of sequence patterns, as quantified by adjacency of identical residues, is present within naturally occurring sequences. In addition, no value of the Mixing Extent appears to be incompatible with either folding or intrinsic disorder. The peak at low mixing extents suggests that there is an increased tendency for identical residues to be adjacent within a sequence. However, the peak height is only about three times the flat, uniform baseline that would result if every sequence was randomly shuffled.

One benefit of a patterning measure defined in terms of abstract strings is the ability to transform the input to emphasize or mask various aspects of a pattern. Here, we coarsen the alphabet to assess the robustness of the distributions shown in Figure B.2 when the criteria for determining symbol equality is relaxed. The twenty amino acids are divided into nine equivalence classes according to physicochemical properties, listed here by their one-letter abbreviation: WFY, ALM, IV, HSTC, QN, ED,
Figure B.2: Normalized histograms of Mixing Extents across folded and disordered proteomes. PDB SELECT 25 represents the folded proteome and DisProt the disordered proteome. The bin width is 0.05.

RK, G, and P. By replacing each amino acid symbol with a unique identifier of its physicochemical class, each protein sequence is converted to a coarse grained alphabet which effectively discounts any boundaries between amino acids belonging to the same physicochemical class. However, as shown in Figure B.3, the Mixing Extent distributions computed for these coarsened alphabets are qualitatively similar to those computed for the full alphabet. This suggests that, across proteomes, some aspects of sequence patterning are directed towards maintaining a distinctive physicochemical profile along the polypeptide chain.
**Figure B.3:** Normalized histograms of nine-element alphabet Mixing Extents across folded and disordered proteomes. The grouping of the twenty standard amino acids into nine classes is given in the text. PDB SELECT 25 represents the folded proteome and DisProt the disordered proteome. The bin width is 0.05.

**B.5 Conclusions**

The Mixing Extent satisfies its design criteria of providing a general metric of sequence patterning for abstract strings that is orthogonal to composition. Efficient algorithms for calculating the Mixing Extent enable its application to real-world sequences and serve as a starting point for developing more detailed patterning measures along with the algorithms for calculating them. The abundance of protein sequences at every Mixing Extent suggests that it will be possible to use it as an order parameter.
in sequence space for both folded and intrinsically disordered proteins. Transformations of the alphabet enable various aspects of patterning to be investigated using the same general measure, and will aid in efforts to understand the relationship between IDP sequence patterning and conformational ensemble properties.

B.6 References


Appendix C

Calculating activity coefficients of ABSINTH electrolytes using grand canonical Monte Carlo simulations

C.1 Introduction

Chemical potentials convey fundamental information about thermodynamic systems. In electrolyte solutions, activity coefficients quantify the deviation of chemical potentials from their ideal values and connect microscopic details of interactions between solution components to macroscopic, measurable properties. Therefore, the calculation of activity coefficients using molecular simulations of electrolyte solutions constitutes a stringent test of the simulated models for solvent, solute, and their interactions. This appendix addresses various practical and theoretical considerations that arise when using grand canonical Monte Carlo simulations to calculate activity coefficients of salts in aqueous solution for direct, quantitative comparison to experimental measurements. Some of the equations presented here are specifically tailored for the ABSINTH [1] implicit solvation model, but many aspects of the discussion are generally applicable. Although the theoretical framework developed here is not fundamentally different from that of previous similar studies [2], we provide greater detail in explicating the mathematical steps leading to the ultimate result.

C.2 Methods

The chemical potential of an electrolyte in solution is unambiguously determined by the solution’s composition and thermodynamic parameters. While its value
may be conventionally subdivided in experimental measurements or theoretical calculations, keeping its fundamental definition in mind can prevent confusion and error caused by the complexity of these conventions. This is true even if its absolute numerical value is never actually measured or calculated. In this appendix, we loosely adhere to the notation and conventions of Robinson and Stokes [3, chapter 2].

C.2.1 Fundamental thermodynamic definitions

Consider a solution consisting of \( n_w \) moles of solvent particles dissolving \( n_s \) moles of solute in thermal equilibrium with a reservoir at temperature \( T \) and pressure \( P \). We consistently use the subscripts \( w \) and \( s \) to refer to the solvent and solute, respectively. In this study, the solute is a salt that completely dissociates into \( t \) distinct ionic species upon solvation; the subscripts \( i \) range from 1 to \( t \) and refer to individual species. All ions of the same species are identical and indistinguishable particles. The salt is characterized by \( t \) stoichiometric integers \( v_i \) giving the number of ions of species \( i \) produced by dissociation of one whole salt; the total number produced is \( v = \sum_{i=1}^{t} v_i \). The solution therefore contains \( n_i = v_i n_s \) moles of ion species \( i \) and \( v n_s \) total moles of ions. The Gibbs free energy \( G(T, P, n_w, n_s) \) is the natural thermodynamic potential for this system, and its partial derivative with respect to \( n_s \) is the salt’s chemical potential:

\[
\mu_s(T, P, n_w, n_s) \equiv \left. \frac{\partial G}{\partial n_s} \right|_{T, P, n_w} \tag{C.1}
\]

Since the addition of one whole salt causes the number of ions of each species to increase by a fixed number, the salt’s chemical potential is the weighted sum of the
individual ion chemical potentials:

\[ \mu_i(T, P, n_w, n_s) = \sum_{i=1}^{t} v_i \mu_i(T, P, n_w, n_s) \]  \hspace{1cm} (C.2)

These individual ion chemical potentials should be regarded as abstract quantities because it would be impractical to titrate the concentration of one ion by itself. Even if this could be accomplished, the results would be specific to the shape of the container due to violation of electroneutrality.

C.2.2 Definition of the activity coefficient and standard states

By convention, the results of electrolyte chemical potential measurements at different concentrations are tabulated as activity coefficients. The natural concentration scale for experiments is the molality \( m = \frac{n}{W_w n_w} \), where \( n \) is the number of moles of solute and \( W_w \) is the molar mass of solvent in kilograms/mol, because it does not vary with pressure or temperature and can be determined without accurate volume measurement. The general relation between absolute chemical potentials and activity coefficients is given in Equation C.3:

\[
\mu(m) = \mu^0(m^0) + RT \ln a(m) = \mu^0(m^0) + RT \ln(m \gamma(m)) = \mu^0(m^0) + RT \ln(m/m^0) + RT \ln(\gamma(m)m^0) \]  \hspace{1cm} (C.3)

This definition uses a hypothetical 1 molal ideal solution as the standard state. \( R = 1.9872 \times 10^{-3} \) kcal/mol is the molar gas constant, \( m^0 \) denotes the standard state concentration of 1 molal, and \( \mu^0 \) is the standard state chemical potential. The
chemical potential $\mu$, activity $a$, and activity coefficient $\gamma$ are functions of concentration, temperature, and pressure, but only the concentration dependence is made explicit in Equation C.3 to simplify the notation. We assume that the temperature and pressure are constant in the remainder of this study.

As an aside, note that texts frequently omit $m^0$ such that the definition is written as in Equation C.4:

$$\mu(m) = \mu^0 + RT \ln m + RT \ln \gamma(m) \quad (C.4)$$

This is technically incorrect because the argument to any function that can be expressed as a power series, such as a logarithm, must be a dimensionless quantity. The concentration $m$ is a physical quantity with units, which implies that $\gamma$ must have units of inverse concentration and contradicts the common statement that $\gamma$ is a dimensionless quantity. In practice, the error does not impact any numerical result because the standard state concentration is always chosen to be exactly one of the appropriate unit. By explicitly dividing the concentration and multiplying the activity coefficient by $m^0$ in Equation C.3, we make the standard state explicit within the definition and maintain its dimensional integrity. This definition is fully compatible with texts that use the erroneous version if it is understood that $\gamma m^0$, rather than just $\gamma$, is the dimensionless activity coefficient.

### C.2.3 Relationship between whole salt and individual ion quantities

The general definition in Equation C.3 applies individually to each ion species:

$$\mu_i(m_s) = \mu^0_i(m^0) + RT \ln(m_i/m^0) + RT \ln(\gamma_i(m_s)m^0) \quad (C.5)$$
The individual ion molalities are multiples \( v_i \) of the overall salt molality, and by Equation C.2 each individual ion chemical potential is multiplied by \( v_i \) in its contribution to the overall salt chemical potential. Therefore, the factors \( v_i \) are prominent in the expression for the whole salt chemical potential:

\[
\mu_s(m_s) = \sum_{i=1}^{t} v_i \mu_i(m_s) \\
= \sum_{i=1}^{t} \left( v_i \mu_i^0(m^0) + v_i RT \ln(v_i m_s/m^0) + v_i RT \ln(\gamma_i(m_s)m^0) \right) \\
= \sum_{i=1}^{t} \left( v_i \mu_i^0(m^0) + v_i RT \ln v_i + v_i RT \ln(m_s/m^0) + v_i RT \ln(\gamma_i(m_s)m^0) \right) \\
\equiv \mu_s^0(m^0) + vRT \ln(m_s/m^0) + vRT \ln(\gamma_s(m_s)m^0) \\
\quad \text{(C.6)}
\]

In the final line of Equation C.6, the whole salt quantities are given their natural definitions in terms of the individual ion quantities:

\[
\mu_s^0(m^0) \equiv \sum_{i=1}^{t} v_i \left( \mu_i^0(m^0) + RT \ln v_i \right) \\
\gamma_s(m_s) \equiv \left( \prod_{i=1}^{t} \gamma_i(m_s)^{v_i} \right)^{1/v} \\
\quad \text{(C.7)}
\]

With these definitions in place, the main subtlety in working with activity coefficients of a strongly dissociating salt is the factor \( v \) that multiplies each term. For a 1:1 electrolyte, \( v = 2 \) and many of the preceding equations become simpler because all the \( v_i = 1 \). Tables of \( \gamma_s \) as a function of \( m_s \) for many different salts have been obtained using a variety of methods [4].
C.2.4 Conversion between concentration scales

We compare activity coefficients calculated using grand canonical Monte Carlo simulations in ABSINTH implicit solvent to the classic data compiled by Robinson and Sinclair [5]. In the grand canonical ensemble, it is the system volume $V$ instead of the pressure that is held constant. This discrepancy is necessary for simulations in ABSINTH implicit solvent because the pressure is ill-defined in the absence of quantified solvent interactions with itself or with container walls. Fortunately, the thermodynamic properties of electrolyte solutions at various concentrations change only mildly over a wide range of pressures [4, 6]; the dependence is weak to the extent that many authors simply omit any mention of pressure. As shown in Chapter 2, ABSINTH has successfully predicted the results of experiments on biomolecular systems performed under atmospheric conditions without any special consideration of the pressure. Therefore, in the thermodynamic limit, it is reasonable to compare electrolyte activity coefficients derived using ABSINTH in a constant volume ensemble simulation against experiments performed under atmospheric pressure [7, 8].

In implicit solvent simulations, the lack of solvent granularity prevents usage of the molality concentration scale because $n_w$ is undefined. Since ABSINTH simulations are performed at constant volume with explicitly represented salt ions, the molarity concentration scale is most suitable. To enable direct quantitative comparisons, we convert the experimental data to the molarity scale, a task which requires the solution mass density $d$ as a function of concentration. Density data for a variety of electrolyte solutions have been compiled as a function of solute mass fraction $f_s$ (which is related to the molality by $m_s = \frac{f_s}{(1-f_s)W_s}$) in the International Critical Tables [9]. We use interpolation to obtain the density at concentrations other than the tabulated ones. Defining $c_s \equiv \langle n_s \rangle / V$ to be the salt molarity, $c^0 \equiv 1$ mol/liter the molar standard
state concentration, $W_s$ the salt molar weight in kilograms, and $y_s$ the molarity scale salt activity coefficient, the necessary conversions are given by Equation C.8:

\[
    c_s(m_s) = \frac{m_s d(m_s)}{1 + m_s W_s}
\]

\[
    c^0 y_s(m_s) = \frac{m^0 \gamma_s(m_s) d(0) m_s}{c_s(m_s)} \quad \text{(C.8)}
\]

Note that $d(0)$ denotes the density of the pure solvent. The conversion formula for the activity coefficient is a consequence of its definition [3], which is analogous to that of the molal activity coefficient:

\[
    \mu^0_s(c^0) \equiv \sum_{i=1}^{t} v_i \left( \mu^0_s(c^0) + RT \ln v_i \right) \quad \text{(C.9)}
\]

\[
    y_s(c_s) \equiv \left( \prod_{i=1}^{t} y_i(c_s)^{v_i} \right)^{1/v} \quad \text{(C.10)}
\]

\[
    \mu_s(c_s) \equiv \mu^0_s(c^0) + vRT \ln(c_s/c^0) + vRT \ln(y_s(c_s)c^0) \quad \text{(C.11)}
\]

**C.2.5 Searching for target concentrations using inverse grand canonical Monte Carlo simulations**

In the grand canonical ensemble, the chemical potential $\mu_s$ is fixed while the number of solutes $n_s$ fluctuates. The simulated concentration $c_s$ is unknown until the ensemble has been sampled adequately to obtain a converged estimate of $\langle n_s \rangle$. Therefore, the general approach is to guess a value of $\mu_s$, run the simulation to calculate $c_s$, and solve for $y_s(c_s)$:

\[
    y_s(c_s) = \frac{1}{c_s} \exp \left( \frac{\mu_s - \mu^0_s(c^0)}{vRT} \right) \quad \text{(C.12)}
\]
Numerical values of the standard state chemical potential $\mu^0_s(c^0)$ can be calculated using Equation C.9 and the ideal solution chemical potential, which is the sum of the ideal gas chemical potential and the solvation free energy $\Delta G_i^{\text{solv}}$ [10]:

$$\Lambda_i \equiv \frac{h}{\sqrt{2\pi W_i R T}}$$

$$\mu^0_i(c^0) = RT \ln(c^0 \Lambda_i^3) + \Delta G_i^{\text{solv}}$$  \hspace{1cm} (C.13)

Care must be taken to maintain dimensional consistency when evaluating Equation C.13, where $\Lambda_i$ is the thermal de Broglie wavelength for ion species $i$, $h$ is Planck’s constant, and $W_i$ is the molar mass of species $i$. In fact, these numerical values are never required because the simulation is more conveniently formulated in terms of the relative chemical potential $\mu^\text{rel}_s \equiv \mu_s - \mu^0_s(c^0)$ compared to $\mu_s$, as shown later.

In practice, we wrap the approach described above in a search procedure that repeatedly attempts entire grand canonical simulations using different guesses for $\mu_s^\text{rel}$ until the target molarity $c^\text{target}_s$ is attained. This is necessary for direct quantitative comparisons at the precise converted concentrations of the tabulated experimental data; interpolating either the calculated or measured activity coefficients would otherwise be necessary. Specifically, we pick a target number $n^\text{target}_s$ and set the volume such that $n^\text{target}_s/V = c^\text{target}_s$, and search until $\langle n_s \rangle \approx n^\text{target}_s$. Using the same $n^\text{target}_s$ while varying $V$ to attain different target concentrations keeps the sampling quality and computational expense more consistent across the concentration range compared to altering $n^\text{target}_s$ at fixed $V$. Despite its computational cost, we chose this grand canonical search-based approach over the more straightforward canonical ensemble Widom insertion technique [11] because it exhibits much lower sensitivity to system size [12]. This is significant because activity data are gathered from measurements
of macroscopic systems which are intractable to simulate in atomistic detail. Not all ensembles are created equal in their ability to approximate the thermodynamic limit [13].

To optimize the search procedure for \( \mu_{s}^{\text{rel}} \), we combine two methods that intelligently update its guessed value based on results from simulations using its current value. Both of these methods have been individually demonstrated to be effective at converging toward a given target concentration given a poor initial guess [14].

The first method uses the excess chemical potential at the attained concentration \( \mu_{s}^{\text{ex}}(c_s) \equiv vRT \ln y_s(c_s) = \mu_{s}^{\text{rel}} - vRT \ln(c_s/c_0) \) to approximate the excess chemical potential at the target concentration, leading to the iterative update scheme:

\[
\mu_{s}^{\text{rel, next}} = \mu_{s}^{\text{rel}} + vRT \ln \frac{n_{s}^{\text{target}}}{\langle n_s \rangle}
\] (C.14)

The second method uses fluctuations of \( n_s \) to compute derivatives of \( \langle n_s \rangle \) with respect to \( \mu_{s}^{\text{rel}} \) and extrapolate the desired value using Newton-Raphson iteration. We extend this approach by computing higher order derivatives and applying a clamped version of Halley’s method [15], a third-order generalization of Newton-Raphson iteration:

\[
F_0 \equiv \langle n_s \rangle - n_{s}^{\text{target}}
\]
\[
F_1 \equiv \frac{\partial \langle n_s \rangle}{\partial \mu_{s}^{\text{rel}}} = \frac{1}{RT} \left( \langle n_s^2 \rangle - \langle n_s \rangle^2 \right)
\]
\[
F_2 \equiv \frac{\partial^2 \langle n_s \rangle}{\partial (\mu_{s}^{\text{rel}})^2} = \frac{1}{(RT)^2} \left( \langle n_s^3 \rangle - 3\langle n_s^2 \rangle \langle n_s \rangle + 2\langle n_s \rangle^3 \right)
\]
\[
\mu_{s}^{\text{rel, next}} = \mu_{s}^{\text{rel}} - \frac{F_0}{F_1} \left/ \max \left( 1, 1 - \frac{F_0}{F_1} \frac{F_2}{2F_1} \right) \right. \] (C.15)

The clamping is introduced because estimates of the \( n_s \) distribution’s higher-order moments are more sensitive to limited sampling and take longer to converge than
the mean. With the clamping in place, the information from the second derivative may decrease, but never increase, the magnitude of the update relative to Newton-Raphson, preventing an imprecise estimate from wildly altering the guess. We then combine the updated guesses $\mu_{s,rel,next}^{rel}$ from each method by simply averaging them. Other groups have successfully demonstrated approaches where information from previous iterations, instead of just the current one, is incorporated when updating the guess [12, 16].

C.2.6 Formulation of grand canonical Monte Carlo for whole salt insertion and deletion

In the grand canonical Monte Carlo simulations that form the core of the search procedure, we allow only simultaneous insertions or deletions of all ions comprising a whole salt. Although it is possible to formulate the ensemble in a way that allows for insertion and deletion of individual ions [17], working exclusively with whole salts has several advantages. It reduces the dimensionality of the search from $t$ variables $\mu_{i,rel}^{rel}$ to one variable $\mu_{s,rel}^{rel}$. In addition to greatly simplifying the search space, it avoids the need for imposition of extrathermodynamic correction terms that penalize deviation from electroneutrality; without this penalty, the individual ion chemical potentials can harbor opposite and compensating errors that mutually amplify each other and prevent the search from converging [12]. Although forbidding individual ion insertions and deletions quenches the momentary fluctuations in net charge that must exist in a finite subvolume of a solution, these fluctuations vanish in the thermodynamic limit. Since experimental measurements are performed on macroscopic systems, the whole salt formulation is most appropriate for comparison with experimental data.

We now derive the acceptance probabilities for whole salt insertion and deletion Metropolis Monte Carlo moves in terms of $\mu_{s,rel}^{rel}$. The derivation closely follows the
reasoning of Valleau and Cohen [2], who adopted Norman and Filinov’s interpretation of the grand canonical ensemble in terms of indistinguishable, unlabeled particles [18]. For simplicity, we limit the derivation to ions with no internal degrees of freedom. In general, a sampling move that takes the current state $0$ to a proposed state $1$ in a Metropolis-Hastings simulation has an acceptance probability $A_{0\rightarrow 1}$ given by Equation C.16 [19], where $M_{0\rightarrow 1}$ denotes the conditional probability of proposing state $1$ given that the system is in state $0$ and $P$ the probabilities of the states in the stationary distribution of the Markov chain:

$$A_{0\rightarrow 1} = \min \left( 1, \frac{M_{1\rightarrow 0} P_1}{M_{0\rightarrow 1} P_0} \right) \quad \text{(C.16)}$$

Consider a grand canonical microstate of the system represented by $(n_s, \mathbf{r}^{vn_s})$, where $n_s$ is redefined here to be the number (rather than number of moles) of whole salts and $\mathbf{r}^{vn_s}$ represents the $vn_s$ position vectors of all the individual ions. If $\Xi$ denotes the grand canonical partition function, $\beta \equiv \frac{1}{RT}$ the inverse temperature, and $E(n_s, \mathbf{r}^{vn_s})$ the energy of the microstate, the stationary probability $P(n_s, \mathbf{r}^{vn_s})$ of the microstate is given by Equation C.17:

$$P(n_s, \mathbf{r}^{vn_s}) = \frac{1}{\Xi} \left( \prod_{i=1}^{t} \Lambda_i^{-3vn_s} \exp(\beta \mu_i v_i n_s) \right) \exp(-\beta E(n_s, \mathbf{r}^{vn_s})) d\mathbf{r}^{vn_s}$$

$$= \frac{1}{\Xi} \left( \prod_{i=1}^{t} \Lambda_i^{-3vn_s} \right) \exp \left( \beta \left( \mu_s n_s - E(n_s, \mathbf{r}^{vn_s}) \right) \right) d\mathbf{r}^{vn_s} \quad \text{(C.17)}$$

Note that explicit multiplication by the volume element $d\mathbf{r}^{vn_s}$ makes both sides of the equation a probability, not a probability density. In the Monte Carlo move set, the overall probability of performing an insertion move must equal that of performing a deletion move; call this probability $M_{\text{grand}}$. From the state $(n_s, \mathbf{r}^{vn_s})$, proposed
insertion of one whole salt involves placement of \( v \) ions at uniformly random positions with the simulation volume, resulting in a state \( (n_s + 1, r^{v(n_s+1)}) \). A factor of \( 1/v_i! \) must also be multiplied for each species to maintain the unlabeled quality of the resulting state. The proposal probability (and not probability density) of the post-insertion state \( (n_s + 1, r^{v(n_s+1)}) \) is therefore given by Equation C.18:

\[
M_{n_s \rightarrow n_s+1} = M_{\text{grand}} \prod_{i=1}^{t} \frac{1}{v_i! V v_i} d r^v_i = M_{\text{grand}} \left( \prod_{i=1}^{t} \frac{1}{v_i!} \right) \frac{1}{V v} d r^v \quad (C.18)
\]

The “backwards” probability of proposing a whole salt deletion that produces the unlabeled original state \( (n_s, r^{v n_s}) \) from state \( (n_s + 1, r^{v(n_s+1)}) \) is one divided by the product of the binomial coefficient that counts the number of ways to choose \( v_i \) ions to delete from a total of \( v_i(n_s + 1) \) candidates:

\[
M_{n_s+1 \rightarrow n_s} = M_{\text{grand}} \prod_{i=1}^{t} \left( \frac{1}{v_i(n_s+1)} \right) \quad (C.19)
\]

Defining the energy change \( \Delta E_{n_s \rightarrow n_s+1} \equiv E(n_s \pm 1, r^{v(n_s+1)}) - E(n_s, r^{v n_s}) \), the combinatorial ratio \( W(n) \equiv \prod_{i=1}^{t} v_i! \binom{v_i}{v_i} = \prod_{i=1}^{t} \frac{n v_i!}{(n v_i - v_i)!} \) and combining Equations C.16, C.17, C.18, and C.19 leads to the acceptance probability for insertion moves in Equation C.20:

\[
A_{n_s \rightarrow n_s+1} = \min \left( 1, \frac{M_{n_s+1 \rightarrow n_s} P(n_s + 1, r^{v(n_s+1)})}{M_{n_s \rightarrow n_s+1} P(n_s, r^{v n_s})} \right)
= \min \left( 1, \frac{V v}{\prod_{i=1}^{t} \Lambda_i^{3 v_i} (v_i(n_s+1))/v_i!} \exp (\beta (\mu_s - \Delta E_{n_s \rightarrow n_s+1})) \right)
= \min \left( 1, \frac{V v}{W(n_s + 1) \prod_{i=1}^{t} \Lambda_i^{3 v_i}} \exp (\beta (\mu_s - \Delta E_{n_s \rightarrow n_s+1})) \right) \quad (C.20)
\]

\( M_{\text{grand}} \) and \( \Xi \) cancel out and are absent from the resulting formula, which is appro-
appropriate for a properly designed Metropolis simulation move. In addition, note that
careful attention to the distinction between probability and probability density was
necessary to see how the volume elements present in the stationary distribution and
the insertion proposal probability combine and cancel to yield a dimensionless quan-
tity. The derivation of the acceptance probability for a whole salt deletion move is
essentially the same, except insertion becomes the “backwards” move and $n_s$ rather
than $n_s + 1$ is the number of whole salts that are candidates for deletion. A special
case arises when $n_s = 0$, when the deletion move must be categorically rejected:

$$A_{n_s \rightarrow n_s - 1} = \begin{cases} 0 & \text{if } n_s = 0 \\ \min \left( 1, \frac{W(n_s) \prod_{i=1}^{\text{vol}} A_{n_s_i}}{\prod_{i=1}^{\text{vol}} v_i^{n_i}} \exp (\beta(-\mu_s - \Delta E_{n_s \rightarrow n_s - 1})) \right) & \text{otherwise} \end{cases}$$

(C.21)

Finally, we expand $\mu_s = \mu_s^0(c^0) + \mu_s^{\text{rel}}$ and substitute the expression for $\mu_s^0(c^0)$ from
Equations C.9 and C.13 to eliminate the standard state chemical potential and ther-
mal wavelengths from the acceptance probability formulae. The final forms of these
probabilities, which are suitable for implementation in code, are given in Equa-
tion C.22, where the whole salt solvation free energy $\Delta G_s^{\text{solv}} = \sum_{i=1}^{t} v_i \Delta G_i^{\text{solv}}$:

$$A_{n_s \rightarrow n_s - 1} = \min \left( 1, \frac{W(n_s) \prod_{i=1}^{\text{vol}} v_i^{n_i}}{\prod_{i=1}^{\text{vol}} v_i^{n_i}} \exp (\beta(-\mu_s^{\text{rel}} - \Delta G_s^{\text{solv}} - \Delta E_{n_s \rightarrow n_s - 1})) \right)$$

$$A_{n_s \rightarrow n_s + 1} = \min \left( 1, \frac{(c^0 V)^v \prod_{i=1}^{t} v_i^{n_i}}{W(n_s + 1)} \exp (\beta(\mu_s^{\text{rel}} + \Delta G_s^{\text{solv}} - \Delta E_{n_s \rightarrow n_s + 1})) \right)$$

(C.22)

Note that the quantity $(c^0 V)^v \prod_{i=1}^{t} v_i^{n_i}$ is constant over the course of the simulation
and only needs to be computed once. As always, care must be taken to ensure di-
mensional consistency when computing it; if the simulation volume $V$ is represented
in cubic Angstroms, then $c^0 = 6.02214 \times 10^{-4} \text{ Å}^{-3}$ must be expressed as the number

212
per cubic Angstrom that corresponds to the standard concentration 1 mol/L. In AB-SINTH, the solvation free energies are conveniently available because they are input parameters to the solvation model. Generalizations of these formulae to molecular ions with structure and internal degrees of freedom are possible, but the proposal probability density for the selection of those internal degrees of freedom during insertion (analogous to $1/V$ for the position) must be known and incorporated.

**C.2.7 Ensemble sampling and analysis methods**

In grand canonical Monte Carlo simulations, insertion and deletion moves are interleaved with particle translation moves in order to enhance sampling of the ensemble. Our implementation contains a standard single-particle move, which picks a random ion and randomly perturbs its position. In 90\% of these moves, the proposed position is uniformly selected at random from a 4 Å radius sphere centered at its original position, while in the remaining 10\%, the proposed position is fully randomized over the entire simulation volume. To enhance sampling of the system, we also implemented two moves that simultaneously perturb the position of multiple ions. One is a “cluster move”, which selects a random subset of the ions present in the system with between 2 and 5, inclusive, elements. The subset of ions is then proposed to translate by up to 2.5 Å in a common random direction, rotate by up to $\pi/32$ radians around a random axis through their centroid, or both. The other multi-particle move is a “scoop move”, which draws a random sphere with radius between 3 and 6 Å within the simulation volume and proposes that all ions within the sphere rotate about a random axis through the sphere’s center by up to $\pi$ radians.

We set the probability of attempting each type of move at each Monte Carlo step to follow the ratio 80 insertion/deletion : 100 simple translation : 10 cluster : 1 scoop.

Initial trials showed that setting $n_s^{\text{target}} = 250$ achieves a good balance between
approaching the thermodynamic limit and minimizing computation time. We adopt several heuristics to reduce the computation time spent on the initial iterations of the search procedure, when \( \langle n_s \rangle \) is likely to be farthest away from \( n_s^{\text{target}} \). First, we use an initial guess of \( 0.6/c^0 \) for the activity coefficient, which corresponds to \( vRT \ln(0.6c^{\text{target}}/c^0) \) for \( \mu_s^{\text{rel}} \), because it tends to make \( \langle n_s \rangle \) small and therefore leads to greater simulation throughput, as the computation time generally scales with the number of particles present in the simulation volume. We also use shorter simulations with lengths of \( 10^5 \) production steps for the first three iterations and \( 5 \times 10^5 \) steps for the fourth iteration, because this is sufficient to obtain a useful update to the guess when the guess is far away from the desired answer. Simulation lengths of \( 5 \times 10^6 \) production steps are used from the fifth iteration onward. All iterations include an additional \( 10^5 \) equilibration steps that occur before the production steps and are discarded from any analyses. Once the equilibration period is over, the \( n_s \) histogram \( N[n_s] \) is accumulated every 50 steps. At the end of each iteration, this histogram is used to compute the central moments in Equation C.15 via symmetric unbiased estimators based on k-statistics [20]. Letting \( N \equiv \sum_{n_s=0}^{\infty} N[n_s] \) denote the total counts in the histogram and \( S_p \equiv \sum_{n_s=0}^{\infty} n_s^p N[n_s] \) the sum of the \( p \)-th powers, Equation C.23 gives the formulae for these estimators:

\[
F_0 = \frac{S_1}{N} - n_s^{\text{target}} \\
F_1 = \frac{1}{RT} \frac{NS_2 - S_1^2}{N(N-1)} \\
F_2 = \frac{1}{(RT)^2} \frac{2S_3^3 - 3NS_2^2 + N^3 S_2}{N(N-1)(N-2)} 
\]

(C.23)

The search proceeds for a minimum of five iterations and continues until \( \langle n_s \rangle \) is within 2.5 of \( n_s^{\text{target}} \), a relative tolerance of 1\%. When this convergence criterion is satisfied,
the guess is updated one final time and $\mu_{rel, next}^s$ is returned as the final answer.

## C.3 Results

### C.3.1 Simulations in ABSINTH overestimate the chemical potential of electrolytes at high concentrations

We implemented the ABSINTH solvation model and the methods described above in a simulation engine specifically designed for calculating activity coefficients in the grand canonical ensemble. In a departure from previous work using ABSINTH, we adopt the Lennard-Jones parameters derived in Chapter 4 and single-ion hydration free energies compiled by Schmid et al. [21]. We also use the “crystal radii” instead of “ionic radii” advocated by Shannon [22] to override the Lennard-Jones $\sigma$ for solvation shell overlap calculations. The natural logarithm of the calculated activity coefficient, which is proportional to the excess chemical potential, is presented as a function of concentration and compared against experimental measurements for NaCl and KCl in Figure C.1. These plots show that the activity coefficients calculated using ABSINTH are generally too high compared to experimental measurements. The discrepancy for concentrations around 125 mM is minor in that the error in excess chemical potential is less than $RT$, but it becomes more severe at higher concentrations.

Both the calculated and measured curves for $\ln(c^0 y_s)$ exhibit the qualitative form of a convex function of concentration that passes through a minimum. For sodium chloride, Rösgen et al. have shown [23] that this behavior can be explained by the combination of classic Debye-Hückel behavior and packing effects. As the concentration is increased, the excess chemical potential first decreases due to increasingly favorable electrostatic interactions, then increases as steric exclusion begins to dominate the interactions at high concentration. Like the primitive model, ABSINTH accounts for both of these effects, but packing effects appear to come into play at
Figure C.1: Comparison of simulated and experimental activity coefficients for NaCl and KCl as a function of molarity.
concentrations that are too low. Since the $\sigma$ and $\epsilon$ parameters for the short-range Lennard-Jones interaction do not represent drastically swollen ion sizes relative to their crystal radii, the apparent packing effects are caused by the 5 Å thick ABSINTH solvation shell that surrounds each ion. In ABSINTH, overlapping solvation shells causes a decrease in the direct mean field solvent interaction along with a concomitant de-screening of the electrostatic interaction. The premature upward growth of the activity coefficient implies that the former effect is not being adequately compensated by the latter effect, leading to an overestimate of the chemical potential at medium to high concentrations.

C.3.2 ABSINTH solvation parameters can be adjusted to improve agreement with experimental measurements

Taking advantage of ABSINTH’s tunable design, we adjust the solvation model parameters in an attempt to improve agreement with experiment. We focus on the parameters $\tau_d$ (FOSTAU) and $\chi_d$ (FOSMID), which respectively govern the cooperativity and midpoint of the switching curve that maps solvent accessible volume fraction to solvation state for the direct mean field interaction. This choice is motivated by a desire to minimize the impact of any changes on systems such as polypeptides that ABSINTH models accurately using its original parameters. Since the reference free energies of solvation are an order of magnitude more favorable for ions compared to model compounds that constitute proteins, altering $\tau_d$ and $\chi_d$ should have a disproportionately greater influence on ions compared to proteins. In addition, initial trials showed that calculated activity coefficients are minimally sensitive to Lennard-Jones parameters and reference solvation free energies. Figure C.2 shows that the calculated activity coefficients for NaCl and KCl with $\tau_d = 0.059$ and $\chi_d = 0.6$, compared to their respective original values of 0.25 and 0.1, are much closer to experimental data.
Figure C.2: Activity coefficients for NaCl and KCl as a function of molarity from experiments and simulations with $\tau_d = 0.059$ and $\chi_d = 0.6$. In addition to the tuned AB-SINTH implicit solvation parameters, the simulations also use an earlier version of the lattice-calibrated Lennard-Jones parameters for the ions compared to the simulations in Figure C.1. The experimental data are identical to that plotted in Figure C.1.
Since these new solvation parameters are obtained by fitting to the activity data, we also test them on the protamines from Chapter 2 for which both simulation and fluorescence correlation spectroscopy data are available. Unfortunately, the new parameters completely disrupt the results for sequences with low net charge per residue, turning collapsed globules into swollen coils. This implies that no single set of ABSINTH solvation parameters simultaneously produces accurate activity coefficients for electrolyte solutions and accurate behavior for IDPs. The difficulty of simultaneously reproducing activity data and other observables has been encountered in other attempts to develop models for ions in implicit solvent. For example, Lenart et al. [24] fit a dielectric saturation model to activity data, but this required Lennard-Jones parameters that caused the first peak of the cation-anion pair correlation function to occur at a physically unreasonable distance below that of a molten salt.

C.3.3 Decreasing the effective radius of solutes improves agreement with experiments

As a baseline assessment of the ion parameters in the absence of ABSINTH implicit solvation effects, we calculate activity coefficients using a simplified model with only Lennard-Jones interactions and screened Coulomb interactions using a dielectric constant of 78.36. In fact, this is equivalent to setting the ion radius used in ABSINTH solvation shell overlap calculations to zero, as this prevents all ions from influencing each other’s solvation shells and therefore results in full solvent accessibility and maximum screening. Figure C.3 plots the activity coefficient curves for NaCl and KCl, which exhibit much better agreement with experimental measurements and excess chemical potentials accurate to within $RT$. In fact, the NaCl curve actually lies beneath the experimental curve, supporting the assumption that dominance of
packing effects seen with the full ABSINTH Hamiltonian is not due to Lennard-Jones parameters. These baselines may serve as a starting point for the development of enhancements that enable ABSINTH to accurately quantify the balance between desolvation and descreening that occurs as ions approach each other. While the baselines are already quantitatively close to the experimental measurements, any enhancements will require enough sophistication to account for the decreased activity coefficients of KCl relative to NaCl at high concentrations despite the larger size of K$^+$ compared to Na$^+$. One possible avenue is the representation of partially desolvated states that enable ions to form stable pairs and low-order clusters without tipping the balance towards formation of insoluble crystals. However, such an approach would increase the number of parameters necessary to specify an ion, and would thus benefit from detailed experimental measurements of the cluster size distribution. Explicit solvent simulations may also provide guidance for these efforts, as laborious free energy calculations have revealed that they exhibit reasonable accuracy in reproducing measured activity coefficients [25].

C.4 Conclusions

The combination of grand canonical Monte Carlo simulations with a search over the input chemical potential is an effective method for calculating activity coefficients of electrolytes in implicit solvent models. Careful and explicit accounting for definitions and standard states are necessary to achieve meaningful quantitative comparisons between simulation and experiment. The ABSINTH implicit solvent model shares the ability of simpler models to predict activity coefficients of electrolyte solutions at low concentration. Its overestimation of activity coefficients at high concentrations can be mitigated by tuning the solvation parameters; however, these modifications lack transferability to other systems. These findings suggest that
Figure C.3: Activity coefficients for NaCl and KCl as a function of molarity from experiments and simulations using a constant dielectric model. The experimental data are identical to that plotted in Figure C.1.
modifying ABSINTH to explicitly account for discrete states of intermediate desolvation may enable it to capture the essential interactions that govern both dilute and concentrated salt solutions. Success along these lines would constitute progress towards a comprehensive molecular understanding of salt water, a goal that has remained elusive over centuries of investigation despite the fundamental importance of the system.

C.5 References


Appendix D

Contributions to other scientific works

This appendix is a compendium of the candidate’s contributions that are not described elsewhere in this dissertation.

D.1 Forces for restraining the radius of gyration during molecular dynamics simulations

We performed multiple umbrella sampling simulations to obtain the potential of mean force (PMF) as a function of $R_g$, the radius of gyration, for polyglycine and glycine-serine block copolypeptides in water and urea. This requires the imposition of a restraint potential that keeps the radius of gyration near its target value $R_g^0$:

$$U_{\text{restrain}}(R_g) = \frac{k}{2} (R_g - R_g^0)^2,$$

where $k$ is a harmonic force constant. In molecular dynamics simulations, the derivative of this potential with respect to each atomic coordinate gives the negative component of the corresponding force along that coordinate. I calculated this force and implemented it by modifying the source code of GROMACS 3.3.1 [1]. In Equation D.1, $n$ is the number of atoms in the molecule to be restrained, $m_i$ is the mass of atom $i$, $r_{i,d}$ denotes the $d$-th component of particle
\( \vec{r}_i \)'s position vector, and \( r_{i,d}^{cm} \) is the \( d \)-th component of the molecule’s center of mass.

\[
M = \sum_{i=1}^{n} m_i \\
I = \sum_{d=1}^{3} \sum_{i=1}^{n} m_i (r_{i,d} - r_{i,d}^{cm})^2 \\
-\frac{\partial U_{\text{restrain}}}{\partial r_{i,d}} = -k \left( \frac{1}{M} - \frac{R_g^0}{\sqrt{IM}} \right) m_i (r_{i,d} - r_{i,d}^{cm}) 
\] (D.1)

The PMFs and other results were published in a paper authored by Hoang T. Tran, Albert Mao and Rohit V. Pappu titled Role of Backbone-Solvent Interactions in Determining Conformational Equilibria of Intrinsically Disordered Proteins [2]. These restraining forces were also used in the explicit solvent calculations described in Section 2.3.2.

**D.2 A reflection move for Monte Carlo simulations that connects kinetically distant proline pucker states**

The \textit{exo} (\( C_{\gamma}\)-up) and \textit{endo} (\( C_{\gamma}\)-down) puckering states of proline rings are separated by an energetically unfavorable planar state. This manifests as a long time scale for interconversion between the puckering states in experiments and broken ergodicity in molecular dynamics or Monte Carlo simulations that only allow local perturbations of the ring degrees of freedom. To preserve ergodicity and enable Monte Carlo simulations to accurately capture the equilibrium thermodynamics of proline puckering, I proposed a simple Monte Carlo move that allows “tunneling” between the energetically favorable pucker states. The move negates all dihedral angles describing the proline ring while keeping all bond lengths and angles constant. This has the effect of “reflecting” the ring across the plane that is roughly
formed by all its atoms except Cγ, thereby converting an instantaneous conformation that belongs to one pucker state into the same conformation in the opposite pucker state. The move was implemented in CAMPARI by Adam Steffen and is described and used in a paper authored by Aditya Radhakrishnan, Andreas Vitalis, Albert H. Mao, Adam T. Steffen, and Rohit V. Pappu titled Improved Atomistic Monte Carlo Simulations Demonstrate That Poly-L-Proline Adopts Heterogeneous Ensembles of Conformations of Semi-Rigid Segments Interrupted by Kinks [3].

D.3 Simulations of the C-terminal tail of single-stranded DNA binding protein (SSB C-t) from Escherichia coli and its fragments

*E. coli* SSB is an example of an IDP that contains a folded core with a disordered C-terminal tail. The disordered tail is 62 residues long and has the sequence QGGAPAGGNIQGGQPQGQPPGGQFGGQFSGQAQRQPQPQQQPPQNEPPMDFDDDIPF. It is rich in both polar and acidic residues, with glycine and glutamine being the two most common amino acids, and the aspartates clustered at the tip of the tail are highly conserved [4]. The phase diagram from Section 2.2.13 predicts that the full tail will adopt compact globular ensembles based on its low net charge per residue. However, the tip of the tail, which is rich in acidic residues and necessary for SSB’s protein interactions [5], would be predicted to adopt extended coil ensembles if sixteen or fewer of its residues are considered in isolation. We prepended the two residues immediately N-terminal to the disordered tail, a glycine and an arginine, to its sequence and divided the resulting 64-residue sequence into eight blocks each containing eight residues. The 36 possible contiguous string of blocks (eight sequences with one block, seven sequences with two blocks, six sequences with three blocks, etc.) were then sim-
ulated as disordered regions in isolation. The simulations showed that all fragments adopted disordered globular conformations.

D.4 An algorithm for performing inverse Maxwell constructions

Scott Crick measured the saturation concentration of various polyglutamine constructs as a function of temperature in order to build a concentration-temperature phase diagram for polyglutamine. We interpreted saturation concentrations as left binodals according to Flory-Huggins solution theory, which requires fitting the Flory interaction parameter \( \chi \) to yield a left binodal that matches the measured saturation concentration. Since performing a Maxwell construction on the Flory-Huggins free energy functional with a given value of \( \chi \) yields the binodal, this fitting procedure can be considered an inverse Maxwell construction. The essence of the algorithm is to wrap the entire Maxwell construction procedure inside a numerical search that takes advantage of the monotonic relationship between \( \chi \) and the position of the left binodal. An initial guess for \( \chi \) is iteratively refined based on whether the left binodal that results from performing a Maxwell construction is above or below the measured saturation concentration. Once the correct value of \( \chi \) is determined, a final Maxwell construction yields the position of the right binodal, and the left and right spinodal are determined analytically. I devised and implemented this algorithm, which was used to generate phase diagrams that will be the focus of a forthcoming paper co-authored by Scott Crick and Rohit Pappu.
D.5 Strategy for elucidating the role of sequence contexts in modulating functional motifs within intrinsically disordered regions

I designed the explanatory figure for a perspectives article [6] discussing the use of phylogenetic relationships to discover functional motifs within intrinsically disordered regions [7]. In addition to describing the work of Nguyen Ba et al., the article advocates the combination of proteome-wide functional annotation with biophysical modeling and machine learning to understand how sequence context modulates function within intrinsically disordered regions.

D.6 References


