A Molecular View of Self-Assembly and Liquid-Liquid Phase Separation of Intrinsically Disordered Proteins

A Molecular View of Self-Assembly and Liquid-Liquid Phase Separation of Intrinsically Disordered Proteins

by

Gregory L Dignon

Presented to the Graduate and Research Committee
of Lehigh University
in Candidacy for the Degree of
Doctor of Philosophy
in
Chemical Engineering

Lehigh University
January, 2020

Approved and recommended for acceptance as a dissertation in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Date

Dissertation Advisor

Committee Members:

Jeetain Mittal, Committee Chair

Anand Jagota

Srinivas Rangarajan

Marcos Pires

Young Chan Kim

Acknowledgements

First I would like to thank my family, my parents Fred and Nancy, my siblings, Danielle, Kelly and Tim as well as their families, and the abundance of friends I have made over the years who have been great support throughout my time at Lehigh. I would also like to acknowledge my church, Graceway Community Church for providing me guidance, as well as a home and a family here in Bethlehem. I would like to acknowledge my girlfriend Shelby for listening to me talk about my research, and for sharing her journey through her own PhD studies with me. I would also like to thank the professors in the chemical engineering department at Lehigh who have provided me with excellent instruction, and to my thesis committee for their time and for their scrutiny, helping me to understand my own research at greater depth. I would also like to acknowledge the current and former members of the Mittal lab, Gül Zerze, Hasan Zerze, Runfang Mao, Ann Maula, Vahid Rahmanian, Elif Irem Senyurt, Lata Rani, Wai Shing Tang, Myrto Perdikari, Chris Rzepa, Roshan Regy, Nina Jovic, Evan Pretti and Huilai Gu who have given me many opportunities to learn, teach, and discuss scientific topics that we enjoy. A special thanks goes to Gül for putting me through the initial training when I started, and providing an excellent example to model my own research style after. I would like to also thank my fellow group members, and the members of the Rangarajan group for their constructive feedback on my presentations in group meetings. I would like to acknowledge the many great collaborators I have had the privilege to do work with, Young C Kim, Nicolas Fawzi, Robert Best, Wenwei Zheng, Ben Schuster, Matt Good, Dan Hammer, and Kristi Kiick and their students. Finally, I would like to thank my PhD advisor, Jeetain Mittal for his daily and continued guidance throughout the progression of my project, for giving me excellent instruction and honest, cosntructive feedback, and for helping me to realize my potential and to become a competent researcher.

Contents

List of Tables

List of Figures

Abstract

Intracellular compartmentalization of biomolecules into non-membrane-bound compartments, commonly referred to as membraneless organelles (MLOs), has been observed for over a century. The past decade has seen a massive surge of research interest on this topic due to evidence that a liquid-liquid phase separation (LLPS) process is responsible for the assembly of biomolecules into liquid-like compartments constituting MLOs. Since the initial discovery, dozens of cellular systems have been explored with this in mind, and have also been shown to have liquid-like properties, owing several unique functions to their liquid-like nature. MLOs such as stress granules may spontaneously form in response to cellular stress, while others such as the nucleolus may form multi-layer architectures that accelerate multi-step assembly processes, similar to an assembly line. The ability of biomolecules to undergo LLPS has been largely attributed to the presence of disordered proteins and nucleic acids. Intrinsically disordered proteins (IDPs) are proteins which lack a native, folded structure while remaining physiologically functional are able to promote LLPS due to their polymeric nature which allows for transient multivalent interactions between many amino acids. In this thesis, I work toward a greater understanding of the relationship between an IDPs sequence, and its ability to undergo LLPS. Using a combination of all-atom and coarse-grained simulations, I make several important contributions of significant and general interest to the field of IDP-driven LLPS. I start by developing a coarse-grained modelling framework which explicitly represents amino acid sequences, and is the first of its kind to directly simulate phase coexistence of IDPs at sequence resolution. I then leverage this model to demonstrate the relationship between a single IDP chain, and a condensed phase of the same IDP, showing that one can predict conditions where LLPS will be possible, simply by observing the single-chain behavior and infinitely-dilute two-chain binding affinity. I then provide a rationalization of the

thermoresponsive behavior of some proteins which undergo lower critical solution temperature (LCST) phase transitions, and how amino acid composition can lead to different thermoresponsive behaviors. Finally, I present atomic-resolution simulations showing the different interaction modes responsible for driving LLPS of two particular IDPs.

Chapter 1

Introduction, Motivation and Background

1.1 Biomolecular Phase Separation[1]

Membraneless organelles (MLOs), while having been first observed over 150 years ago[1], have recently gained much interest since several demonstrations showing that many MLOs have liquid-like properties[2, 3], and form through a process of liquid-liquid phase separation (LLPS)[4, 5]. LLPS has been shown to occur in a variety of biological contexts, facilitating a wide array of functions requiring compartmentalization[6–8] including pathological formation of disease-causing aggregates[6, 9–12]. MLOs differ from membrane-bound organelles in their ability to spontaneously form and dissipate[2, 13], and their permeability[14, 15]. To gain a greater understanding of the normal and pathological functions of MLOs requires a clear view of the molecular interactions underlying LLPS, and how different biomolecules may contribute to the process of phase separation.

1.1.1 Role of Intrinsic Disorder in Phase Separation

Recent studies have linked protein intrinsic disorder to membraneless organelles (MLOs), showing that the proteome for MLOs has a significantly greater fraction of proteins containing intrinsi-

[1] Adapted from article currently under review.

cally disordered regions (IDRs) than the overall proteome[16]. Intrinsically disordered proteins (IDPs) are proteins which do not adopt a stable folded structure, yet are able to carry out biological functions[17]. They are also highly abundant, composing a large fraction of the eukaryotic proteome[18–21]. Different classes of proteins generally tend toward being disordered, such as typical IDPs which are rich in charged amino acids[19], elastin-like polypeptides (ELPs), which are more enriched in hydrophobic amino acids[22] and prion-like domains[10, 23, 24] which generally have a simple repetitive sequence, and are composed largely of only a few different amino acid types. Each of these are generally enriched in glycine or proline residues which disfavor formation of normal secondary structures.

IDPs have been suggested as being important to LLPS because of their ability to form many contacts with one another simultaneously, having high multivalency[25]. Indeed, the length of an IDP has been shown to correlate with its ability to phase separate[15, 26], having phase diagrams in agreement with polymer theories such as Flory-Huggins[27, 28]. In vitro studies have shown that a fully disordered protein may undergo LLPS, and that it remains disordered while in the phase-separated state[27, 29], forming weak interactions promiscuously between all types of amino acids[30]. Another advantage of IDPs is that the amino acids are more exposed, and therefore, more accessible to post-translational modification[31] which is a major regulator of biomolecular phase separation[32–36].

When studying IDPs, researchers may infer characteristics of phase behavior from simple theory, simulation or experiment from the properties of single chains[13, 37]. This is because the same interactions driving compaction (or otherwise) of a single chain tend to also stabilize the protein-rich phase in LLPS, thus the degree of collapse is expected to be correlated with the propensity to phase separate. Single molecule experiments, scattering experiments or simulations of IDPs can be used to identify the average size of an IDP in solution[38–40]. This is related to the backbone flexibility of the protein[41], as well as the overall strength of interactions between its amino acids, and the overall chain length.

1.1.2 Role of Folded Domains in Phase Separation

In addition to IDPs and IDRs, folded proteins and domains contribute to many functions of MLOs. Folded domains involved in LLPS include RNA-recognition motifs (RRMs) which bind to specific sequences of RNA[13, 42], oligomerization domains[15, 43, 44], and other domains which carry out the intended function of the MLO such as metabolic catalysis[45], promoting gene expression[46], and recruiting specific cargo molecules[15, 34]. There are some cases where IDRs have even been shown to inhibit LLPS while the folded domains are the major driving force of phase separation[13].

Going back many decades, X-ray crystallography studies have observed liquid-liquid phase separation at some conditions during screening for protein crystallization[47]. This, however, generally requires very high protein concentration, and rather extreme conditions, unlike many IDRs which may phase separate at much lower concentrations[26, 35]. Folded proteins may aggregate or crystallize, leaving only a small window of conditions where LLPS may occur[48]. However, IDRs, including those involved in LLPS, have also been known to be prone to aggregation and formation of disease-causing inclusions[11, 17, 49].

Many proteins involved in LLPS, such as ribonucleoproteins (RNPs) include multiple folded domains tethered together with disordered linkers[24]. These folded domains may contribute significantly to phase separation by oligomerizing multiple protein molecules together and effectively increasing the multivalency and number of interactions a single “particle” is able to form[43]. Taking advantage of this, researchers have engineered proteins including a light-activated oligomerization domain[50], thus enabling induction of phase separation in a controlled manner inside living cells[46, 51]. RNPs even more commonly include RRMs which selectively bind to particular regions of RNA and can promote LLPS in the presence of these particular RNA sequences[13, 42, 52]. Partially folded structures may also contribute to the phase separation through folding upon binding to specific binding partners[53, 54], or promotion of secondary structure upon self-association[55]. Inclusion of short helical motifs within an ELP also contributes to phase separation with significant hysteresis[56], having a considerably higher saturation temperature (T_{sat}) upon heating compared to cooling.

Some folded domains in phase separating proteins are relatively passive and do not contribute appreciably to the ability of the protein to phase separate. Such domains may have an orthogonal function, such as enzymes[45, 57] and RNA-remodelling helicase domains[58, 59]. Many studies use protein constructs containing green fluorescent protein (GFP) or other similar fluorescent protein domains in order to visualize LLPS within cells[32, 42]. Importantly, different fluorescent tags have been shown to incorporate into droplets of the LAF-1 RGG with different preferences[15] indicating that in some contexts, the inclusion of fluorescent tags may alter LLPS.

Another advantage of IDPs that makes them preferable for driving LLPS is that they can interact promiscuously with a large number of binding partners (this may be why they occur at protein interaction network "hubs" with a significantly higher frequency than folded proteins[60, 61]). IDPs may interact with other IDPs, sometimes with very high affinity, while remaining fully disordered and having no specific bound complex[62]. Many IDPs also interact with folded domains in a specific manner, by adopting a folded structure, usually via an induced fit mechanism[17, 53, 54], though they may simply interact with folded domains and remain disordered[63, 64]. Self-complementary RNA structures also play a role in LLPS by imparting an identity to the MLO it is incorporated in, and preventing merges with other MLOs containing different folded RNAs[65, 66].

1.1.3 Amino Acid Sequence-Dependence of LLPS

Since weak multivalent interactions between disordered and folded protein domains are the major driving forces of phase separation, it is important to understand exactly what are the different modes of interaction which cause proteins to assemble, demixing from their normal solvated state. Biology has provided proteins with an incredible arsenal of amino acids with differing side-chain chemistries, and an even more extensive library of PTMs[67]. The result is hundreds of different types of amino acid derivatives in addition to the 20 canonical amino acids, and many different possible interaction modes arising from these[24, 68–70]. A full understanding of each of the interaction modes, and how they contribute to or detract from a system's ability to phase separate needs to be well understood in order to appreciate the implications of protein composition and sequence.

Proteins and nucleic acids make use of diverse chemistries in order to drive self-association and incorporation or exclusion of other molecules, and to control dynamical and transport properties of phase-separated assemblies[24, 30, 67, 71–75]. Some amino acids may interact through several different interaction modes, which may work cooperatively to provide even stronger binding[30]. Since the components of MLOs are highly dynamic, and usually disordered, it is a major challenge to directly determine which interaction modes are contributing, and the relative importance of each to LLPS[29, 33]. Atomic-resolution simulations provide a promising path forward for this as they can be used to observe all of the different interaction modes directly [30, 35, 76–78]. The current challenge in using simulations is the cost of running atomic-resolution simulation on a large assembly of many proteins, and so far this has only been achieved in one study[78]. Studies identifying interactions within proteins may be useful in identifying which interactions contribute most significantly, and how small perturbations can be made to the sequences to greatly alter the macroscopic phase behavior, and how naturally occurring mutations may have significant physiological and biophysical repercussions[11, 23].

In addition to amino acid composition, the arrangement and sequence of amino acids also has an important role in driving phase separation of biomolecules. Many decades of research have been devoted to relating protein sequence to structure, but for disordered proteins, the role of sequence is not as well understood[19, 79]. This is largely due to a lack of structural information from experiment, thus necessitating alternative measurements to be used for IDPs such as size measurements (R_g, R_h and ν), and their propensity to self-associate, aggregate, or phase separate. Indeed, phase separation may serve as an excellent descriptor for studying the sequence determinants of disordered proteins, nucleic acids, and their interactions[80, 81].

To explore the effects of the arrangement of charged amino acids, Das et al. used all-atom implicit solvent simulations to demonstrate the wide range of single molecule behaviors for a set of proteins having identical composition[79]. The specifically designed sequences are composed of 25 positively charged lysine residues, and 25 negatively charged glutamate residues arranged differently throughout the sequence, with extremes being a strictly alternating dipeptide repeat ($[\mathrm{KE}]_{25}$) and a block copolyampholyte ($[\mathrm{K}]_{25}[\mathrm{E}]_{25}$)[79]. To quantify the degree of charge segregation, they devised a parameter, κ where sequences with low κ values have well-mixed charges, and are more extended, while sequences with high κ values are more blocky (charge-segregated) and are more collapsed[79]. Sawle and Ghosh then used mean field theory to support this assessment, and proposed an alternative charge patterning metric, termed sequence charge decoration (SCD)[82]. A large negative value of SCD indicates a block copolyampholyte, while a value near zero would indicate an alternating polyampholyte. SCD is more general than κ, as it works reasonably well for non-polyampholyte sequences[83], and may also be positive for sequences having a non-zero net charge. Since the single-chain size of the protein is generally correlated with its phase separation propensity, it also follows that the charge patterning values of these polyampholytic sequences are highly correlated with their critical temperatures as calculated by theory[84] and simulation[85, 86]. Importantly, charge patterning has been demonstrated to have a significant impact on the phase separation of biological IDPs such as Ddx4, which contains 25% charged amino acids[27, 87]. While these two charge patterning metrics are highly correlated with each other, some sequences may be designed where one metric predicts collapse, while the other predicts extended configurations. For such sequence, neither metric is effective at predicting its single-chain or LLPS behavior[85], demonstrating that both may be limited in their predictive capabilities in special circumstances. Other metrics have also been used to quantify degree of charge patterning such as charge fluctuations[88], and average "run" length[89].

There is also evidence that the effects of patterning are non-negligible for other types of interactions, particularly involving hydrophobic and aromatic amino acids. Previous studies have looked at the arrangement of hydrophobic and hydrophilic amino acids to find that folded proteins need to have particular patterning in order to collapse properly[90]. Computational work has also shown how sequence correlation of hydrophobic residues can be used to design disordered proteins of identical hydrophobic and hydrophilic composition with very different single chain behavior[91], which should also result in differences in phase separation[92]. A hydrophobic correlation parameter may be used similarly to the SCD and κ metrics in order to make predictions about IDP single chain behavior and phase separation, though when comparing, it will also be important to take into account the distance dependence of such interactions[93]. Some disordered proteins with a very blocky nature may even function as biological surfactants, promoting mixing of polar and nonpolar molecules[94].

Patterning of amino acids is very important to a protein's ability to phase separate, and also toward biological functions through molecular recognition and condensate selectivity. Lin et al. have shown that differences in charge patterning between polyampholytic sequences results in drastically different partitioning of the two components into two or three distinct phases where sequences which are similar cooperatively phase separate, while drastically different patterning results in separation of the two into separate phases[95]. In addition, sequence patterning may also serve as a mode of recognition between disordered proteins and folded proteins with patchy surfaces[63, 96] which could result in additional selectivity of condensates for both disordered and folded domains.

1.1.4 Stimulus-Response of Phase-Separated Granules

Many perturbations have been shown to alter the phase separation of diverse proteins[30, 42, 46, 97, 98]. Depending on the amino acid sequence, different stimuli may either promote or disrupt phase separation, and may change the way in which multiple components mutually or exclusively phase separate. In Fig. 1.1A we show an example of a 2-component system at different values of a control variable which could be temperature, salt concentration, pH, etc. The two components have different propensity to phase separate as a function of the control variable, where component A may phase separate at a wider range of conditions than B. Fig. 1.1A shows the single component binodal phase diagrams of each of the two components, having a region where LLPS is permitted at low values of the control variable, and a region above the critical point where the system is in a single continuous phase. At conditions where both may phase separate, it is likely that droplets may form that contain both components, however, in the region between the two critical points, it is unclear whether component B will be able to incorporate into droplets of A. To visualize this further, Fig. 1.1B; shows a 2-component phase diagram at three conditions highlighted in Fig.1.1A. We observe a cooperative condensation of both components into a single condensed phase for the two lowest values of the control variable (green & orange), and at conditions where only A may phase separate, we observe a scaffold-client phase diagram (yellow) where a condensed phase of component A incorporates a finite concentration of B, even though B is incapable of phase separating on its own at such conditions. This example shows how altering conditions in solution may lead to different phase behaviors, and that some components may still be incorporated into condensates even if they would not undergo LLPS in isolation.

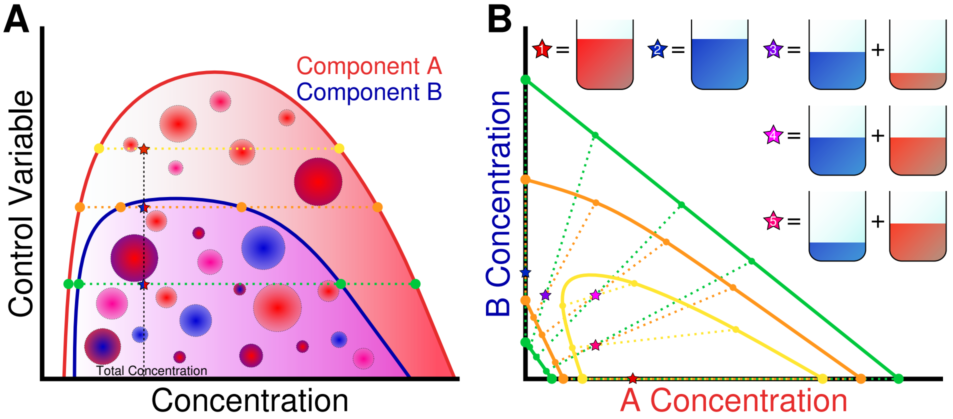

Figure 1.1: Co-phase separation of two species with similar self- and cross-interactions. A) Single component phase diagrams for Component 1 and Component 2 with tie lines at three values of the control variable: both components may phase separate (green); both 1 and 2 may phase separate, but 2 is nearing its critical point (orange); and the region where only 1 is able to form a condensed phase. B) Multicomponent phase diagram of mixtures of Components 1 and 2, with control variable indicated by color. Stars indicate different experiments conducted at different relative total compositions of the two components where 1 and 2 only contain a single component, and 3-5 contain a mixture of the two. Tie lines show the resulting concentrations within the two phases.

While cells generally exist in a narrow range of temperature, changes to interactions in response to temperature is still very important to understanding membraneless organelles, particularly the role of different interaction modes due to each one's distinct temperature response[99]. An increase in temperature may induce phase separation of proteins through active processes which accelerate at higher temperatures, or through thermodynamically-driven, and reversible LLPS[97]. The thermodynamically-driven phase separation which is promoted at higher temperatures results in a lower critical solution temperature(LCST)-type phase transition. Garcia-Quiroz and Chilkoti have provided a comprehensive characterization of composition-dependent phase behavior of IDPs, demonstrating that sequences containing many polar and aromatic amino acids generally are follow an upper critical solution temperature (UCST) phase transition, while those containing more hydrophobic amino acids follow LCST transitions[26]. Some protein sequences such as An16-resilin even show a reentrant phase behavior, where LLPS occurs at both low and high temperatures with a region of miscibility in between[100], typically referred to as having an hourglass phase diagram[101, 102].

With increasing temperature, the loss of chain entropy accompanying IDP collapse or phase separation will increase, which itself may fully explain the UCST phase transitions. However, to observe an LCST phase transition, it is important to consider the temperature dependence of solvent-mediated interactions, particularly with the different types of amino acids[103]. Privalov and Makhatadze conducted solvent transfer experiments on small molecule analogues of the 20 different amino acid side chains and backbone to show that the free energy of solvation is temperature-dependent, and increases with temperature[104]. Amino acids becoming more insoluble with increasing temperature could, in principle, overcome the effect of chain entropy and allow for LCST phase transitions. The difference in temperature-dependence between different types of amino acids based on how different interaction modes are strengthened or weakened by temperature would explain the ability to switch between UCST and LCST phase transitions based on overall composition. Indeed, temperature-dependent solvation free energy of hydrophobic molecules is also non-monotonic, having an initial increase of "hydrophobicity" up to a turnover point, and subsequent decrease[105], which may be explained by the dominance of enthalpy at low temperatures, and entropy at high temperatures[106, 107]. By fitting this temperature dependence to a thermodynamic equation, Dill et al. developed a theoretical model to explain protein folding and thermal stability, showing that temperature-dependent interactions can be used to explain increasing stability with increasing temperature, and cold denaturation[108]. Models such as this one can be quite helpful in elucidating the sequence determinants of temperature-controlled phase behavior[109].

Salt is an important solution additive that can be used to tune phase separation as it is easily controlled in vitro, and is perhaps more physiologically relevant than large changes in temperature. Early studies have shown that for some proteins, increasing salt concentration may induce (salt-out) phase separation[29, 55] or prevent (salt-in) phase separation[27, 110]. An obvious effect of increasing ionic strength is the screening of electrostatic interactions[111]. However, for a sequence such as the low complexity (LC) domain of FUS which is nearly devoid of charged amino acids, it is unlikely that charge screening is the only factor contributing to the large effect salt concentration has on its phase separation. Brangwynne et al. discuss several other interaction modes which are affected differently by salt concentration[93]. The identity of the salt ions also plays an important role in the effects on phase separation. The Hofmeister series ranks different anions and cations by their ability to solubilize or precipitate proteins[112], and have been shown to have the same effect on the phase separation of elastin-like peptides[113]. Ions with higher valency such as transition metals also will also facilitate phase separation for biomolecules having an abundance of the opposite charge[48, 114]. With all of these considerations, salt may present one of the most tunable handles to perturb phase separation.

Another stimulus likely to induce changes to phase behavior is the solution pH. Kroschwald et al. showed that reduced pH as a result of cellular starvation causes the formation of stress granules (SGs) in vivo and in vitro[42]. Interestingly, they found that pH-induced phase separation results in SGs which are more dynamic than those induced by heat-shock, implying that different assembly mechanisms may be occurring[42]. Changes to a system's pH will undoubtedly alter the charges within a protein, and depending on the types and arrangement of charged residues within the sequence, this may also have strong a impact on the charge patterning of the sequence. Thus pH might also be used as a powerful tool for tuning LLPS and selectivity through altered net charge and charge patterning[101].

Other factors that the cell uses to control LLPS are ATP[4, 98] and poly ADP-ribose (PAR)[115, 116]. One process by which ATP may modulate phase separation is by acting as a hydrotrope, thus solubilizing nonpolar groups of the proteins and preventing, or reducing phase separation when driven by hydrophobic interactions[117]. Another consideration is that ATP-driven reactions, particularly phosphorylation of amino acid side chains may also drive or prevent phase separation[118, 119]. PAR may seed phase separation, particularly in cases where phase separation is required to aid in repair of damaged DNA[115].

LLPS may also be mediated by various small molecules such as 1,6-hexanediol[12, 46, 73, 120, 121], chemical chaperones[122], and large molecular crowders including polyethylene glycol (PEG)[123–125]. Braun et al. look at effects of solvent isotope content, and find that D_2O promotes hydrophobicity-driven phase separation of BSA more strongly than H_2O, having important implications for NMR experiments which commonly use D_2O as a solvent[48]. Oxidation also plays an important role in LLPS in various ways. Reed et al. show that oxidative croslinking of a cysteine residues within a designed oleosin protein can facilitate LLPS by promoting dimerization and increasing the multivalency of the protein[126]. They also find that the position of the cysteine residue within the sequence can control the degree to which LLPS is promoted[126]. In contrast, oxidation of methionine side chains has been shown to prevent LLPS of the yeast ataxin-2 protein[127]. Thus, oxidizing and reducing environments are able to promote, or abrogate LLPS depending on protein compositions.

The stimulus-response of biomolecular condensates may currently be somewhat unpredictable, but an understanding of the molecular interactions underlying LLPS, and how such interactions are perturbed by various stimuli will go a long way in making it more predictable. This, however, is very challenging, and it is not clear what is the best way to quantify interactions. Solvation free energy is a useful metric, but only considers the interactions of the amino acid with solvent, and not with other amino acids[108], and use of small molecule analogs may also neglect the polymeric effect[104]. Alternative strategies which may provide the field with much-needed insights include bioinformatics[128, 129], or experimental[130] or computational[131, 132] characterization of binding energies of all amino acid pairs, and their dependences on relevant stimuli.

1.2 Computational Techniques for Phase separation[2]

1.2.1 Spatiotemporal resolutions of simulations

Biomolecular simulations can be conducted using highly diverse models and at different resolutions, each with their own strengths and weaknesses. In general, there is a trade-off between model detail and simulation efficiency, meaning that the greater the accuracy, the smaller the system that can be simulated. An appropriate resolution must be sufficiently detailed in order to accurately capture the properties of interest, while also being computationally tractable. Fig. 1.2 provides a summary of the different spatial resolutions applied to proteins and protein phase separation.

The most detailed resolution is quantum mechanics (QM) which explicitly accounts for electrons and orbitals, and is mostly limited to very small systems, such as short peptide sequences. Hybrid methods such as quantum mechanics/molecular mechanics (QM/MM) may be used to

[2] reproduced from ref.[133]

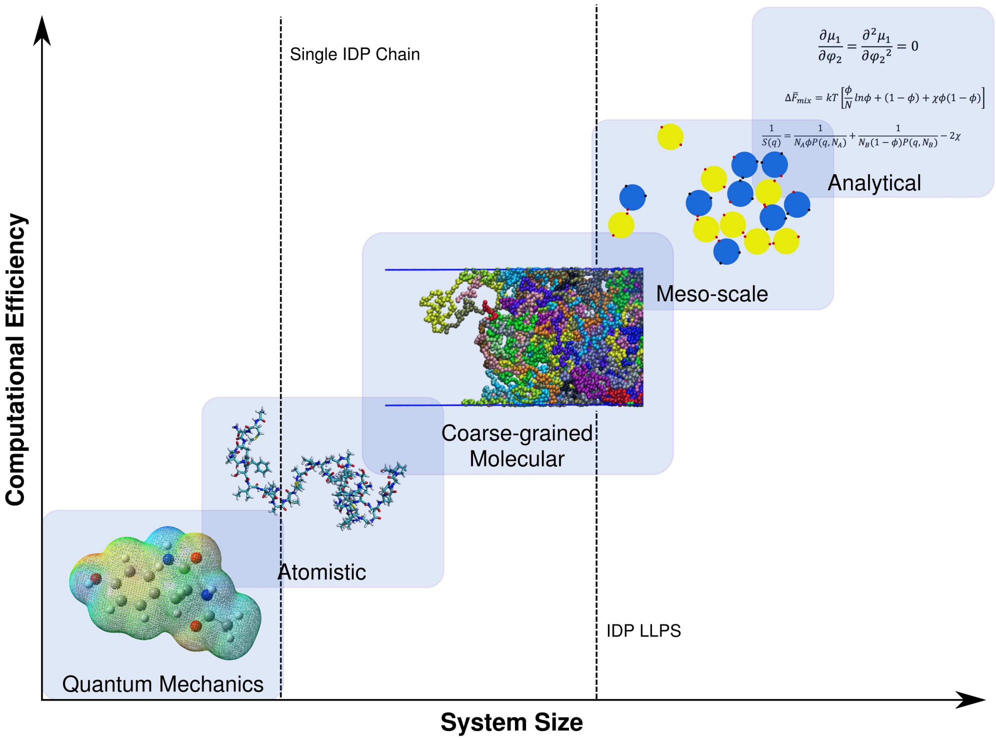
Single IDP Chain
$\frac{\partial \mu_1}{\partial \varphi_2} = \frac{\partial^2 \mu_1}{\partial {\varphi_2}^2} = 0$
$\Delta \bar{F}_{mix} = kT\left[\frac{\phi}{N} ln\phi + (1-\phi) + \chi\phi(1-\phi)\right]$
$\frac{1}{S(q)} = \frac{1}{N_A \phi P(q,N_A)} + \frac{1}{N_B(1-\phi)P(q,N_B)} - 2\chi$
Analytical
Meso-scale
Coarse-grained
Molecular
Atomistic
Quantum Mechanics
IDP LLPS
Computational Efficiency
System Size

Figure 1.2: Overview of simulation resolutions. Higher-resolution techniques give more detailed insight into chemical systems, but are hindered by low computational efficiency, and limits to system sizes which can be studied.

extend the size of the system, however QM regions are still limited to several hundred atoms[134]. One important contribution from QM calculations in the study of disordered proteins has been in constructing classical force field for atomic resolution simulations[135]. QM methods can also be used to improve and extend existing atomistic force fields to include non-canonical amino acids such as those with post-translational modifications[67].

Reducing the level of complexity to classical mechanics, the highest resolution can be achieved by using all-atom simulations with explicit representation of solvent molecules. These simulations have been applied to many biomolecular systems due to their high level of detail, and sufficient computational efficiency to consider a single IDP consisting of several hundred amino acids, or a small assembly of shorter IDPs. Such simulations are able to provide structural characteristics of protein sequences[55, 136, 137] as well as detailed information on inter-residue interactions[78, 103] in good agreement with experimental measurements[138]. Rauscher et al. simulated 27 copies of a 35-residue elastin-like peptide (ELP) for a combined simulation time of 165 μs and verify the proposed liquid-like nature of ELP assemblies, showing that association is driven by nonspecific hydrophobic contacts and hydrogen bonds[78]. To date, this is the only such study for a large assembly of phase separating IDPs at atomic resolution. However, the system size and amount of sampling required to converge on reasonable results are highly cumbersome, and thus, inaccessible to most research groups. One way to reduce this barrier to sampling would be to develop/use atomic resolution force fields with implicit solvent using mean field theory [139].

To further overcome the obstacle of simulating large IDP assemblies, coarse-grained (CG) models are commonly employed in which a group of atoms may be represented collectively as a coarse-grained "bead"[140]. The degree to which a protein can be coarse-grained is flexible, and ranges from multiple beads per amino acid to multiple amino acids per bead[141], following the same trade-off described earlier between model detail and simulation efficiency. CG models commonly account for interactions between protein and solvent molecules implicitly by modifying the protein-protein interactions accordingly, further reducing the computational cost. CG models can also be system specific, being optimized to the experimental data of one particular system[142, 143], or can be more general-purpose, focusing on transferability and applicability to all IDP sequences[144–146]. Simulations of proteins in CG representation have been successfully applied to the study of IDP phase separation and assembly, including multiple beads per residue[56, 147], single bead per residue[88, 144], and multiple-residues-per-bead[148–150]. For the purpose of elucidating sequence-encoded phase separation, the balance lies at a single-bead-per-residue (residue-level) model which minimizes the computational cost while explicitly representing amino acid sequences. Dignon et al. proposed a general purpose, residue-level model which considers IDPs as flexible chains, ignoring secondary structure, and accounts for all 20 canonical amino acids based on either amino acid hydrophobicity[144] or bioinformatics-based contact potentials[151]. This model has successfully been implemented to reproduce sequence-dependent phase behavior of disordered proteins [33, 35]. This framework also accommodates for introduction of non-canonical amino acids, improved interaction potentials, and imposition of secondary structure either through rigid body constraints, or combined angle and dihedral potentials[144]. To date, the residue-level CG model is the most detailed model that can practically simulate the IDP phase coexistence.

Considering many of the proteins involved in biomolecular LLPS are intrinsically disordered[93], polymer theories may also be applicable to the problem. Lin et al. combined Flory-Huggins theory with a random phase approximation (RPA) and successfully captured the interactions between charged amino acids[83]. They further saw a strong correlation between the radius of gyration (R_g) of their corresponding critical temperature, observing phase separation for polyampholytic chains with different charge patterning [37].

1.2.2 Advanced sampling of phase coexistence

Even with well chosen models, efficient sampling of phase behavior remains a non-trivial task. One classic strategy is to improve sampling by constraining chains of polymers onto a simple lattice. Brute force lattice Monte Carlo simulations have been used at residue-level to study the phase behavior of short polyampholytic sequences to determine the effects of charge patterning on phase separation[86]. Other studies used a much coarser model, representing multi-domain proteins and RNAs as chains of interaction sites on a lattice, and parameterized to specifically capture behaviors observed in experiment[148, 149, 152]. Representing particles on a lattice, however, will be limited in its ability to capture densities in the condensed phase[85]. Representing chains off-lattice would therefore provide a more accurate representation protein chain, which we find to justify the additional challenge to sampling.

Another common approach for sampling phase coexistence is grand canonical Monte Carlo (GCMC) which involves attempting insertions and deletions of molecules randomly[153]. One weakness of GCMC is that the acceptance probability of inserting into a liquid-density phase drops rapidly as chain length increases, making study of IDPs prohibitive without the use of lattice coordination and/or other enhanced sampling techniques. One such technique is configurational bias Monte Carlo[154], which can be used to find the “holes” in the dense phase. On-lattice GCMC simulations using conformational biasing have been successful for systems of polymers up to 1000 residues[155]. Jacobs et al. utilize on-lattice GCMC with multicanonical biasing, and observed the effects of interaction strengths and number of unique components on phase separation, showing that intermolecular interactions have a greater influence than the number of components[156]. Another popular method is Gibbs ensemble Monte Carlo (GEMC), in which particles are modeled in two separate boxes of varying size where particles may be transferred from one to the other, thus yielding two continuous "bulk" phases in coexistence[155].

Another efficient method to simulate the phase behavior is to use a slab geometry, where two coexisting phases are simulated in an elongated simulation box with periodic boundary conditions, having two planar interfaces perpendicular to the elongated axis[85, 144, 157]. This strategy can be used with virtually any representation of IDPs on- or off-lattice. Jung et al. have also shown the use of slab geometry on multi-component systems, leading to convergent results in excellent agreement with studies using semi-GCMC and GEMC[157]. These observations suggest that the slab sampling method could be highly beneficial to the study of LLPS of IDPs and other biomolecules.

1.3 Thesis Organization

In this thesis, I discuss my work using computational methods at atomistic, and coarse-grained resolutions to determine the behaviors of IDPs in different conditions and environments, and how they undergo self-assembly and phase separation. First, I discuss the interactions of a disordred protein, hIAPP, with lipid bilayers of different compositions, and demonstrate that association of IDPs with lipid membranes may induce demixing of different lipid types. I then discuss the development of a coarse-grained modeling framework, specifically designed to interogate the sequence-to-phase-behavior relationship of IDPs. It is the first coarse-grained model to our knowledge that couples sequence-specificity with a sufficiently coarse-resolution to directly observe phase coexistence, and to calculate binodal phase diagrams representing LLPS. This model highlights the differences in interactions that are coupled with phosphorylation of particular residues of the FUS LC protein, and also allows us to demonstrate the slowing effect of a large globular helicase domain in LAF-1 phase separation. I then take this coarse-grained modelling framework and demonstrate how the fundamental relationship between intramolecular interactions driving single-chain compaction, and intermolecular interactions driving LLPS can be determined in a consistent way. We show that the θ-solvent temperature is highly correlated with the critical temperature of phase separation, and that this correlation is not specific to our model, or even to our coarse-grained modeling framework, suggesting it likely holds true in experiment. In addition to single-chain properties, two-chain properties, as measured by the osmotic second virial coefficient (B_{22}) can also be used to predict conditions where LLPS may or may not be observed. These insights are of huge relevance to the field of biomolecular condensates as it provides simple descriptors that can be calculated to determine whether LLPS should be possible at the tested conditions, and provides experimental and theoretical researchers a high-throughput method of quickly testing many sequences and conditions for how they may alter LLPS propensity. Next, building upon the original coarse-grained model I introduce a temperature-dependent interaction potential for the different types of amino acids, which more accurately represents the solvent-mediated interactions between amino acids. By including this temperature-dependent term, the model is able to distinguish between protein sequences which undergo LCST or UCST phase transitions from a large library of sequences tested by experiment[26]. This model used in conjunction with an empirical predictor allows for the testing of millions of sequences to determine the predicted shape of phase diagrams for sequences enriched in different amino acid types. Finally, I consider the intermolecular interactions occurring between different pairs of amino acids, and highlight the presence of diverse interaction modes in two particular IDP sequences. Quantitative information on the interactions occurring within a condensed phase of protein is unprecedented, and would provide the highly-detailed information required to understand the sequence, and composition determinants of IDP phase separation.

The results presented in this thesis provide the community with a coarse-grained model which can be used to observe the sequence-dependent phase separation of IDPs, and may be used to identify important factors that promote phase separation. Since simulations are conducted using molecular dynamics simulations, the framework may be used not only to determine equilibrium thermodynamic properties of a phase-separated assembly, but also the dynamic properties of the components composing a condensate. This framework has the potential to be extended to account for effects of salt, pH, and other factors that perturb LLPS of proteins. Using the insights from all-atom simulations one may determine the interactions that drive association of different proteins. All of these observations may be applied to multicomponent systems, which is highly relevant to the field, considering that most MLOs are composed of dozens to hundreds of different components. Finally, the coarse-grained simulations may be used to generate starting configurations for atomic-resolution simulations of condensed proteinaceous phases, which may then be conducted and analyzed using the same analysis procedure as presented in the all-atom chapters of this thesis.

Chapter 2

Lipid Membranes' Response to Interactions with IDPs[1]

2.1 Introduction

Aggregation of proteins is commonly resultant in a variety of degenerative disorders including Alzheimer's disease, Amyotrophic Lateral Sclerosis (ALS), and Huntington's disease[159, 160]. For many of these aggregates, the proteins involved are naturally unfolded and exist as a heterogeneous ensemble of disordered or partially disordered conformations at physiological conditions. Such proteins do not require a definite folded structure to be functional and for many, a lack of structure is advantageous toward its physiological function[161, 162]. These intrinsically disordered proteins (IDPs) are capable of carrying out many important biological functions including cell signaling, DNA recognition[60], and formation of membraneless organelles[5]. The heterogeneous nature of IDPs allows for interactions with a larger range of other molecules as they are able to adopt structures complementary to multiple binding partners[60, 163]. Where most experimental techniques may only provide averaged information on the conformational ensemble, molecular dynamics simulation allows for the detailed characterization of the full ensemble, thus making it a popular technique to study IDPs[55, 103, 164–173].

Many IDPs are also amyloidogenic, proteins which aggregate and take the form of fibrils

[1] reproduced from ref.[158]

composed of cross-linked β-sheet structures[174]. In type 2 diabetes, accumulations of amyloid fibrils are usually found in the pancreatic islets, and are primarily composed of the 37-residue protein, human Islet Amyloid Polypeptide (hIAPP or amylin)[175]. While the toxic activity of amylin is still not well understood, evidence points to disruptions and permeabilization of the cell membranes of insulin-producing pancreatic β-cells as being the primary toxic behavior[176–178]. hIAPP has also been shown to form oligomeric aggregates prior to the nucleation and growth of fibril structures, which have implications both as growth sites or precursors for aggregates, and as the cause of membrane disruption and toxic behavior[179, 180]. The region from residue 20-29 has been most commonly proposed as the amyloidogenic region responsible for promoting nucleation and aggregation, as this is where the majority of differences between amyloidogenic human amylin (hIAPP) and non-amyloidogenic rat amylin (rIAPP) occur[181–183].

In bulk solution, hIAPP predominantly adopts disordered coil-like conformations, but will transiently populate extended α-helical conformations[183–185]. The presence of lipids results in increased helical content detected in the peptides, and acceleration of fibril formation[185–189]. It is likely the lipid interactions aid in the formation of pre-fibrillar oligomers which can then facilitate the formation of amyloid fibrils[190]. Characterization of the mechanism for such a helix-to-β-sheet transition has yet to be attained. It has also been suggested that the peptide is able to insert below the phospholipid head groups of lipid membranes,[191] resulting in the stabilization of amphiphilic helices by allowing the sidechains access to both hydrophobic tails and hydrophilic head groups of the lipids[192]. NMR experiments using sodium dodecyl sulfate (SDS) micelles and simulation studies provide additional support for this finding[193–195]. It has also been shown that insertion propensity is not significantly influenced by the type of head groups, but more predominantly by the overall content of charged lipids[191]. The nature of membrane disruption continues to be deliberated with proposed mechanisms including pore formation [196] and lipid extraction [197].

The increased rate of hIAPP aggregation in the presence of membranes is further influenced by the fraction of negatively charged lipids headgroups composing the membrane[186, 191, 198]. Eukaryotic β-cell membranes are generally negatively charged, having anionic lipid fractions within the range of 1-10%. With prolonged exposure to high concentration of glucose, common to people with diabetes, anionic lipids fractions in the cell membrane can be as high as 30%[199]. To date, the majority of studies done on the effects of membrane composition on hIAPP adsorption have been done using experimental techniques,[191, 200–202] while there has been little computational effort until recently[203, 204].

The purpose of this study is to obtain well-sampled equilibrium ensembles of hIAPP monomers in the presence of lipid bilayers of varying anionic composition, thus providing atomistic details contributing to the understanding of membrane-mediated amyloid formation. We focus on the role of lipid headgroup charges, and other contributing factors such as the ability of different lipid types to rearrange within the bilayer. The equilibrium ensembles were found to capture both inserted (strongly adsorbed) and uninserted (weakly adsorbed) states, showing a high propensity for interactions with the membrane. We observe different effects on the conformational ensemble and stabilization of helical structures for each of the bilayers tested. In the case of a mixed anionic-zwitterionic membrane, the lipids displayed the ability to locally demix, and have each lipid type cluster around specific residues of the peptide. Resultant ensemble structures of amylin also show significant similarity with experimentally determined structures in commonly sampled α-helical regions.

2.2 Simulation and Analysis Methods

2.2.1 System Properties

All simulations use full length (37-residue) human Islet Amyloid Polypeptide (hIAPP) with amidated C-terminus and disulfide bond between Cys2 and Cys7. [205, 206] The N-terminus and side chains for arginine and lysine are protonated, and histidine is protonated at Nϵ, but not Nδ resulting in a neutral charge for histidine and protein net charge of +3. Lipid bilayers were constructed using varying fractions of Dioleoyl-phosphatidylcholine (DOPC) and Dioleoyl-phosphatidylserine (DOPS). The headgroups PC and PS were selected as they are the most common zwitterionic and anionic head groups in pancreatic islet cells[207]. These headgroups also are commonly used in the literature, usually as POPC/POPS or DOPC/DOPS. We use the latter as the SLipids topology for POPS was not readily available.

Simulations were conducted in the presence of three different membranes: pure DOPC, pure DOPS, and a mixed membrane with a 7:3 ratio of DOPC to DOPS. Equilibrated models of pure DOPC and DOPS membranes were downloaded from the Stockholm Lipids website [208–210], and the mixed DOPC/DOPS membrane was constructed using CHARMM-GUI's bilayer builder[211]. Each bilayer contains two leaflets of 64 lipid molecules for a total of 128 lipids per system over an area of roughly 50 nm^2, allowed to fluctuate via pressure coupling.

2.2.2 Simulation Protocols

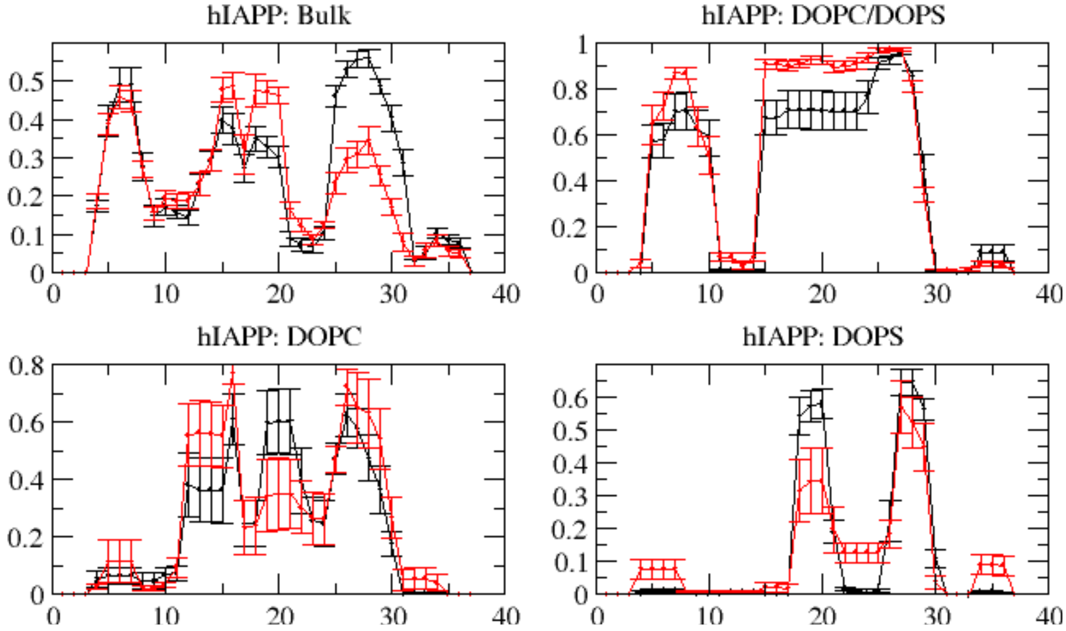

Figure 2.1: Comparison between first half (black) and second half of simulation (red).

Molecular dynamics simulations were conducted using GROMACS 4.6.7 [212] with Amber03w force field [213] for protein and TIP4P/2005 water model, [214] which have been found to work well for protein folding and disordered proteins, [167, 213, 215] and SLipids force field [208–210] was used for lipids. Validity of the SLipids force field in combination with TIP4P/2005 water model is discussed in Appendix B for POPC bilayer. A single REMD simulation of amylin in bulk solution was conducted in the same way as described in a previous publication [183] to allow for comparison with amylin's bulk ensemble. For simulations containing bilayers, semiisotropic pressure coupling was used, allowing the bilayer-normal axis allowed to fluctuate with respect to the other two axes which were coupled using Parinello-Rahman barostat [216]. Particle mesh Ewald method [217] was used for electrostatic interactions. To improve sampling efficiency, parallel tempering was used in conjunction with well-tempered ensemble (PT-WTE)[218–220]. Replicas were kept at constant temperature using Langevin dynamics with a friction coefficient of 1.0 picoseconds[221].

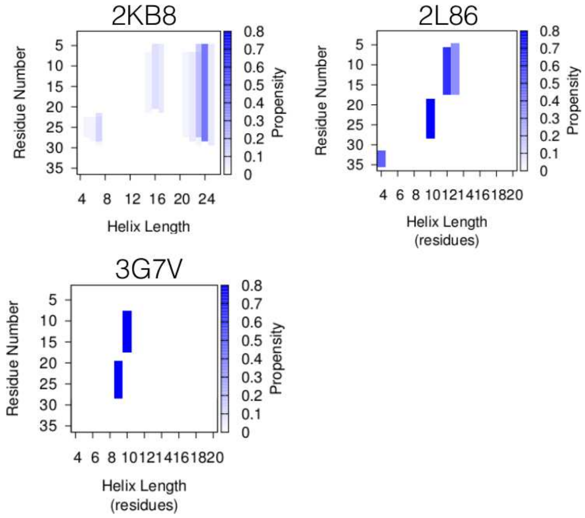

Figure 2.2: Secondary structure maps done on pdb files of 2 NMR structures (top)[193, 194] and one X-ray crystallography structure (bottom)[222] of hIAPP.

Table 2.1: Alignment of helical structures from available pdb structures of hIAPP. RMSD is calculated from residues 8-17 (red) and 20-28 (blue).

RMSD (nm)	2KB8	2L86	3G7V
2KB8	—	0.234	0.201
2L86	0.185	—	0.166
3G7V	0.181	0.134	—

To test for convergence, the secondary structure ensemble average of two halves of the simulation were compared and show no quantitative difference (Fig. 2.1). We then considered the agreement of simulation ensembles to experimental data using structural information from the protein data bank. We calculated secondary structure maps for three solved structures of hIAPP, two being adsorbed to SDS miceles (PDB IDs: 2KB8[193] and 2L86[194]) and one fused to maltose-binding protein (PDB ID: 3G7V[222]) and find that all three structures share two helical regions in common (Fig. 2.2). Alignment of the experimental structures at these two regions (res 8-17 and res 20-28) results in relatively low RMSD values between each pairing (Table 2.1) indicating these helices may also be present in bulk solution, or when interacting with lipid membranes. The simulation ensembles were found to show sampling of helical structures similar to these (Fig. 2.3).

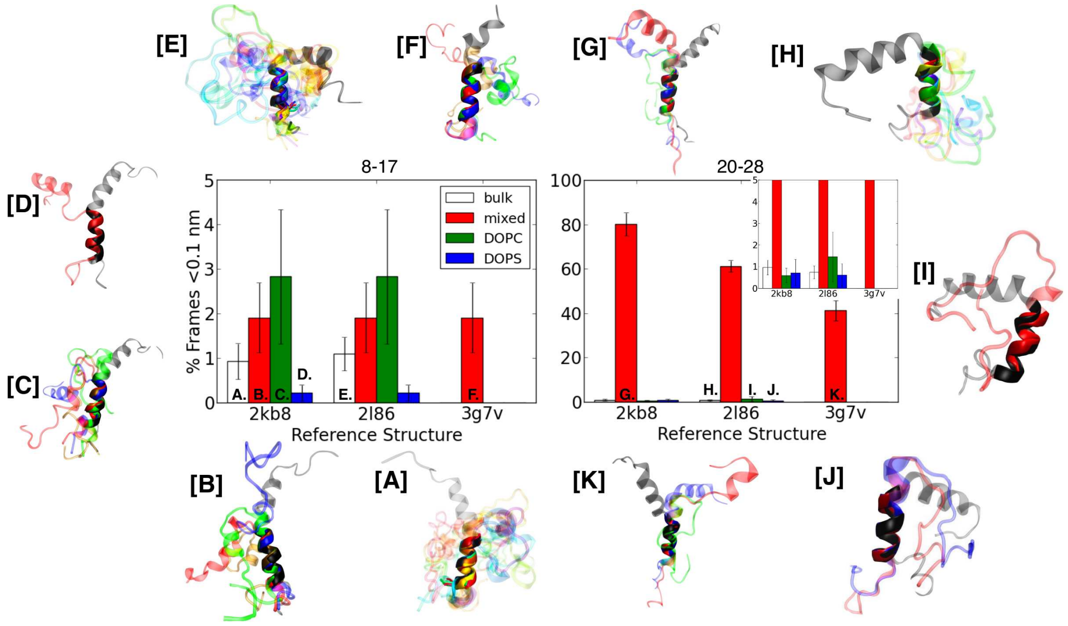

Figure 2.3: Tabulation of low-RMSD conformations within each of the equilibrium ensembles. Snapshots of select clusters are included. The insert at the top right shows a zoomed view of the comparison with the 20-28 region of the peptide. Surrounding snapshots show the different clustered conformations contributing to the overall agreement percentage. Different contributing conformations are colored according to the VMD color scheme.

Table 2.2 shows a list of the simulation systems with total simulation time, and system size used in this study.

Table 2.2: List of simulations used in this work, number of atoms and simulation time before discarding equilibration time.

Protein	Environment	# Atoms	Time (ns)
hIAPP	Bulk	13190	300
hIAPP	DOPC	50012	155
hIAPP	DOPS	42948	157
hIAPP	7:3 DOPC/DOPS	62049	163

2.2.3 Analysis

Secondary structure calculations, density calculations, center of mass distances, contact analysis and radial distribution functions were all done using GROMACS analysis modules[212]. For secondary structure definition, the DSSP algorithm[223] was used. Secondary structure maps (ss-maps) are calculated following the algorithm by Iglesias et al. [224] using the DSSP definition of the secondary structures. Visualization and RMSD alignment were done using VMD 1.9.1[225]. All results presented in this work are calculated from the replica set to 300K. Error bars for ensemble averages were obtained using block averages with five blocks.

2.3 Results

2.3.1 hIAPP Monomer Adsorbs Below Membrane Headgroups

To compare between bulk solution, and membrane-adsorbed ensembles, we first must ensure that the simulations are capturing membrane-bound states. We initially look at the simulation in the presence of the zwitterionic DOPC bilayer as its net neutral charge may be unfavorable for the positively charged amylin peptide. Fig. 2.4A shows the density profile of protein along the membrane-normal axis, from which we can see that amylin is interacting with the membrane with a strong preference for the region below the choline headgroups. Bulk solution is the least populated region in this simulation, to the extent that there is not sufficient sampling to make comparisons between bound and unbound ensembles using only data from the near-membrane simulations. In order to make comparisons between the near-membrane and bulk ensembles of amylin, we use a separate simulation conducted in bulk solution. Since the majority of the near-membrane ensemble is comprised of membrane-adsorbed states, the results have been interpreted as only representing amylin's strongly adsorbed ensemble, where much of the peptide has inserted below the lipid headgroups, and weakly adsorbed ensemble, where the peptide is primarily located at or above the surface of the membrane.

The peak of the density curve lies roughly 1.5 nanometers from the bilayer center, well below the surface, and at a depth that allows for interactions with both the hydrophilic head groups and the hydrophobic tails of the lipids. The density histogram is scaled to the average z-dimension of the simulation box, which would correlate with the thickness of the bilayer. The profile drops to zero before it reaches the bilayer center, which is in agreement with studies showing that amylin inserts into the membrane, but does not span the full width of the membrane.[191, 200]

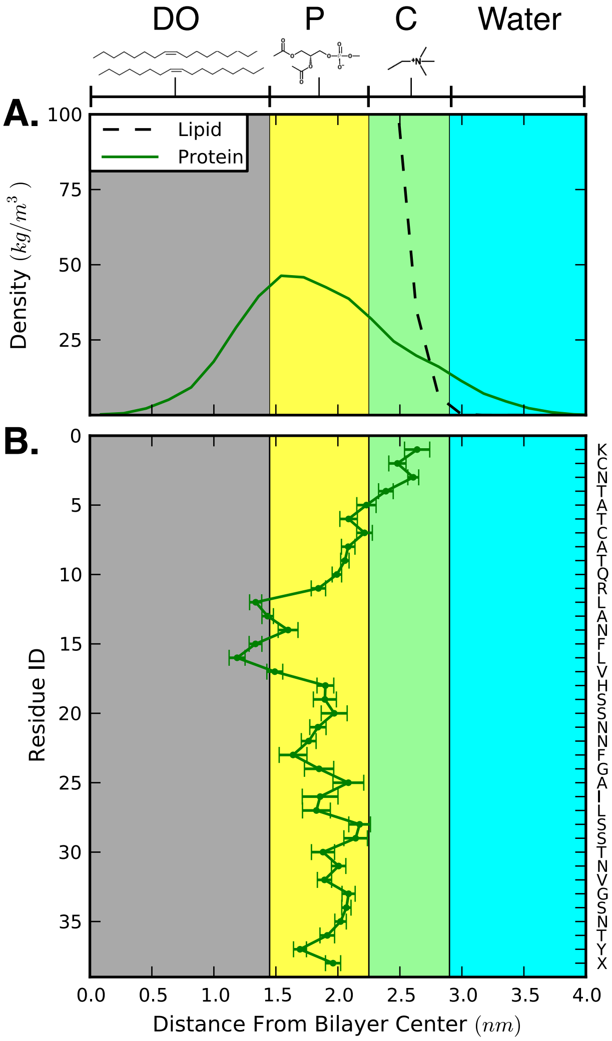

Figure 2.4: The preference of amylin to associate with different regions of the bilayer. A) Average ensemble density of amylin and DOPC lipid. B) Per-residue average distance from bilayer center. Residue X represents the C-terminal capping NH_2 group. Different environments represented by different colors: Lipid tail groups (grey), Ester/Phosphate groups (yellow), Choline headgroups (green), bulk solution (blue). Region cutoffs were determined using the average coordinates of atoms in each group.

Since amylin contains a mix of hydrophobic, polar and charged residues distributed throughout its length, we find that the interface, having access to both hydrophilic (by polar groups of membrane or water) and hydrophobic (by hydrophobic tail group of membrane) groups, is a favorable environment over bulk solution.

Further details of the protein-lipid interactions were obtained by expanding to a per-residue view of the average depth within the membrane as shown in Fig. 2.4B. The deepest insertion occurs at hydrophobic residues L12 and F15-V17, while the N-terminal, which contains a charged lysine residue followed by several polar residues is the least deeply inserted region of the peptide. This is consistent with the most recent NMR study by Nanga et al.[194] of full-length amylin in the presence of SDS micelles where the determined structure was shown to have much of the C-terminal embedded into the micelle surface with the N-terminal partially exposed to solution. They also show that much of the embedded C-terminal end, residue 21-28, is helical and parallel to the surface of the micelle.

2.3.2 Helix Stabilization With Membrane Adsorption

It is expected that the adsorption to the bilayer may have a significant effect on amylin's conformational ensemble, and to quantify this difference, we calculate the secondary structure propensity for the peptide. The DSSP algorithm was utilized to assign secondary structure to each residue for each frame of the bulk and near-membrane simulations. In Fig. 2.5A, the ensemble average of helical propensity (combined α-helix, 3(10)-helix, and 5(10)-helix), β-sheet propensity, and disordered propensity (combined β-bridge, coil, bend and turn) are shown, and an expanded view of secondary structure propensity for each amino acid of the sequence is shown in Fig. 2.5B. The overall helical content is unchanged by the presence of the bilayer, though the per-residue helicity does appear to be affected. The N-terminal region, residue 1-10, experiences a decrease in helicity and is also the region where the peptide is least strongly bound to the membrane. The already low β-sheet propensity present in bulk solution reduces to near zero with the addition of the bilayer. Miller et al. [183] find a correlation between amyloidogenicity of sequence and helical propensity in residues 22-31, a region where we also see a small increase in helicity near-membrane.

The difference between the two ensembles is further investigated by looking at the individual helices contributing to the average helix propensity, shown using secondary structure maps in Fig. 2.5C. To create secondary structure maps (ss-maps) for helicity, the per-residue helical propensity is expanded to show each independent helical segment of the sequence that is contributing to the overall helical content. This is done by showing not only the length of independent helices, but also which residues are involved, and the fraction of the ensemble that the helical structure occupies. Summing along the x-axis of the ss-map plot will yield the residue-by-residue helical content as in

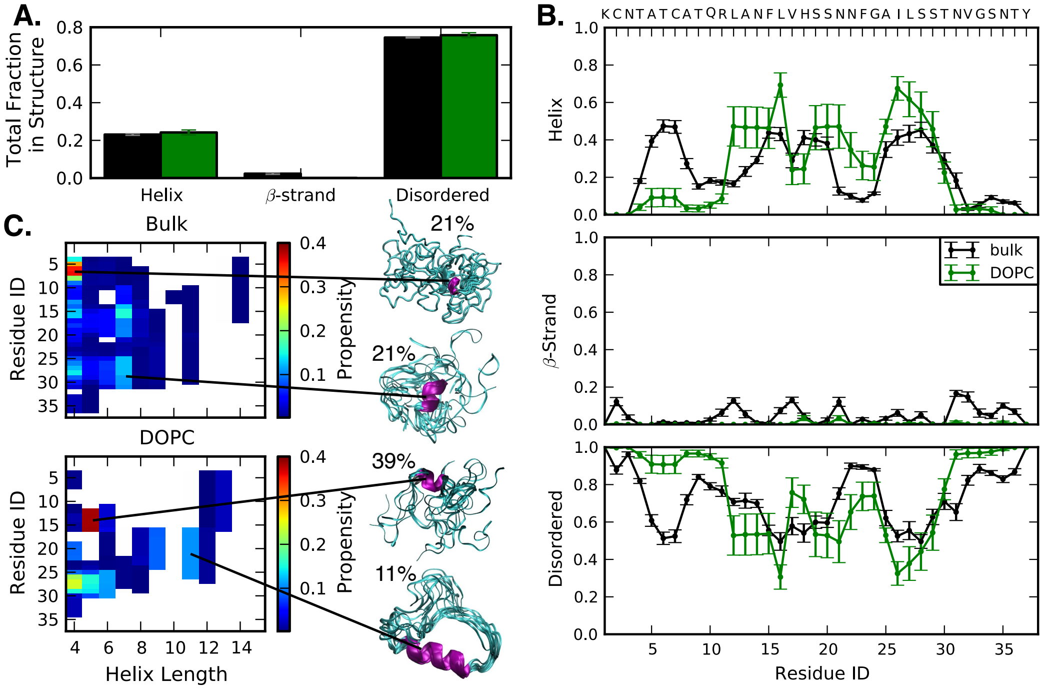

Figure 2.5: Secondary structure of amylin ensemble in bulk and near-membrane. A) Average fraction of residues occupying each type of secondary structure. B) Per-residue secondary structure. C) Secondary structure maps and clusters corresponding to certain individual helices. Regions where helical propensity is zero are set to white for clarity.

Fig. 2.5B. Since the β-strand content is comparably negligible, ss-maps are presented for helical content only.

In the bulk ensemble, the majority of helical content is composed of short helices that occur transiently throughout the length of the peptide, with longer helical conformations only being sampled rarely. In the near-membrane ensemble, the number of independent helices sampled is less than that in bulk, and there are specific regions of the peptide where helical content is enhanced, while many helical conformations observed in the bulk ensemble are not sampled at all. Of particular interest is the 11-residue helical segment spanning residue 16-26 which overlaps with the previously identified region having been linked to amyloidogenicity[183], and is the most prevalent extended helical conformation at about 11% of the population.

2.3.3 Effect of Membrane Composition on Adsorption and Structure

While the presence of the membrane has a significant effect on amylin's conformational ensemble through stabilization of specific helical structures, the lipid headgroup composition can vary widely. Two new membranes, both containing anionic DOPS lipids are used to probe if and how the lipid charge density can influence hIAPP adsorption and secondary structures. For the first membrane, a 7:3 ratio of zwitterionic to anionic DOPS lipids was used as it has been used in other studies,[191, 195] and as it is roughly the same charge fraction as biological membranes having been exposed to elevated levels of glucose[207]. And for the second, we test the extreme case with a bilayer composed purely of anionic DOPS.

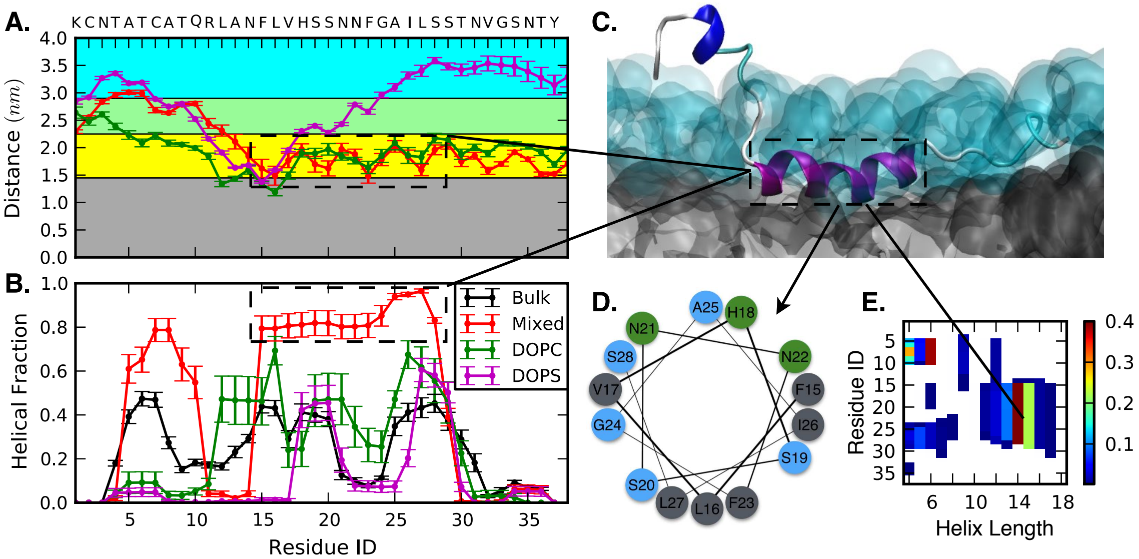

Figure 2.6: Membrane adsorption and helical content of amylin for the different near-membrane cases. A) Average per-residue coordinates showing insertion propensity with each membrane. B) Per-residue secondary structure of amylin in bulk and in the three membrane cases. C) Snapshot of high probability conformation and its position relative to the membrane in the mixed DOPC/DOPS case. Bilayer shown in surface representation with tail group region in grey, and headgroup region in blue. D) Helix wheel to show relative positions of helix residues with types: hydrophobic (grey), intermediate (blue), hydrophilic (green). E) Secondary structure maps for the mixed membrane case showing highly enhanced extended helices with little difference in the length of independent helical structures sampled.

From Fig. 2.6A, it is apparent that for the extreme case of a fully anionic membrane, amylin is merely adsorbing to the surface rather than inserting below the surface as with the other membranes. The exception to this lack of insertion is the one hydrophobic region which was also most deeply bound in the DOPC case. The mixed membrane ensemble shows little difference in insertion propensity to the DOPC membrane in the C-terminal half, yet shows greatly enhanced helical content throughout much of that region from residues 15-28 (Fig. 2.6B). A visualization of amylin sampling an extended helical conformation and its location with respect to the bilayer is given in Fig. 2.6C. Fig. 2.6D shows the residue types and their respective orientations along the helix axis in this conformation. Secondary structure map analysis was repeated for the mixed membrane in Fig. 2.6E. Most notably, the mixed membrane shows a high propensity to form long extended helices spanning about 15 residues.

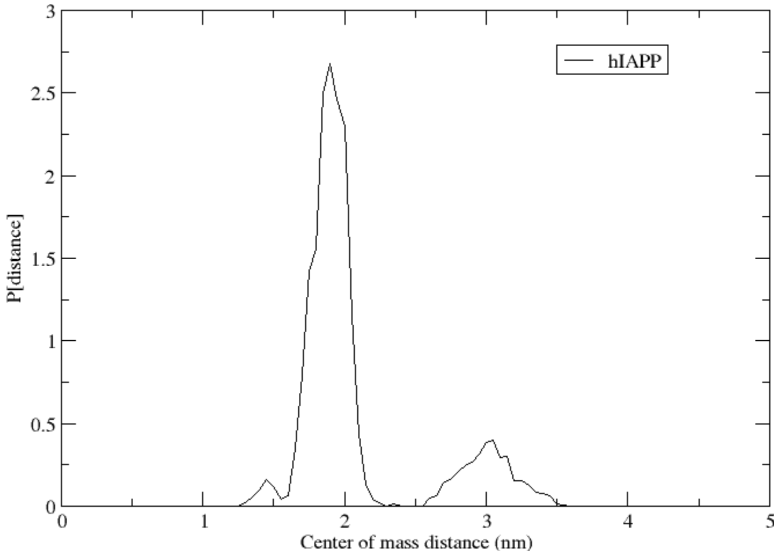

Figure 2.7: Center of mass distance distribution between protein and bilayer center for mixed DOPC/DOPS bilayer simulation.

To understand if the insertion of the peptide below the headgroups of the lipid is coupled with the helix stability, the trajectory for the mixed membrane case was split into two sub-ensembles, strongly-bound and weakly-bound based on center of mass distance between protein and membrane, and secondary structure propensity was calculated for each one separately, shown in Fig. 2.9. Using a cutoff distance of 2.5 $\mathring{A}$ is effective at separating conformations where a significant fraction of the peptide is adsorbed below the headgroups, and those where there is little to no insertion below the headgroups (Fig. 2.7 and 2.8). From the split trajectory, it is clear that amylin is highly prone to sample extended helical conformations when strongly adsorbed. The weakly-bound ensemble only shows one greatly enhanced helical region which is also present,

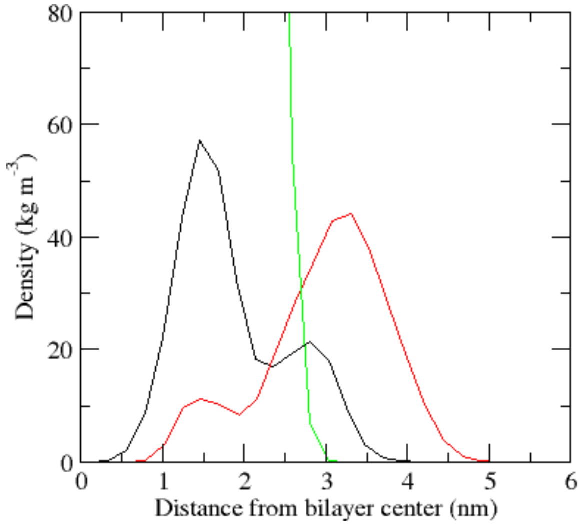

Figure 2.8: Weakly and Strongly-bound sub ensembles of IAPP near the mixed DOPC/DOPS bilayer. Strongly-bound (black) and weakly-bound (red) are both well separated, and the lipid density (green) falls to zero between the two peaks.

though to a lesser extent, in the bulk ensemble.

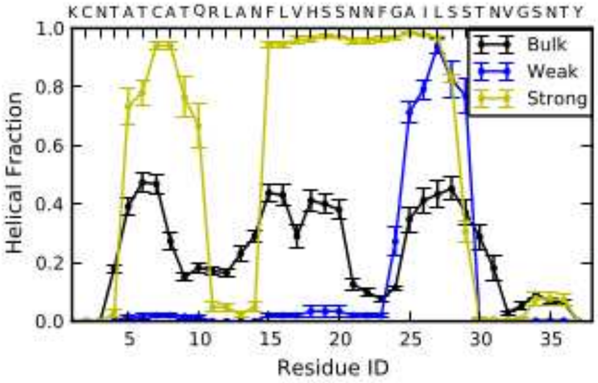

Figure 2.9: Per-residue helical fraction in sub-ensembles of mixed membrane, compared with bulk ensemble.

The addition of 30% DOPS lipids to the membrane results in very high stabilization of membrane-bound helical structures, but when the DOPS composition is increased to 100%, the helical stabilization decreases even lower than with the pure DOPC membrane as shown in Fig. 2.6B. The initial increase in stabilization could potentially be attributed to the favorable interactions between the positively charged peptide chain and the net-negative headgroup region. It is possible that the decrease in helical content for the DOPS membrane could be due to the high net charge disfavoring the presence of hydrophobic sidechains within the membrane or disruptions to the backbone hydrogen bonding needed for the formation of helical structures. Another interesting observation is that the average distance profile for the highly helical region, (Fig. 2.6A) looks very similar between the mixed and DOPC cases while the helical propensity greatly differs. A possible explanation for this is that the mixed lipid membrane is contributing to helix stabilization by allowing the DOPC and DOPS lipids to locally demix and optimize interactions with different amino acids.

2.3.4 Lipids Demix Locally to Allow Stabilization of Helical Conformations

To test whether lipid demixing is occurring, we take a look at the contacts being formed between the bound residues and the different types of lipids in the strongly-bound mixed membrane sub-ensemble. Contacts have been defined as two atoms with a distance less than 0.6 nanometers, and the contact fraction is calculated as the percentage of conformations in the ensemble where there is at least one contact between the two groups. The contact propensity for each residue is given in Fig. 2.10 and shows great similarity between the two types of lipids throughout the sequence except for one small region near the center, residues 17-22, where there is a significant drop in contacts with DOPS lipids. Using a shorter cutoff of 0.25 nanometers also shows similar reduction of contacts with PS lipids in the center of the sequence. The amino acids within this region are mostly polar, meaning that the reduction of contact propensity is likely not related to unfavorable interactions between hydrophobic amino acids and charged lipids.

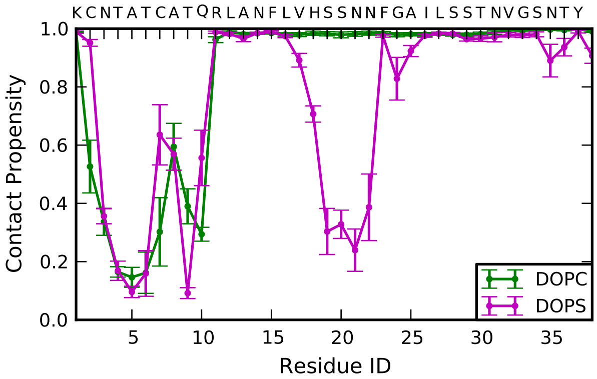

Figure 2.10: Contact propensity for amylin in its strongly-bound state with the two types of lipids composing the bilayer.

Additionally, the radial distribution function for each residue with the two lipid types was

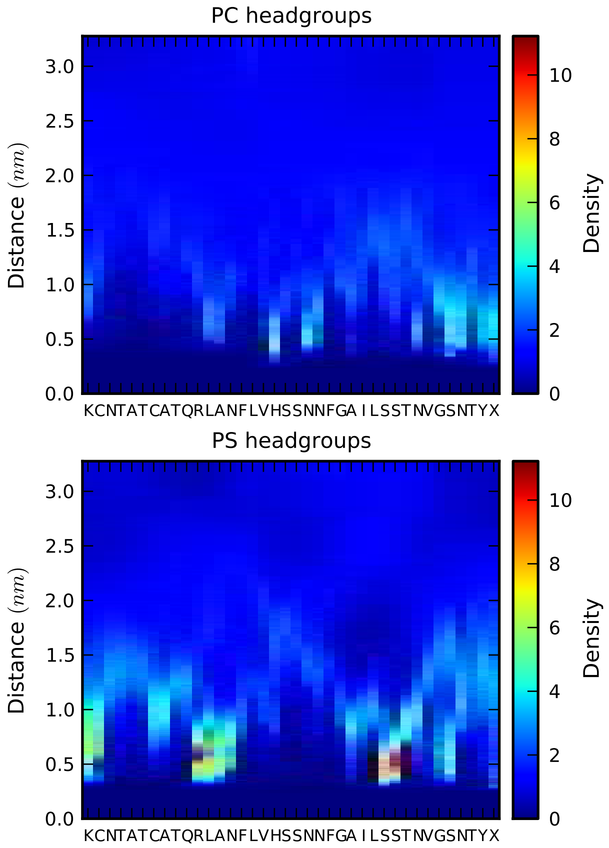

Figure 2.11: The probability of each lipid type being within a certain distance of each residue of the peptide sequence. Height of the histogram is represented by the color, and generally converges to 1 for each residue. Residue type X represents C-terminal-capping NH_2 group.

calculated and shown in Fig. 2.11. The propensity of DOPC headgroups to associate with H18 is apparent, and is similarly high for several residues throughout the helix segment. There also appears to be a reduction of DOPS density near the region where the bound helix primarily occupies, and increased density at S28 where the extended membrane-bound helical conformation ends. There is also greatly increased DOPS density at R11 near the membrane-water interface. To further visualize the demixing activity, the trajectory was centered on a single residue, H18, and density maps were calculated for PC and PS headgroups for the membrane, showing clustering of PC headgroups around the center (Fig. 2.12). The delocalization and clustering of specific lipid headgroups around specific membrane-adsorbed residues appears to be important for the stabilization of the bound-helical structure. Such ability to rearrange components is not present with the homogeneous DOPC bilayer, which could explain the lower helical content despite the similar adsorption.

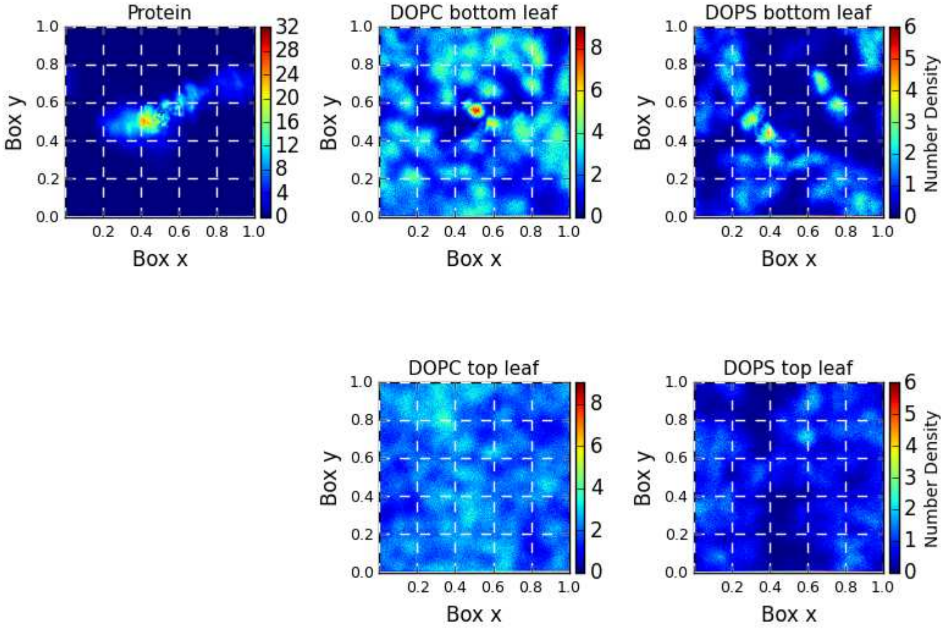

Figure 2.12: Density maps of DOPC and DOPS lipids on each leaflet with the trajectory centered at H18 of the peptide

2.4 Conclusions

We have conducted atomistic simulations of amylin-near-membrane and generated well-sampled equilibrium ensembles for three different membrane compositions. The conformations sampled by the peptide resemble those determined by experiment in a range of different conditions (Fig. 2.3). From these ensembles we determine that the presence of a lipid bilayer has an impact on the α-helical conformations sampled by the peptide. The induced changes range from reduction of overall helicity with pure DOPS, to rearrangement of helical propensity with pure DOPC, to enhancement of extended helical conformations with 7:3 DOPC/DOPS mixture. Despite the large difference in helicity, the average depth profiles of amylin for the pure DOPC and mixed membrane cases look quite similar, especially in the 15-28 region, implying that adsorption alone is not sufficient for imposing helical stability. Considering the localization of each lipid type around specific residues in the mixed membrane ensemble, it is likely that each residue of the peptide is selectively forming contacts with specific head groups, resulting in conditions more favorable for both insertion and helical stabilization (Fig. 2.13).

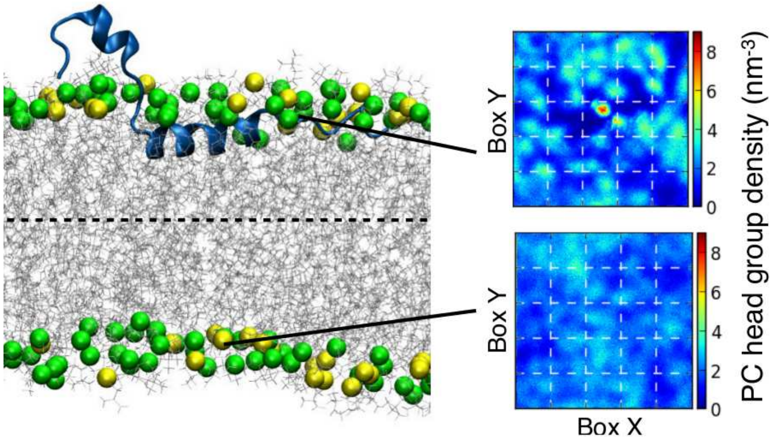

Figure 2.13: Visualization of a chain of hIAPP adsorbed to a bilayer of 7:3 DOPC/DOPS and representative lipid density maps for the side which is interacting with the protein vs. the side that is not. This shows that interactions of IDPs with lipid bilayers may induce demixing of lipids such that more preferable interactions may occur between lipid headgroups and proteins.

Specifically, the PS headgroups localizing around the end of the helix at residue 28 may be resulting in stabilization of the helix through interactions with the free backbone carbonyl groups. Considering the relative isolation of the rest of the helical structure from PS headgroups, it's possible the DOPS bilayer is not allowing insertion at all because the helical structure is disrupted by the highly charged environment, but is necessary for strong adsorption. The localization of PS around R11 is likely due to the positively charged sidechain, and may be allowing for easier passage across the water/membrane surface.

Future work on this system will be directed toward connecting the membrane-binding effects on amylin's conformational ensemble with the accelerated aggregation seen in experiment. To accomplish this, simulations of multiple amylin monomers must be conducted in the presence of lipid bilayers, which could be significantly more computationally demanding. Interactions between membrane-bound monomers may also, give insight into mechanisms of membrane disruption involved in β-cell death.

Chapter 3

Developing a Coarse-Grained Model to Simulate IDP Phase Separation[1]

3.1 Introduction

Intracellular compartmentalization is essential for normal physiological activity. This is commonly accomplished through isolation by lipid membranes or vesicles, but can also be achieved without the use of a membrane via membraneless organelles[4, 45, 226]. These organelles include processing bodies[227], stress granules[29, 120, 226, 228] and germ granules [2, 87] in the cytoplasm, and nucleoli[148] and nuclear speckles[229] in the nucleus. It has recently been established that many of these membraneless organelles can be described as phase separated liquid-like droplets[2, 230]. The process of liquid-liquid phase separation (LLPS) allows these organelles to spontaneously coalesce and disperse, and is important for many biological functions, such as response to heat shock and other forms of stress[13, 29, 231], DNA repair[11, 115], regulation of gene expression[232, 233], cellular signaling[226, 234], and many other functions requiring spatial organization and biochemical regulation [148, 235–237]. LLPS has also been implicated as a precursor to the formation of hydrogels[238] and fibrillar aggregates[11, 120], suggesting possible relevance to the pathogenesis of many diseases including Amyotrophic Lateral Sclerosis (ALS) and Frontotemporal Dementia (FTD)[11, 239].

[1] reproduced from ref.[144]

Experimental studies have characterized different properties of biological LLPS, and have shown that many systems share several common characteristics. First, the formation and dissolution processes can be tuned by the cellular environment such as changes in temperature, pH and salt concentration [110], by post-translational modification such as phosphorylation [32, 234], and by mixing with other biomolecules such as proteins [123], RNA [65, 240, 241], and ATP [4, 241]. Second, the concentrated phase has liquid-like properties, including fusion, dripping, wetting [110] and ostwald ripening [240], and its viscosity is typically several orders of magnitude higher than that of water [2, 4, 110]. Third, LLPS is commonly driven or modulated by low complexity (LC) intrinsically disordered regions (IDRs) of the protein sequence [26, 29, 110], suggesting similarities to the well-characterized LLPS of polymer mixtures [242]. It should be noted that a disordered domain is not necessary for LLPS to occur[13], and indeed LLPS is known to occur for folded proteins during crystallization or purification[47]. Folded domains along with IDRs have also been shown to modulate LLPS properties [50]. Lastly, some proteins involved in LLPS process are also able to form fibril structures[11, 120], suggesting a possible connection between the liquid-like droplet and solid fibril states. However, the molecular level understanding of LLPS cannot be easily obtained by experimental methods due to difficulty of obtaining structural properties even in the concentrated phase [29], and the cumbersome process of screening mutations [80].

A number of recent theoretical and simulation studies have addressed protein phase separation. Jacobs and Frenkel used Monte Carlo simulations to study multiple-component phase separation and found that the phase boundary is very sensitive to intermolecular interactions, but less dependent on the number of components in the system [156]. Lin and Chan applied the random phase approximation to treat electrostatic interactions[83] and Flory Huggins theory for mixing entropy and other interactions. They were able to capture the sequence specificity of charged amino acids and found that the dependence of the phase boundary of the IDP Ddx4 on salt concentration can be explained by considering only electrostatic screening in their model[243]. It was also found that the monomer radius of gyration (R_g) is correlated with the corresponding critical temperature in both theoretical work [37] and by experiment [13]. This supports the hypothesis that fundamental polymer physics principles can be used to understand LLPS[93]. However, a computational framework capable of capturing the general sequence specificity including both hydrophobic and electrostatic interactions and molecular details on both intra- and inter-molecular interactions is still missing. All-atom simulation has the potential of fulfilling both tasks [244, 245] with the use of force fields suitable for intrinsically disordered proteins (IDPs) [138, 246]. Such a force field has been recently applied to study the monomer properties of TDP-43 which is known to undergo LLPS[55]. However, computational efficiency imposes limits on the use of all-atom representation for simulating LLPS directly. Even the use of coarse-grained simulations requires well-designed sampling methods to overcome the enthalpy gap between the two phases [247, 248].

In this work, we introduce a general computational framework for studying LLPS, combining a residue based potential capable of capturing the sequence specific interactions and the slab simulation method capable of achieving convergence for phase transition properties including critical temperature, and protein concentration in dilute and concentrated phases. To demonstrate the capabilities of the model, we have selected two model proteins: the LC domain of RNA-binding protein, Fused in Sarcoma (FUS), and DEAD-box helicase protein, LAF-1, both of which are able to phase separate in vitro and in vivo[11, 29, 110]. Mutations of FUS have been shown to be highly relevant to the pathogenesis of ALS[249, 250] and display the ability to alter the kinetics of both droplet formation and aggregation into fibrils [11]. In addition, both the full length and disordered domain of LAF-1 phase separate in vitro[110], allowing us to explore the impact of a large, rigid domain on the LLPS behavior.

3.2 Simulation and Analysis Methods

3.2.1 Coarse-grained Model Development

All-atom simulations are unable to reach the time scales needed to study phase separation with current state-of-the-art computational hardware resources and sampling methods. We therefore introduce a coarse-grained representation of the protein, in which each residue is represented as a single particle (Fig. 3.1A). The model takes into account the chemical properties of the 20 naturally occurring amino acids, listed in Table 3.1, thus making it sequence specific. The potential energy function contains bonded, electrostatic, and short-range pairwise interaction terms. Bonded interactions are modelled by a harmonic potential with a spring constant of 10 kJ/Å^2 and a bond length of 3.8 Å. Electrostatic interactions are modeled using a Coulombic term with Debye-Hückel[111] electrostatic screening to account for salt concentration, having the functional form:

$$E_{ij}(r) = \frac{q_i q_j}{4\pi D r} \exp(-r/\kappa), \tag{3.1}$$

in which κ is the Debye screening length and $D = 80$, the dielectric constant of the solvent medium (water). For all the simulations for which phase diagrams are generated, a Debye screening length of 1 nm, corresponding to an ionic strength of approximately 100 mM, is used. When determining R_g for IDPs from the literature, ionic strength is set to match that from the experimental results, as listed in (Table 3.2). The short-range pairwise potential accounts for both protein-protein and protein-solvent interactions. Here we have introduced two different models: the first is based on amino acid hydrophobicity [251, 252] and uses functional form introduced by Ashbaugh and Hatch [253]; the second is based on the Miyazawa-Jerningan potential [128] with the parameterized functional form taken from Kim and Hummer [151].

Table 3.1: The amino acid parameters used in the HPS model. σ is the diameter of the amino acid used in the short-ranged pair potential. λ is the scaled hydrophobicity from Kapcha et al.[252].

Type	Mass (amu)	Charge	σ (Å)	λ
ALA	71.08	0	5.04	0.730
ARG	156.20	1	6.56	0.000
ASN	114.10	0	5.68	0.432
ASP	115.10	-1	5.58	0.378
CYS	103.10	0	5.48	0.595
GLN	128.10	0	6.02	0.514
GLU	129.10	-1	5.92	0.459
GLY	57.05	0	4.50	0.649
HIS	137.10	0.5	6.08	0.514
ILE	113.20	0	6.18	0.973
LEU	113.20	0	6.18	0.973
LYS	128.20	1	6.36	0.514
MET	131.20	0	6.18	0.838
PHE	147.20	0	6.36	1.000
PRO	97.12	0	5.56	1.000
SER	87.08	0	5.18	0.595
THR	101.10	0	5.62	0.676
TRP	186.20	0	6.78	0.946
TYR	163.20	0	6.46	0.865
VAL	99.07	0	5.86	0.892

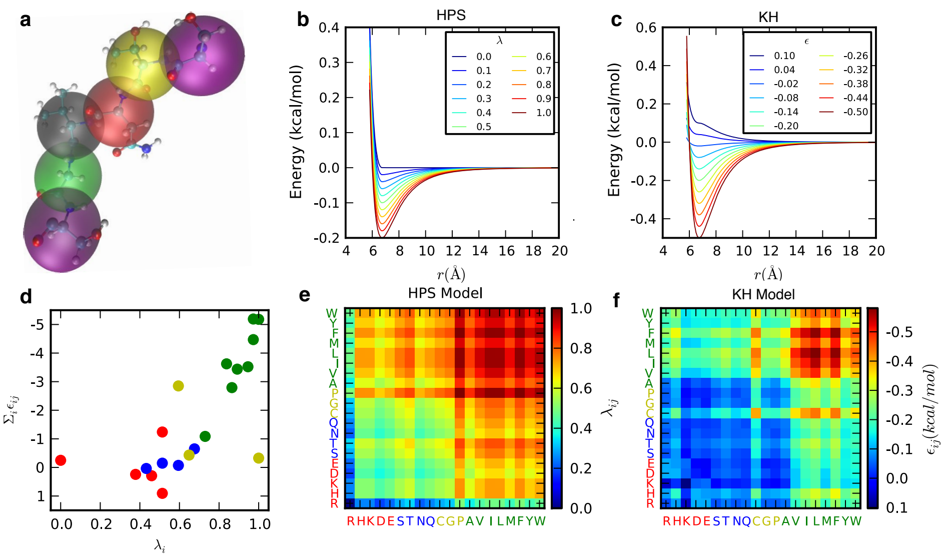

Figure 3.1: Schematic of the two knowledge-based potentials used for short-range pairwise interactions. A) Each amino acid is treated as a single particle. B, C) Potential energy functional form for HPS and KH models at different interaction strengths, plotted with a constant σ value of 6 Å. D) Correlation between the amino acid interaction strength ($\Sigma_i \epsilon_{ij}$) in KH model and hydrophobicity (λ_i) in HPS model, colored by the side-chain properties of amino acids (i.e., red for charged, blue for polar, green for hydrophobic and yellow for other amino acids). E, F) The pairwise interaction parameters used in HPS and KH models shown in color maps with blue being most repulsive interactions and red being most attractive.

Table 3.2: List of intrinsically disordered or unfolded proteins with experimentally determined R_g. The radii of gyration of ACTR and hNHE1cdt were measured at 5°C and 45°C in the experiment and have been interpolated to 25°C for comparison with the other proteins.

	Protein	Chain length	[Ion] (mM)	R_g, expt (nm)	Method	P_{charge}	Hydrophobicity
a	CspTm	67 (54)	42	1.37 (0.07)	FRET [254]	0.313	0.676
b	IN	60 (57)	50	2.25 (0.11)	FRET [254]	0.267	0.648
c	ProTα-N	112 (56)	42	2.87 (0.14)	FRET [254]	0.563	0.555
d	ProTα-C	129 (55)	42	3.70 (0.19)	FRET [254]	0.488	0.573
e	R15	114 (94)	128	1.72 (0.09)	FRET [38]	0.325	0.616
f	R17	100 (94)	128	2.29 (0.11)	FRET [38]	0.340	0.647
g	hCyp	167 (164)	85	2.00 (0.10)	FRET [38]	0.234	0.679
h	Protein-L	64 (64)	128	1.65 (0.14)	FRET [255]	0.266	0.682
i	ACTR	71	199	2.51 (0.13)	SAXS [256]	0.254	0.644
j	hNHE1cdt	131	199	3.63 (0.18)	SAXS [256]	0.298	0.671
k	sNase	136	17	2.12 (0.10)	SAXS [257]	0.331	0.659
l	α-synuclein	140	156	3.3 (0.3)	FRET [258]	0.279	0.678

Hydrophobicity scale (HPS) model.

The first model uses a hydrophobicity scale from the literature [252] to describe the effective interactions between amino acids. For use in the coarse-grained model, the atomic scale is first summed up to obtain a residue scale and is then scaled to the range from 0 to 1. The hydrophobicity values, λ, used for the 20 amino acids can be found in Table 3.1. The arithmetic average is set as the combination rule for both the pair interactions λ between two amino acids and the size σ, of the amino acids (i.e., hydrophobicity scale $\lambda_{i,j} = (\lambda_i + \lambda_j)/2$ and amino acid size $\sigma_{i,j} = (\sigma_i + \sigma_j)/2$). The combined pairwise interaction strengths for each amino acid pair are shown in Fig. 3.1E. The Ashbaugh-Hatch functional form[253] which has previously been applied to the study of disordered proteins[259], allows the attractiveness of the interactions to be scaled by λ (Fig. 3.1B), and is described by,

$$\Phi(r) = \begin{cases} \Phi_{LJ} + (1-\lambda)\epsilon, & \text{if } r \leq 2^{1/6}\sigma \\ \lambda\Phi_{LJ}, & \text{otherwise} \end{cases} \tag{3.2}$$

in which Φ_{LJ} is the standard Lennard-Jones potential

$$\Phi_{LJ} = 4\epsilon\Big[\Big(\frac{\sigma}{r}\Big)^{12} - \Big(\frac{\sigma}{r}\Big)^{6}\Big]. \tag{3.3}$$

The pair potential for the least hydrophobic amino acid at a λ value of 0 consists of only the repulsive term, making it equivalent to the Weeks-Chandler-Andersen functional form[260]. The model contains one free parameter ϵ, which determines the absolute energy scale of the short-ranged interactions and is set to be constant across all pairs. To determine the optimal ϵ, R_g was calculated for a set of IDPs (Table 3.2) using our model, and compared with available experimental R_g data. Obtaining accurate estimates of R_g from FRET and SAXS experimental data requires some care, as has recently been noted[261–264]. Since FRET probes an intramolecular pair distance, inferring R_g requires the assumption of an underlying polymer model with known pair distance distribution and related R_g. It has been shown that the commonly used Gaussian chain model works reasonably well for IDPs in the absence of chemical denaturants, but it breaks down when such denaturants are added[261, 265]. This is because the polymer scaling exponent $\nu \approx 1/2$

for IDPs without denaturants present, so that the Gaussian chain is a reasonable approximation for the denaturant-free conditions we are concerned with. We obtain the R_g using a Gaussian chain model with a dye correction of 9 residues, as previously described[38, 262]. For SAXS, Guinier analysis is challenging because the approximation is only valid for a small range of q where the data tends to be noisy; when fitting a larger range of scatter angles, it tends to underestimate the R_g[261]. A proper treatment of SAXS data requires a model that can also fit data at wider angles[39, 261–263]. Despite the limitations of the presently used data set, we expect that the systematic errors introduced by data analysis methods are still substantially smaller than the the deviation of the fit from experiment. However, a finer optimization of the model may require both the FRET and SAXS experimental data to be more accurately analyzed.

Fig. 3.2 and 3.3 show that an ϵ of 0.2 gives the greatest similarity to the experimental size of these unfolded proteins. In order to test if the model can capture the degree of collapse for folded and disordered sequences, we generated 131 sequences of 100 amino acids with properties covering a wide range of net charge and hydrophobicity values, and determined R_g from simulation. In Fig. 3.4, we present the R_g of these sequences in a Uversky type plot [19] and in a Pappu type plot [266], both of which have been widely used to characterize sequence properties of proteins.

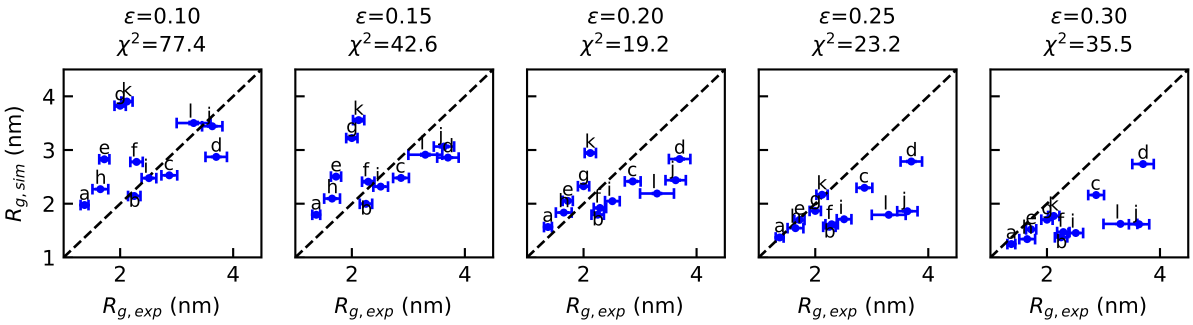

Figure 3.2: Comparison of R_g between simulations and experiments with different ϵ parameters for HPS model. The deviations χ^2 between the simulations and experiments are shown in the title. The list of the proteins and legends can be found in Table 3.2

R_g values are observed ranging from 1.5 to 6.0 nm, and the predictions are good for naturally occurring test sequences. The larger R_g values obtained for some of the synthetic sequences are outside the range observed for natural sequences in Fig. 3.3, however this is because the extreme synthetic sequences are essentially polyelectrolytes which are rare in nature. Although we do not have experimental data for such sequences, we note that the model still makes accurate predictions for the most charged protein in our data set, Prothymosin α-N (Table 3.2C), which has a net charge of -43 (-0.384 per residue), mean hydrophobicity of 0.555, and R_g of 2.87 nm. It is clear that the HPS model describes the known sequence-specific features of the disordered proteins, that is, a small mean hydrophobicity scale and a large mean net charge. The Uversky plot in Fig. 3.4 shows a correlation of R_g with both hydrophobicity and mean charge per residue as seen in experiment[38]. It does appear that the correlation is stronger with net charge, while both factors were correlated with scaling exponents in earlier work [38]. This is partly because our sampled sequences span a larger range of charge, and also because charge and hydrophobicity are correlated in naturally occurring sequences, making it harder to separate their respective contributions. Even so, the correlation with charge does appear to be better in experiment [38].

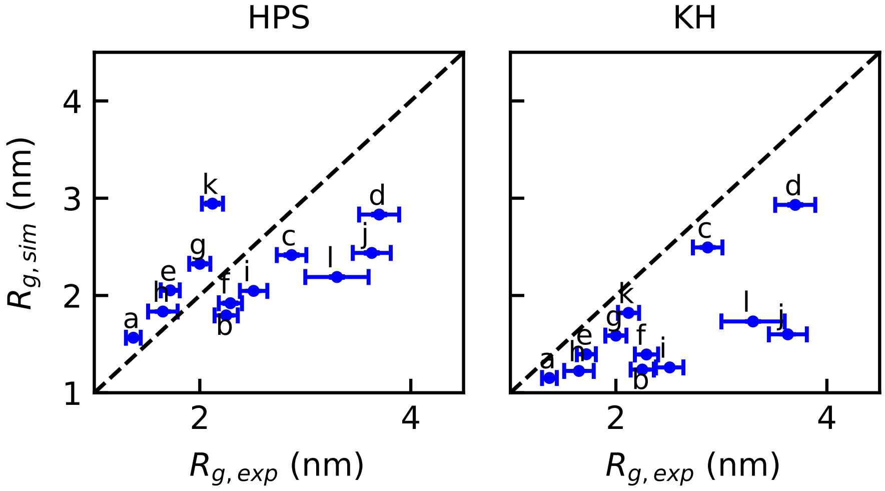

Figure 3.3: Parameterization of coarse-grained models: Comparison between radius of gyration of various intrinsically disordered proteins from experiment, and from simulation with the optimal parameters.

Kim-Hummer (KH) model.

A different model for short-range interactions has been previously developed and parameterized by Kim and Hummer to describe protein-protein interactions, using a variety of experimental

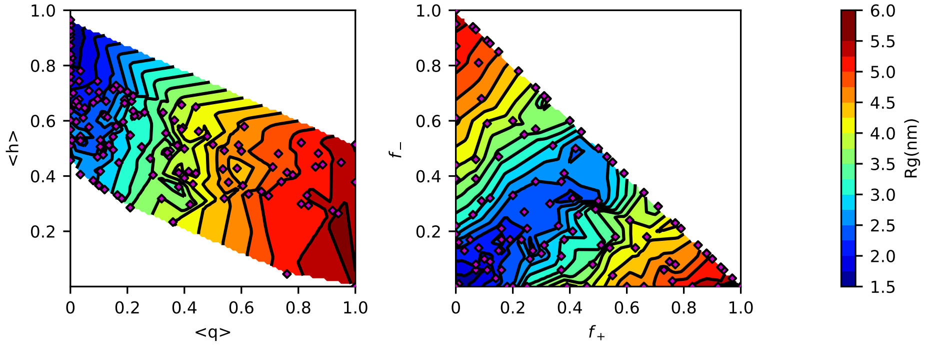

Figure 3.4: Randomly generated sequences of 100 amino acids follow the general trends expected from an Uversky type plot (left) and Pappu type plot (right). Axes are: mean hydrophobicity per residue $\langle h\rangle$, mean net charge per residue, $\langle q\rangle$ and fractions of positively f_+ and negatively f_- charged residues. For both plots, the color represents average R_g, and contour lines are spaced every 0.25 nm. The location of each tested sequence is represented by a purple diamond.

data including the osmotic second virial coefficient of lysozyme and the binding affinity of the ubiquitinCUE complex[151]. The potential function they used can be expressed in terms of Ashbaugh-Hatch potential function (Eq. 3.2 and 3.3), where

$$\epsilon = |\alpha(\epsilon_{MJ} - \epsilon_0)|, \tag{3.4}$$

and

$$\lambda = \begin{cases} 1, & \text{if } \epsilon_{MJ} \leq \epsilon_0 \\ -1, & \text{otherwise} \end{cases} \tag{3.5}$$

ϵ_{MJ} is from the Miyazawa-Jerningan statistical contact potential [128]. Regarding the choice of α and ϵ_0, the original literature identifies six sets of parameters, differing in the treatment of interactions involving buried residues. Here we employ parameter set D ($\alpha = 0.228$ and $\epsilon_0 = -1.00$ kcal/mol) for IDR, which generates a reasonable estimate of R_g for a list of IDPs (Fig. 3.3), and parameter set A ($\alpha = 0.159$ and $\epsilon_0 = -1.36$ kcal/mol) for the helicase domain, which was parameterized for interactions between folded proteins [151]. The correlation between the parameters of the HPS and KH models for IDR is shown in Fig. 3.1D. We repeat the analysis previously done with the HPS model on the same set of 100mers (Fig. 3.5) to provide additional insight into how the two models compare with regard to relative interaction strength of hydrophobic and electrostatic interactions. Both attractive and repulsive forces are stronger in the KH model than in HPS, thus there is a stronger dependence of R_g on hydrophobicity, especially for sequences with low charge.

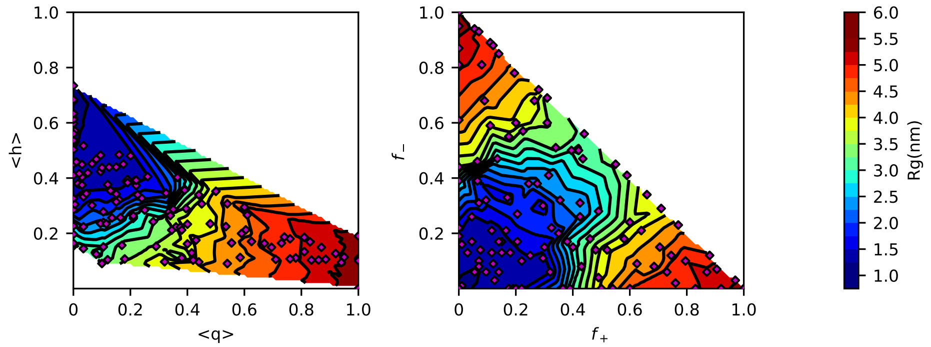

Figure 3.5: Randomly generated 100-mers in a Uversky (left) and Pappu (right) plot to show the dependence of R_g on charge and hydrophobicity using the KH model.

3.2.2 Simulation framework

Slab method.

In order to determine the phase diagram of the disordered proteins, we utilize a method [247, 267], in which the high-density (concentrated) phase, with surfaces normal to z, is simulated in equilibrium with the low-density (dilute) phase as visualized in Fig. 3.7C. This allows the determination of the equilibrium density (or concentration) of proteins in each phase and consequently, the critical temperature, as described in more detail below. This initial equilibration is conducted for 100 ns in the NPT ensemble, starting from a dispersed phase of protein chains with periodic boundary conditions at 150 K, maintained by a Langevin thermostat with a friction coefficient of 1 ps^{-1}, and 1 bar, maintained by a Parrinello-Rahman barostat[216]. A time step of 10 fs is used for all the simulations. The box size is first scaled to roughly 15 nm (25 nm for full length LAF-1) for both x and y axes and then equilibrated along the z-axis using anisotropic pressure coupling. Depending on the protein of interest and the pairwise potential, the length of the z-axis can vary. The x- and y- dimensions were set to 15 nm which is sufficient to prevent to most of the chains ($>$ 99% estimated by a random-coil model for a 170-residue chain) from interacting with its periodic image. Then the z-dimension of the box was extended to 280 nm ($\sim$ 20 times larger than the initial z-dimension box size). Simulations are then conducted at multiple temperatures for ~ 5 μs using constant temperature and volume with a Langevin thermostat. The temperature is gradually increased from 150 K to the targeted temperature over the first 100 ns. The next 1 μs of simulation is discarded as equilibration, and the remainder (at least 4 μs) is used for further analysis. Simulations were conducted using the LAMMPS[268] and HOOMD-Blue v2.1.5 [269] software packages in order to benefit from both CPU and GPU resources.

We took several measures to verify that the initial configuration, system size and number of steps are sufficient to obtain well-converged thermodynamic properties of the system. First, we find that a simulation starting from a fully dispersed configuration, in which chains are put far from each other, but having the same periodic box geometry, will eventually coalesce to form a concentrated phase and generate a similar density profile (after 4μs) to a simulation starting from a slab-like initial configuration (Fig. 3.6). Therefore a slab-like initial configuration reduces the length of the simulation required for convergence. Second, we do not see a quantitative difference of the results between the two halves of a 10μs simulation, suggesting 5 μs is sufficient for convergence of the system. Third, we have also found that a system with 100 chains of length ~ 160 is sufficiently large to avoid finite-size effects, as the results are identical to those from a similar set of simulations containing 200 chains.

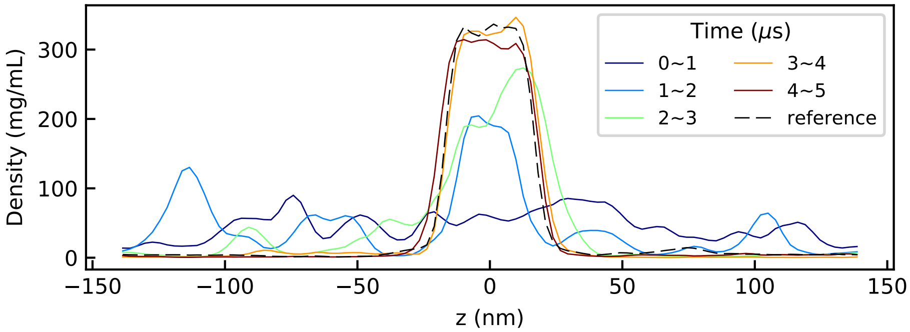

Figure 3.6: LAF-1 simulation started from dispersed state at 210K with KH-D model showing coalescence to a slab conformation after about 4 μs. The colored lines show the density profile at different time ranges throughout the simulation. The black line shows the simulation starting from an initial slab configuration as a reference.

Slab density profile.

To determine the density profile along z, we first center the trajectory on the slab for each frame. The slab is defined as the cluster with the largest number of chains. Clustering was done according to center-of-mass-distance between chain pairs, where chains with center-of-mass distances less than 5 nm are considered to be in the same cluster except for full length LAF-1, with which we use a cutoff of 7 nm due to its larger size. The density profile along z is then generated as shown in Fig. 3.7A and 4.1A. If phase separation occurs, we obtain the protein concentration of the dilute and concentrated phases (ρ_L or ρ_H) by using the average concentrations when $|z|$ >50 nm or $|z|$ <5 nm respectively. Protein concentration is reported in units of mg/mL.

Phase diagram.

The critical temperature T_c can be obtained by fitting

$$\rho_H - \rho_L = A(T_c - T)^\beta \tag{3.6}$$

where β is the critical exponent which is set to 0.325 (universality class of 3D Ising model[270]) and A is a protein-specific fitting parameter. For fitting to this equation, a specific range of temperatures must be used. The minimum fitting temperature, T_1, is chosen as the lowest temperature where ρ_L is nonzero. The maximum fitting temperature T_2 must be below the critical temperature as Eq. 3.6 can only describe the behavior below T_c (Fig. 4.1C). To determine the optimal value for T_2 we calculate the relative error of T when fitting T as a function of $\rho_H - \rho_L$ using different test values of T_2. This error will be large if T_2 is greater than T_c (Fig. 4.1D). We can then obtain a typical phase diagram as shown in Fig. 3.7B and 4.1B, in which the ρ_L and ρ_H when $T < T_c$ are determined from averaging different regions of the slab density profile as described above and T_c and the corresponding ρ_c are from fitting Eq. 3.6 (Fig. 4.1C). Fig. 3.7C shows visualizations of the different states of coexistence captured by these simulations. When the system is above T_c, the slab evaporates to a supercritical protein solution. When the temperature is below T_c, we see coexistence of two phases: one phase with free monomers and the other with many proteins in a condensed, liquid-like assembly. The number of free monomers decreases with decreasing temperatures to concentrations comparable with protein concentration in the dilute phase observed by experiment[29, 110]. The critical temperatures for all sequences presented in this work are listed in Table 3.3.

Table 3.3: Summary of slab simulations and critical temperatures obtained.

System	Model	N_{residue}	N_{chain}	T_c (K)
FUS				
WT	KH	163	100	260.3
WT	HPS	163	100	344.4
WT	HPS	163	200	346.1
6E mutant	HPS	163	100	326.6
6Ep mutant	HPS	163	100	327.2
6Es mutant	HPS	163	100	326.4
12E mutant	HPS	163	100	280.3
LAF-1				
IDR	KH	168	100	223.6
IDR	HPS	168	100	247.2
Folded	KH	437	100	260.9
Full length	KH	708	100	253.5
Repeated fragment of FUS				
$[\text{FUS40}]_1$	HPS	40	480	309.6
$[\text{FUS40}]_2$	HPS	80	240	336.5
$[\text{FUS40}]_3$	HPS	120	160	348.5
$[\text{FUS40}]_4$	HPS	160	120	356.1
$[\text{FUS40}]_5$	HPS	200	96	363.7

We have also fit our simulated phase diagram with Flory-Huggins theory[271, 272] by using Eq. S11 of the reference[27]. There are three fitting parameters used in the original literature[27]: A and B are the temperature-independent and dependent terms in the interaction strength χ whereas ρ is the protein density. We expect that in our coarse-grained simulation, the entropic contribution to χ will be negligible. Indeed, we find that if we allow three fitting parameters, A is usually one or two orders of magnitude smaller than B/T. In order to improve the robustness of the fitting, we therefore set A to be zero and only use two free parameters, B and ρ. We list ρ and χ calculated from B of each sequence in Table 3.4.

Table 3.4: List of parameters for fitting to Flory-Huggins theory.

System	Model	Protein density (mg/mL)	$k_B T \cdot \chi$ at T=300K (kcal/mol)
LAF1 IDR	KH	1010.07	0.270
LAF1 IDR	HPS	1369.87	0.298
FUS WT	KH	1247.92	0.307
FUS WT	HPS	1547.03	0.410
FUS 6E	HPS	1553.05	0.340
FUS 6Ep	HPS	1433.90	0.391
FUS 6Es	HPS	1410.66	0.396
FUS 12E	HPS	1691.11	0.325
$[\text{FUS40}]_n$	HPS	1286.82	0.437

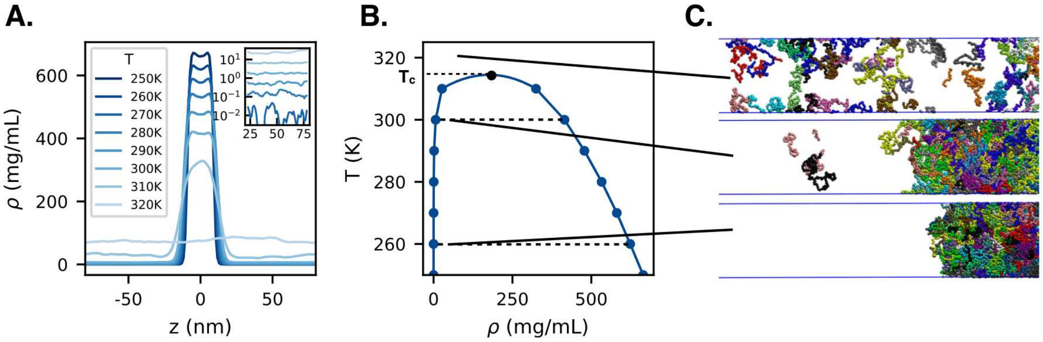

Figure 3.7: Determining phase diagram from CG simulation. A) Density profiles calculated along elongated z-axis of simulation box. Inset shows concentration in vapor phase on log scale, and that it is nonzero. B) Coexistence curve of protein plotted using concentrations from density profiles. C) snapshots of coexistence simulation at different temperatures corresponding to phase separated and dispersed single-phase systems. Figure adapted from ref. [92].

Simulations with folded domain.

Proteins which undergo LLPS usually contain multiple domains, including both folded and disordered domains[25]. Recently, Riback et al. found that poly(A)-binding protein Pab1 exhibits LLPS behavior in the absence of its disordered domain, but does not in the absence of the folded domains [13], contrary to the notion that intrinsic disorder is necessary for phase separation. Since both intrinsically disordered and folded domains can form favorable intermolecular interactions stabilizing the high density phase, it is only natural that they may both contribute to the LLPS behavior, and the contributions may be different from protein to protein. We use full length LAF-1 which contains a folded domain and two disordered domains, as a test case to see how the proposed framework will accommodate folded proteins.

The structure of the folded domain (helicase) of LAF-1 has not yet been solved, so we have used homology modelling and the Modeller v9.17 package[273] to embed the LAF-1 helicase sequence into its homologue with a solved crystal structure, VASA[274] (Fig. 3.8).

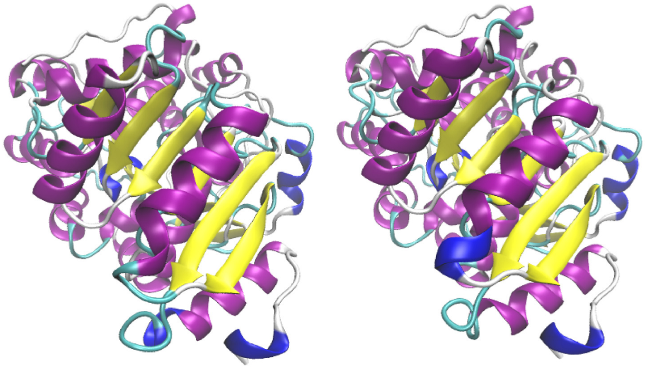

Figure 3.8: Homology modelling of helicase domain of LAF-1 using the structure of VASA. Left: the structure of VASA residue 202-621 (PDB ID: 2DB3[274]); Right: the structure of LAF-1 helicase domain residue 187-623 from homology modelling.

Here we employ the KH model with parameter set A ($\alpha = 0.159$ and $\epsilon_0 = -1.36$ kcal/mol) for all interactions involving the helicase domain, and parameter set D for disordered-disordered interactions as before. The reason for this is that a 12-6 potential allows buried residues to make a significant contribution to binding energies of folded domains, which will have a stronger effect on the affinity than the specificity of the interactions. Model A was parameterized including such interactions for folded proteins, and is therefore appropriate for use in our model in describing interactions involving folded proteins. Model D was parameterized using a screening term to reduce the effect of buried residues, and is therefore appropriate for describing interactions between disordered regions where all residues are essentially fully exposed. A universal set of parameters would require a different functional form.

When the structure of the folded domain is modelled, we treat the helicase as a rigid body (i.e., “fix rigid” command in LAMMPS or “md.constrain.rigid” command in HOOMD-Blue) in the simulation so that the structure of the folded domain is preserved. Interactions between residues within the same rigid body are neglected. The mass of the rigid body is scaled to be 0.5% of the original mass in order to accelerate rigid body dynamics. When calculating the density of the folded domain, the mass is scaled back to match the mass of the original folded domain with all residues. The folded domain can in principle also be simulated using harmonic restraints instead of rigid constraints, which would allow additional flexibility. However there is a clear advantage for using rigid body dynamics in terms of computational efficiency.

3.3 Results

3.3.1 Phase separation of FUS and its phosphomimetic mutants

As a first application of our model to LLPS, we use the prion-like LC domain of the protein FUS (FUS-LC) which is sufficient to induce LLPS in vitro in the absence of other biomolecules [11]. FUS-LC is an ideal system to test our model as it is fully disordered and displays very low secondary structure content[29]. The sequence is largely uncharged, with only 2 anionic aspartate residues within its 163 amino acid sequence. To test for sequence-specific effects, we conducted simulations for several different variants of the FUS-LC peptide, wild-type and four phosphomimetic mutants where a set of the 12 naturally phosphorylated threonine or serine residues are mutated to glutamate[275]. The first of these mutants is the 12E mutant, which contains all 12 glutamate substitutions, and does not undergo LLPS under similar conditions to FUS WT [33]. We additionally test the 6E mutant reported in the same work [33], and two designed variations of the 6E mutant, termed 6E' and 6E* which maximize and minimize, respectively, the clustering of charged residues within the sequence under the constraints of preserving the amino acid composition of 6E, and only mutating naturally occurring phosphorylation sites[82].

Utilizing the slab method, determine the range of temperatures at which the simulated FUS chains separate into two phases, and calculate the coexistence curve using both the HPS and KH models. The concentration of the dilute phase gives the predicted critical/saturation concentration of the protein, the concentration above which it will begin to form droplets in solution. The concentration of the dilute phase is on the order of 0.1-10 mg/mL over the tested temperature range, consistent with typical concentrations used to observe phase separation of FUS WT in vitro [29] (∼1-5 mg/mL). We find that the critical temperature differs between the two models for FUS WT. However the coexistence curves and the phase diagrams are qualitatively similar, as are the intermolecular contact maps (Fig. 3.10).

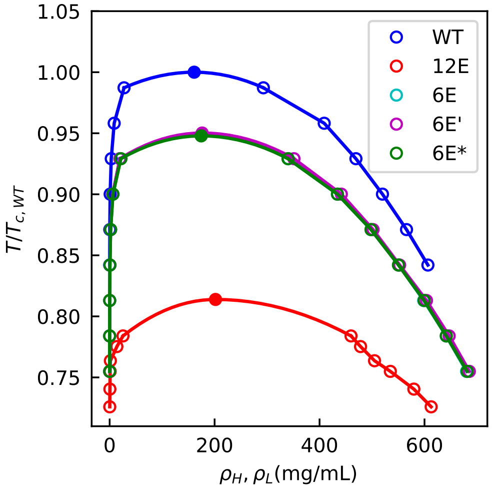

Figure 3.9: Phase diagram for FUS WT, 6E variants and 12E. Temperatures are scaled by the critical temperature of FUS WT.

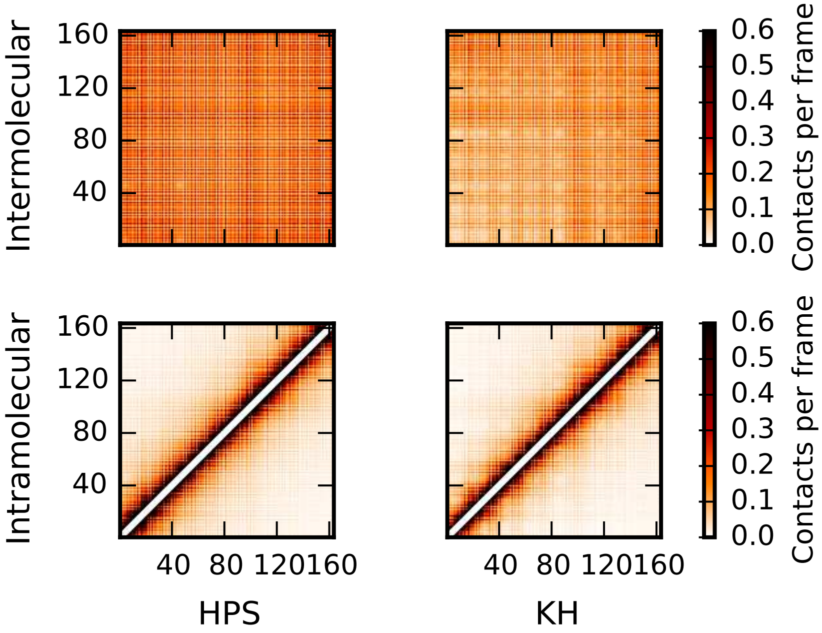

Figure 3.10: Inter- (upper) and intra-molecular (lower) contact maps for FUS WT at 260 K using HPS (left) and KH models (right).

To evaluate the impact of the phosphomimetic mutations, we determine the phase diagram for FUS WT, 6E, 6E', 6E*, and 12E using the HPS model (Fig. 3.9). The 12E mutant phase separates at a much lower temperature, with the critical temperature smaller than even the lowest temperature at which we can observe coexistence between two phases for FUS WT (due to the prohibitively small concentration of the low-density phase). This is consistent with the experimental observation that FUS 12E is unable to phase separate in contrast to FUS WT at similar conditions[33]. The 6E mutants all lie between the two extreme cases, and have nearly identical phase diagrams. While the difference of just 6 amino acids results in a greatly altered phase-separation ability from wild type, the rearrangement of these mutations does not have such an effect. However, these mutations were done under very strict constraints which do not allow for much change in the degree of charge clustering. We also calculate the inter-chain contacts, defined as two amino acids of different chains within $2^{1/6}\sigma_{ij}$ of each other. There are no specific contacts formed in either of the cases (3.10), suggesting that LLPS of FUS WT is not driven by a specific region within the protein sequence. However, when comparing the different 6E mutants at the same temperature, the degree to which different regions of the peptide interact are greatly affected (3.11). This shows that despite having nearly identical phase diagrams, the interactions involved in phase separation can vary. The average interaction strength per residue χ can also be obtained by fitting the phase diagram to Flory-Huggins theory[271, 272]. We find that there is a clear decreasing trend of χ from 0.410 to 0.325 kcal/mol with increasing number of phosphomimetic mutations (Table 3.4).

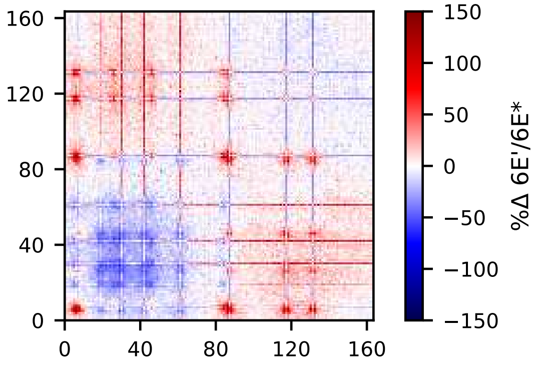

Figure 3.11: Intermolecular contacts for FUS 6E' divided by that of FUS 6E* showing how the overall number of contacts forming within the slab changes between the two sequences.

To further check the liquid-like nature of the concentrated phase, we calculate the mean squared displacement (MSD) as a function of time using NVT simulations for WT, 6E and 12E at 500mg/mL (Fig. 3.12). For each, there is a linear region with non-zero slope suggesting that the concentrated phase is liquid-like, and not a solid aggregate. The diffusion coefficient from fitting the linear region is ~3x10^{-6} cm^2/s, three orders of magnitude larger than measured in the experiment (4x10^{-9} cm^2/s [29]) as can be expected from a coarse-grained simulation and as we are using an increased relaxation time with Langevin dynamics. Finally, we check monomer radius of gyration in both the dilute and concentrated phase and find that chains in the concentrated phase are generally more extended than those in the dilute phase (Fig. 3.13).

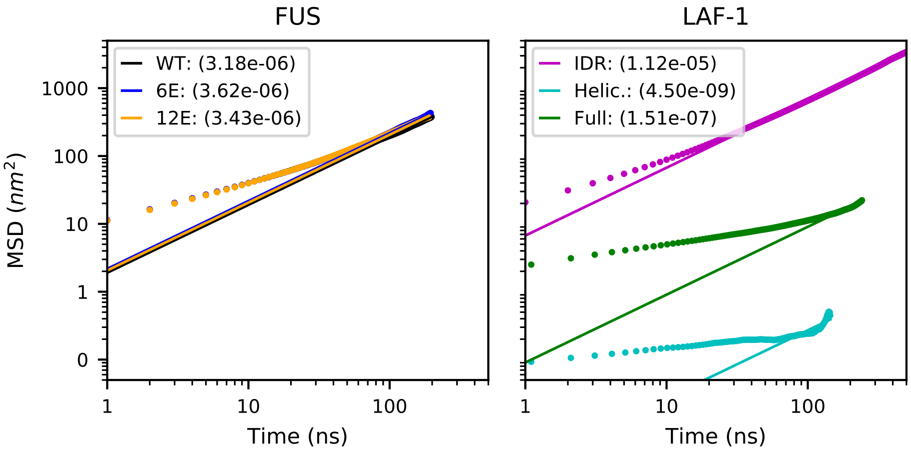

Figure 3.12: Mean squared displacement (MSD) as a function of time for FUS variants at 260K and 600 mg/mL (left), and LAF-1 at 210K and 260, 535 and 500 mg/mL for IDR, helicase, and full length respectively.Linear fits were calculated using all data points after 100 ns. Diffusion coefficients are included in parentheses in units of cm^2/s.

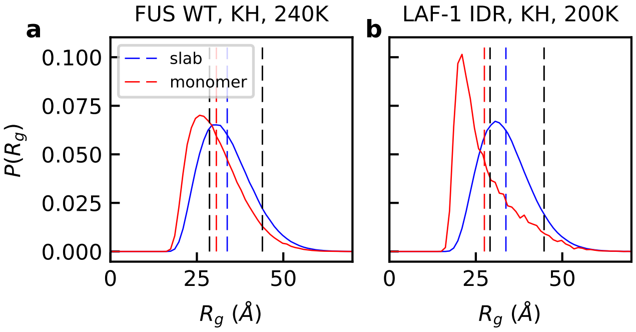

Figure 3.13: Radii of gyration of the disordered proteins inside (blue) and out of (red) the slab. A) FUS WT with KH model at 240K. B) LAF-1 IDR with KH model at 200K. Black lines show the R_g from the random coil or excluded volume chain with the same chain length.

3.3.2 Phase separation of IDR and full length LAF-1

Next, we apply our model to DEAD-box helicase protein, LAF-1, which has been shown to phase separate as both its IDR and as full length, including a 437 residue folded domain, in vitro[110]. To test the effect of inclusion of folded domains, three variants of LAF-1 sequences have been simulated, including the N-terminal IDR of LAF-1, the helicase domain, and full length LAF-1 with both the IDR and folded domain as well as the prion-like C-terminal domain which is also disordered. The IDR sequence is of similar length to FUS, but contains a larger fraction of charged amino acids, (~26%) compared to FUS WT (~1%), and FUS 12E (~9%), and includes both attractive and repulsive electrostatic interactions. For LAF-1 IDR, we simulated the phase diagram with both KH and HPS models. As was the case for FUS WT, the phase diagrams are qualitatively similar between the two models.

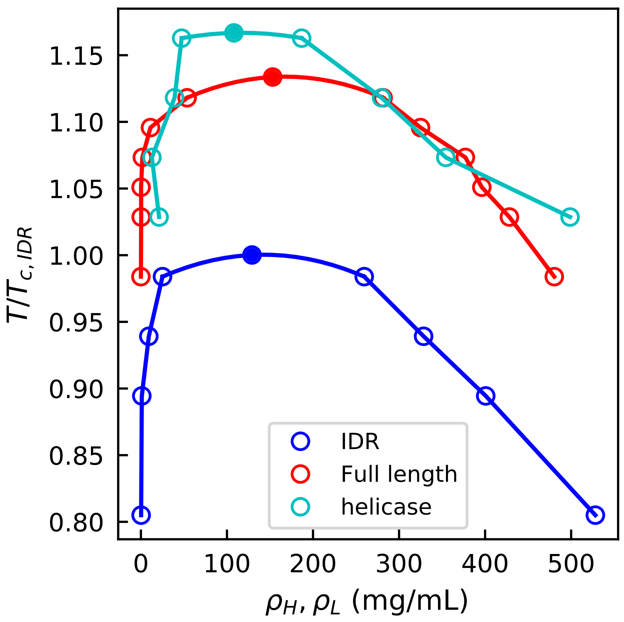

Figure 3.14: Phase diagram of IDR (blue) full length (red), and helicase domain (cyan) of LAF-1. Temperatures are scaled by the critical temperature of IDR LAF-1.

In Fig. 3.14 we compare the phase diagrams of the full length and IDR regions of LAF-1. The phase diagram for the full length protein is shifted toward higher temperatures, and suggests a smaller saturation concentration as compared to the LAF-1 IDR alone at the same temperature. The results for the helicase domain alone also clearly show phase separation (Fig. 3.14). The experimental phase boundary in ~ 120 mM NaCl is ~ 0.05 mg/mL for full length LAF-1, but ~ 0.4 mg/mL for the isolated IDR[110]. Even though we cannot accurately estimate the low protein concentrations in the dilute phase so as to quantitatively compare with the experimental values, we do see an increase in the saturation concentration when adding the folded domain as has been seen by experiment. We note that the concentrations obtained from the high density phase are much higher than recently estimated by Wei et al. [14], however, they are quite comparable with those measured by Brady et al. for the similar DEAD-Box Helicase protein Ddx4 [27]. Fitting the phase diagram to Flory-Huggins theory, we obtain the average interaction strength per residue, χ, of LAF-1 IDR (Table 3.4). The χ is 0.270 kcal/mol for KH model and is 0.298 kcal/mol for HPS model, both comparable to 0.3 kcal/mol obtain from experimental Ddx4 data [27].

The reason for the change of critical temperature upon inclusion of the folded domain is likely two-fold. First, the folded domain contains more hydrophobic residues with an average hydrophobicity of 0.664 (0.579 for the surface residues) in contrast to 0.520 for LAF-1 IDR (3.1), therefore strengthening the intermolecular attraction. In addition, providing more interaction sites per chain generally favors a higher critical temperature, because more interactions can be formed with a smaller loss of entropy, an effect commonly referred to as multivalency [80]. The impact of multivalency on the phase coexistence will be investigated explicitly in the next section.

In the concentrated phase, we also investigate the intermolecular contacts in Fig. 3.15. Unlike the case of FUS, there are regions along the sequence where there is a relatively high propensity to form contacts, (residue 21 to 28, RYVPPHLR) and (residue 13 to 18, NAALNR). These regions are present in both the IDR with the KH and HPS model (Fig. 3.15A and B) and in the full length protein (Fig. 3.15C). The central region of these two segments is composed of uncharged amino acids, suggesting the importance of hydrophobic patches in the sequence even with a large fraction of charged residues. As is shown in both 1D and 2D contact maps (Fig. 3.15A, C and d), the pattern, and number of contacts within the IDR look similar in both the IDR and the full length LAF-1 simulations. This suggests that the key residues contributing to the droplet formation are the same for the disordered peptide with and without the folded domain (Fig. 3.15D). Additionally, the disordered part of the protein (including both the N-terminal and C-terminal disordered regions) contributes more contacts than the folded domain in the simulation

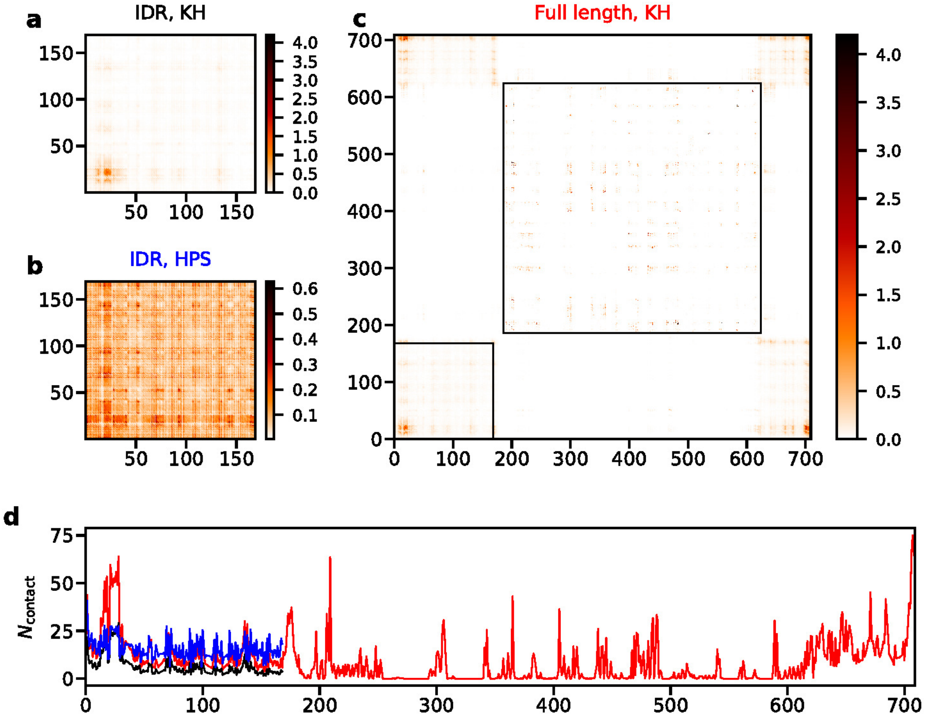

Figure 3.15: Number of intermolecular contacts per frame for LAF-1 with different models at 220K. A) Contact map of IDR LAF-1 with KH model. B) Contact map of IDR LAF-1 with HPS model. C) Contact map of full length LAF-1 with KH model. Black boxes illustrate the N-terminal IDR and the folded domain. D) Number of intermolecular contacts per residue per frame for IDR LAF-1 with KH model (black), IDR LAF-1 with HPS model (blue) and full length LAF-1 with KH model (red).

of full length LAF-1, consistent with the experimental observations that the disordered region of LAF-1 is the driving force for the LLPS [110].

We additionally calculate the mean squared displacement (MSD) as a function of time for all the three variants of LAF-1 (i.e., IDR, helicase and full length) using NVT simulations at concentrations predicted for the condensed phase at 210K, to see how the different regions affect the diffusion of the protein within the concentrated phase. There is a linear region with non-zero slope for all the variants (Fig. 3.12) suggesting liquid-like behavior. The IDR has a much larger diffusion coefficient than both the full length and the helicase domain of LAF-1 making it the most mobile of the three. This is likely due to its flexibility as well as the lower concentration. The diffusion coefficient for full length LAF-1 is an order of magnitude larger than that of just the helicase domain, further supporting the importance of the flexible region for maintaining liquid-like behavior of proteins inside the droplet.

3.3.3 Multivalency of IDRs

Multivalency has shown to be important in driving LLPS in experimental studies [80, 235] where proteins with a higher number of repeated units begin to form droplets at lower concentrations. Usually multivalency is used to describe a certain number of specific interaction sites per molecule. For polymers, there is inherently a large number of possible interactions between molecules, so for well-mixed sequences specific residue-residue interactions are less likely to play a role in assembly. Nonetheless, increasing the chain length will (for a given sequence composition) increase the number of available interaction sites per chain, and thus, increase multivalency of the system.

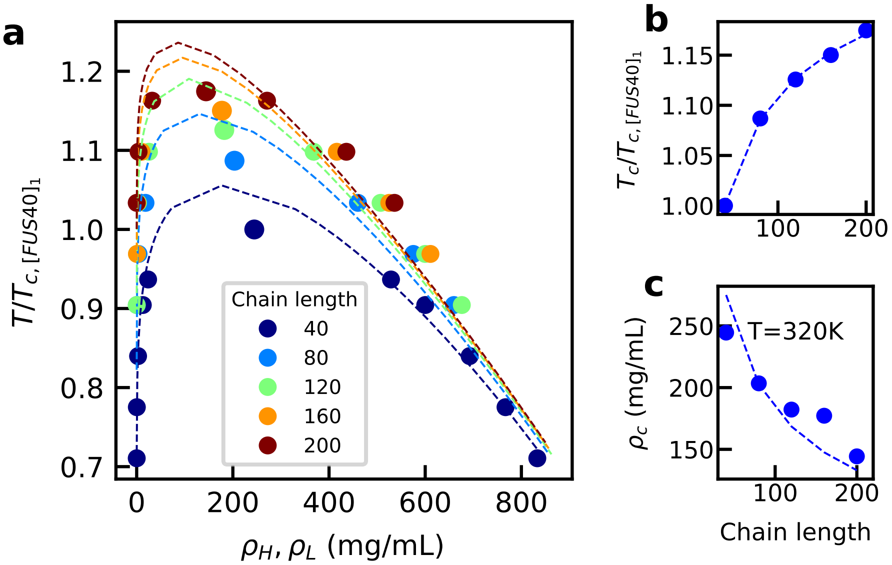

Figure 3.16: Phase separation of truncated FUS fragments of different lengths. A) Phase diagram for each peptide. Dashed lines show the fitting to binodal of Flory-Huggins theory. B) The critical temperature. Dashed lines show the fitting using relation $T_c \propto N/(\sqrt{N}+1)^2$ with prefactor as the fitting parameter. C) The critical concentration. Dashed lines show the fitting using relation $\rho_c \propto 1/(\sqrt{N}+1)$ with prefactor as the fitting parameter. Temperatures are scaled by the critical temperature of $[\text{FUS40}]_1$.

In order to investigate the mechanism of such behavior, we use a model system where we take the first 40 residues from FUS LC and make several repeated units in the form of $[\text{FUS40}]_n$, in which n=1, 2, 3, 4 and 5. We then conduct multiple slab simulations for each of these sequences, keeping the total number of atoms constant (see detailed system size in 3.4). The phase diagrams of $[\mathrm{FUS40}]_n$ in Fig. 3.16A show that the phase boundary shifts to higher temperatures and lower concentrations with increasing chain length.

To understand the mechanism of such dependence, we apply Flory-Huggins theory[271, 272], which has previously been used to understand IDP phase separation[27, 83, 87, 243], to fit the phase transition properties obtained by molecular dynamics simulations when varying the chain length N. If we assume that each solvent molecule occupies one lattice position, we can fit all five phase diagrams from different chain lengths to the binodal of Flory-Huggins theory using the same set of average interaction strength per residue χ and protein density ρ (Fig. 3.16A and 3.4). Since there is analytic solution for the critical temperature and concentration from Flory-Huggins theory: the critical temperature $T_c \propto N/(\sqrt{N}+1)^2$ and the critical concentration $\rho_c \propto 1/(\sqrt{N}+1)$, we can also fit our simulated T_c and ρ_c as a function of the chain length with these approximating equations (assuming prefactor as the fitting parameter), as shown in Fig. 3.16B and C. These results suggest that the phase diagram dependence on the chain length can be described by Flory-Huggins theory. The term that is sensitive to changes in chain length is the mixing entropy per segment. With increasing the chain length, the mixing entropy per segment decreases, and therefore the critical temperature increases. It would then be easier to observe LLPS with a longer chain at the same temperature, in the sense that the dilute-phase concentration is smaller, consistent with experimental observations [235].

This factor should be considered when making mutations to protein sequences with the aim of understanding the molecular origin of LLPS: in general, chain truncation or extension will disfavor or favor LLPS, respectively, regardless of the sequence-specific effects. Similarly, when cutting a larger protein into fragments in order to evaluate the contribution of each to driving LLPS, it is expected in general that longer fragments will be able to phase separate at a higher temperature.

3.4 Conclusions

We have introduced a general framework for conducting molecular dynamics simulations of LLPS leading to protein assemblies constituting many membraneless organelles. Coarse-graining to amino-acid-resolution gives access to length and time scales needed to observe this phenomenon, and to achieve convergence of thermodynamic observables (i.e., phase diagram, critical temperature and protein concentration in the dilute and concentrated phases) while preserving sequence-level information, thus allowing observance of changes induced by mutations to the protein sequence. The force fields utilized in this work are based on previously determined, knowledge-based potentials, parameterized to accurately represent radius of gyration of disordered proteins, but the framework is also flexible to incorporate other residue-based pairwise interaction potentials. The two force fields generate similar intermolecular contact maps within the concentrated phase, suggesting that description of the weak nonspecific interactions in IDPs can be captured easily with different models as compared to the description of specific interactions involved in binding between folded proteins.

We have tested the framework and the two force fields with two model systems, which undergo phase separation in vitro, yielding phase diagrams, thus giving the critical temperature, and saturation concentration at the tested temperatures. Despite that simplicity of the currently used potentials, and the fact that they were exclusively optimized based on the properties of monomeric proteins, we demonstrate the ability to predict how various perturbations to the system can change the LLPS. In the case of FUS LC, the model is able to capture the experimentally observed variation of phase diagram when introducing mutations. In LAF-1, the model is able to capture the experimentally observed difference between the phase separation of full length and truncated disordered-only sequences. We also show that the inclusion of the disordered parts function to increase the diffusion of LAF-1 within the condensed phase.

We have also investigated an important feature of LLPS regarding the dependence of phase behavior on chain length, which is well established in polymer physics and was previously observed in experiment [235]. We show that there is an upward shift in the phase diagram (temperature-concentration) with increasing chain length. At a given temperature, the saturation concentration will be higher for shorter chain lengths. Both the critical temperature and concentration are in good agreement with Flory-Huggins theory and therefore suggest the behavior can be explained by relative loss of entropy. With this in mind, if the phase behavior of a protein of interest cannot be observed in vitro, making repeated units might be a convenient way to shift the phase diagram enough that LLPS will be observable under more reasonable experimental conditions. One must also consider this effect when making changes to protein length, such as His tags, or cleavage of a certain section of residues, and how just the change in chain length may affect the coexistence.

Additionally, we measure certain important properties of proteins within the concentrated phase for the two model systems such as intermolecular contact propensities, which are quite difficult to resolve experimentally. With FUS LC, the intermolecular contacts are evenly distributed throughout the length of the peptide, suggesting that non-specific hydrophobic interactions are largely responsible for driving the phase-separation. For LAF-1, we observe enhanced intermolecular contacts within a specific region (residue 21-28), largely composed of hydrophobic amino acids, suggesting that even though LAF-1 containts 26% charged residues, hydrophobic interactions are still an important driving force for LLPS.

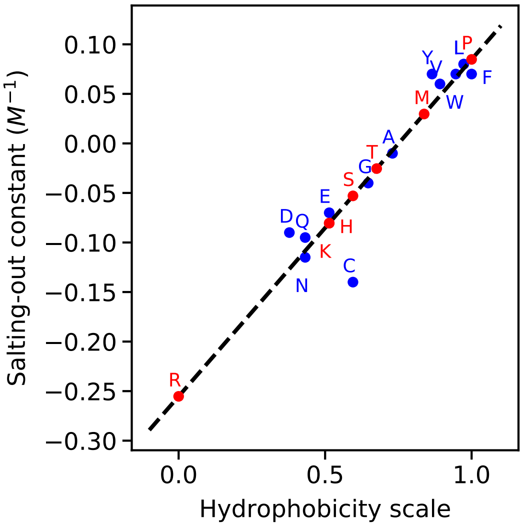

Figure 3.17: The correlation between salting-out constant and hydrophobicity scale. Black line shows the linear fitting curve between these two parameters. Blue dots show the data from literature for sodium chloride[276–279] and red dots show the estimate from linear interpolation or extrapolation.

There are some features that cannot be captured in the presented model, but can be added in the future work. First the absolute temperature of the simulation is not comparable to the experiment. The phase behavior at the lower critical solution temperature, which is observed in some disordered peptides experimentally [26], cannot be captured, either. Both require the addition of a temperature dependent solvation energy term into the framework, and more comparison with experimental R_g data (or other relevant data). Second, we have not fully tested the ionic strength dependence of the current model because of the breakdown of Debye-Hückle electrostatic screening at high ionic strength, even though the trend of LAF-1 experiment when varying salt concentration is captured in the current model. However, we do not see any ionic strength dependence for FUS LC, which is inconsistent with the experiment[29]. To capture salt dependence in proteins with negligible charged amino acid content, it may be necessary to include a description of "salting-out" effects, i.e., the change of solubility with salt concentration as captured by the Hofmeister series. In 3.17, we show that the literature-known amino acid specific salting-out coefficients [276–279] are strongly correlated with the hydrophobicity scale and therefore it may be possible to model the salting-out effect with an additional energy term using the same hydrophobicity scale. In the future, we would also like to introduce additional handles (such as a structure-based potential for intramolecular interactions) to allow for conformational changes within the folded parts of a chain. This will allow us to study LLPS of proteins with small populations of folded regions that are important for self-assembly.

Chapter 4

Relating single-chain properties to phase separation[1]

4.1 Introduction

Liquid-liquid phase separation (LLPS) of proteins and nucleic acids has recently drawn significant attention due to its relevance to physiological functions[2, 233–235], disease pathology[11, 120, 238] and design of self-assembling materials with tunable properties[26, 280, 281]. One major reason for this interest in LLPS is its relation to intracellular compartmentalization via the formation of membraneless organelles[27, 29, 110, 148]. Experimental studies suggest that the main driving force for LLPS for many of the proteins involved comes from their intrinsically disordered domains, i.e. those which lack a stable folded structure [29, 110]. A working hypothesis is that the disordered domains drive LLPS under physiological conditions, thus increasing the effective concentration of other (folded) domains (e.g. RNA-binding domains) which carry out additional functions by recruiting other biomolecules such as RNA[27, 55, 65, 110, 237]. Also, many proteins that are known to form dynamic liquid-like assemblies through LLPS can also form pathological solid-like aggregates. It has been suggested that this high-density phase might help overcome nucleation barriers and promote aggregation [11, 120]. The effects of known disease-related mutations, which promote both LLPS and formation of solid-like aggregates, provide evidence in support of this

[1] reproduced from ref.[92]

hypothesis[11, 55]. LLPS-susceptible protein sequences are also promising for the design of new materials for a variety of applications[15, 26, 280].

Relating sequence to phase separation properties is a major goal in this work, and will allow for the exploration of the proteome to identify sequences which may participate in LLPS, and will aid in directed design of peptide sequences having desired assembly, and material properties. An excellent study by Quiroz et al. has already provided much evidence toward the relationship between amino acid composition and phase behavior by measuring the demixing temperature for a large number of sequences[26]. Several studies have also addressed the relation between the degree of collapse of an IDP in dilute conditions and the phase boundaries. Lin and Chan observed a correlation between the radius of gyration (R_g) of an isolated protein and the LLPS critical temperature (T_c) for a collection of polyampholytic sequences[37, 95]. Riback *et al.* found a similar correlation between the single chain R_g and the demixing temperature (T_{demix}) for sequences of differing hydrophobicity[13]. Both of these studies, despite focusing on different aspects of protein sequence properties suggest that the R_g of a single chain in dilute solution is linked to the temperature associated with phase separation. However, such a correspondence is not expected to hold well for proteins of different lengths due to the differences in the scaling of R_g and T_c as a function of chain length.

LLPS of intrinsically disordered proteins (IDPs) shares common features with the well-established field of polymer solution phase behavior, suggesting that existing polymer physics principles might be a good starting point toward a better understanding of the underlying phenomenon[26, 83, 93, 243]. For example, the collapse of a single, isolated homopolymer chain and polymer solution phase separation are related via the effective monomer-monomer interaction strengths in the Flory-Huggins theory[271, 272]. Panagiotopoulos *et al.* have shown that the temperature of the coil-to-globule transition (T_θ) and T_c are equal in the limit of infinite chain length[155]. Wang *et al.* applied self-consistent field theory to also show the correspondence between T_θ, the Boyle temperature (T_B) at which the second virial coefficient $B_{22} = 0$, and T_c [282]. We investigate whether these theories in polymer physics can also be applied to understand LLPS of finite-length, heteropolymeric IDPs.

Our goal is to find a general relationship between T_θ, T_B, and T_c for a variety of IDP sequences.

To accomplish this, we use a recently developed coarse-grained (CG) computational framework which is capable of obtaining all three of these properties [144]. We have shown previously that this approach is sufficient to capture qualitative trends in phase behavior as induced by sequence mutations or by inclusion of a folded domain[33, 35, 144]. We calculate these three characteristic temperatures for 20 different IDP sequences and mutants with diverse hydropathy, charge, patterning and sequence length from experimental and theoretical studies on IDP phase separation[14, 29, 35, 55, 144]. We further calculate T_θ for a set of 30 synthetic polyampholyte sequences and compare with T_c determined from theoretical methods to show this correlation is not specific to our CG framework[37]. We observe a strong linear correlation among T_θ, T_B and T_c. This highlights the utility of polymer physics principles in discovering predictive models of protein LLPS. Moreover, since it may be easier to obtain T_θ and T_B via simulations or some experimental methods, they can serve as proxy descriptors of LLPS, and allow for high-throughput screening of sequences.

4.2 Simulation Methods and Analysis

4.2.1 Coarse-Grained Model

We employ our recently developed C_α-based model, where proteins are represented as flexible chains, and each amino acid residue is considered as a single particle. Bonds are modeled using harmonic springs with a spring constant of 10 kcal/(mol Å^2) and a bond length of 3.8 Å. Long-range electrostatics are modeled using a Coulombic term with Debye-Hückel electrostatic screening[111], having the functional form:

$$E_{ij}(r) = \frac{q_i q_j}{4\pi D r} \exp(-\kappa r), \tag{4.1}$$

in which $\kappa^{-1} = 10\AA$, the Debye screening length corresponding to approximately 100 mM salt at room temperature, and $D = 80$, the dielectric constant of water. Nonbonded pairwise interactions are modeled using one of the two knowledge-based potentials we have previously applied to these systems[144].

The first pairwise interaction model, the hydrophobicity scale (HPS) model is based on amino acid residue hydrophobicity from Kapcha and Rossky[252], and applied to a Lennard-Jones-like functional form which can be used to scale the strength of interactions based on hydrophobicity[253]:

$$\Phi(r) = \begin{cases} \Phi_{LJ} + (1-\lambda)\epsilon, & \text{if } r \leq 2^{1/6}\sigma \\ \lambda\Phi_{LJ}, & \text{otherwise} \end{cases} \tag{4.2}$$

in which Φ_{LJ} is the standard Lennard-Jones potential and λ represents hydrophobicity. ϵ is set equal to 0.2 kcal/mol in order to minimize deviation of R_g from multiple FRET and SAXS experimental measurements of unfolded proteins[144].

The second model used is the Kim-Hummer (KH) model which was derived from the Miyazawa-Jernigan pair potential[128] for use with weakly binding folded proteins[151]. The KH model can be expressed as:

$$\Phi(r) = \begin{cases} \Phi_{LJ} + 2\epsilon, & \text{if } \epsilon > 0 \text{ and } r < 2^{1/6}\sigma \\ -\Phi_{LJ}, & \text{otherwise} \end{cases} \tag{4.3}$$

where positive values of ϵ will result in a fully repulsive potential. The model was parameterized by the experimental osmotic second virial coefficient of lysozyme and the binding affinity of the ubiquitinCUE complex[151].

4.2.2 Simulation Methods

Slab configurations were initially generated by conducting 100 ns simulations at constant temperature and pressure, starting from a dispersed phase of protein chains with periodic boundary conditions at 150 K and 1 bar, maintained by a Langevin thermostat and a Parrinello-Rahman barostat[216]. The x- and y- dimensions were set to $\sim$ 15 nm which is sufficient to prevent chains from interacting with their periodic images. The z-dimension of the box is then extended to $>$ 200 nm. Production simulations were conducted for ~ 5 μs at constant temperature and volume. The first 1 μs of simulation was discarded as equilibration, and the remainder is used to calculate the density profile, the phase diagram and T_c. All slab simulations were conducted using HOOMD-Blue v2.1.5 [269]. The errors of the T_c were estimated by using a block average with 5 blocks.

In order to obtain T_θ, single-chain simulations were conducted at a range of temperatures using replica exchange molecular dynamics (REMD)[218], with a temperature list of 150.0, 170.1, 193.0, 218.9, 248.3, 281.7, 319.5, 362.4, 411.1, 466.3, 529.0, and 600.0 K. For the 9 polyampholyte sequences where T_θ falls outside this range, we ran additional simulations with an extended temperature range. Simulations were conducted in cubic boxes with periodic boundaries, large enough that a protein chain will not encounter its periodic image, and temperature was maintained using a Langevin thermostat. All single-chain simulations were conducted using LAMMPS[268]. For each temperature we estimated ν by fitting to:

$$R_{ij} = b|i-j|^{\nu}. \tag{4.4}$$

An alternative way of obtaining ν by using only the radius of gyration is through equation [40, 283, 284]

$$R_g^2 = \sqrt{\frac{\gamma(\gamma+1)}{2(\gamma+2\nu)(\gamma+2\nu+1)}}bN^{\nu} \tag{4.5}$$

in which $\gamma \approx 1.1615$[285]. We further estimated T_θ by interpolating the temperature at which ν=0.5. The errors of T_θ were estimated by using a block average method and dividing the entire trajectory into 5 blocks.

In order to obtain T_B, first the potential of mean force (PMF) of two protein chains was calculated via Monte Carlo (MC) method using an umbrella sampling strategy. A harmonic biasing potential was applied to center of mass distance, d, between the two proteins with a spring constant of 0.1 kcal/(mol Å^2). The center of the distance, d_0, for umbrella sampling varied from 0 Å to 102.9 Å with an interval of 3.4 Å for $d_0 < 40$ Å and an interval of 6.9 Å for $d_0 > 40$ Åso that the density of umbrella windows is doubled for the distances at which the two IDPs are in close contact. The weighted histogram analysis method (WHAM) was then used to merge the umbrella sampling data and compute the PMF[286]. The corresponding radial distribution function $g(r)$ was calculated from PMF and B_{22} is obtained from that using the following equation:

$$B_{22} = 2\pi \int_0^{\infty} [1-g(r)]\, r^2 dr. \tag{4.6}$$

The errors of B_{22} were estimated by using a block average with 5 blocks. In order to determine T_B considering the errors of B_{22}, we follow a bootstrapping strategy: by first generating 1000 sets of B_{22} data at the temperatures simulated taking into account the errors of B_{22}; second linearly interpolating the temperature at which B_{22}=0 and at last obtain T_B and the errors from the mean and standard deviation of the 1000 trials.

4.2.3 Fitting scheme for T_c

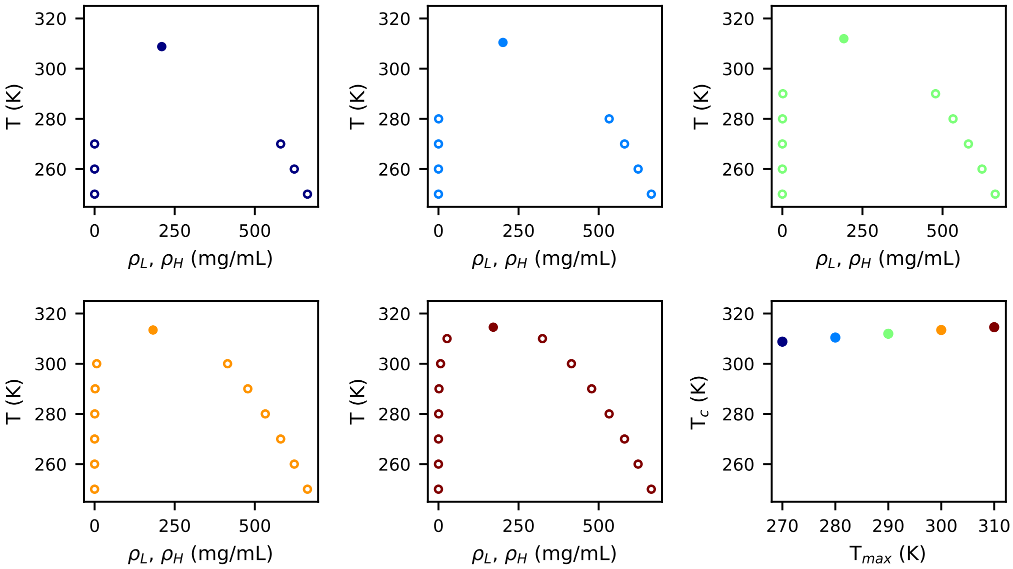

Figure 4.1: The number and location of data points has a small influence on the accuracy of the extrapolated T_c value. As the highest temperature used for fitting ($T_{\max}$) gets closer to T_c, the extrapolation is generally better.

We have described the fitting scheme for obtaining T_c from the density profile in our previous work (3)[144] and will briefly discuss here using FUS WT with KH model as an example. The critical temperature T_c can be obtained by fitting

$$\rho_H - \rho_L = A(T_c - T)^{\beta} \tag{4.7}$$

where β=0.325 is the critical exponent [270], and ρ_H and ρ_L are the concentrations of the high- and low-density phases, respectively. A is a protein-specific fitting parameter. Since we only have a rough estimate of the critical temperature for a specific IDP sequence based on their molecular properties in isolation, and their sequence composition, we always run simulations at more temperatures than usually necessary. The minimum fitting temperature ($T_{\min}$) is selected as the lowest temperature at which ρ_L is nonzero in the simulation, whereas the maximum fitting temperature ($T_{\max}$) is determined by checking the fitting errors [144]. However, we find that the fitting of T_c is largely insensitive to the number and location of temperatures used for fitting (Fig. 4.1).

4.2.4 Formation of a slab

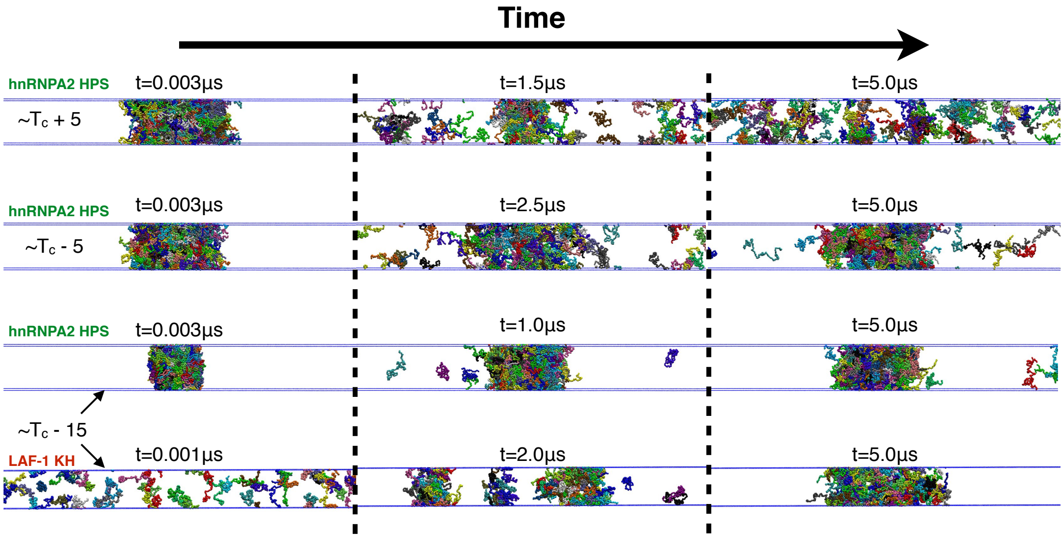

Figure 4.2: Time evolution of simulations starting from slab configuration at 300, 310 and 320K for hnRNPA2, and starting from continuous dispersed phase of LAF-1 at 210K. The slab breaks up at temperatures above T_c, while it remains phase separated at temperatures below T_c. When starting from a dispersed phase, the system eventually relaxes to a slab at temperatures below T_c after sufficient time.

To further elaborate on the validity of the extrapolated T_c, we present simulation snapshots at few time points for several temperatures in the vicinity of the computed T_c (Fig. 4.2). It is quite clear from these snapshots that the system tends to form a single phase above the T_c and remains in a two-phase coexistence below T_c, as one would expect if the computed T_c value was accurate. Moreover, a system initiated from fully dispersed protein chains at a temperature below T_c forms a dense protein phase (slab) though the process itself may take a long time thereby making it more efficient to start the simulations from a slab configuration (Fig. 4.2). The final results though will be independent of the starting configuration as we have previously shown[144].

4.2.5 Slab method comparison with other sampling methods

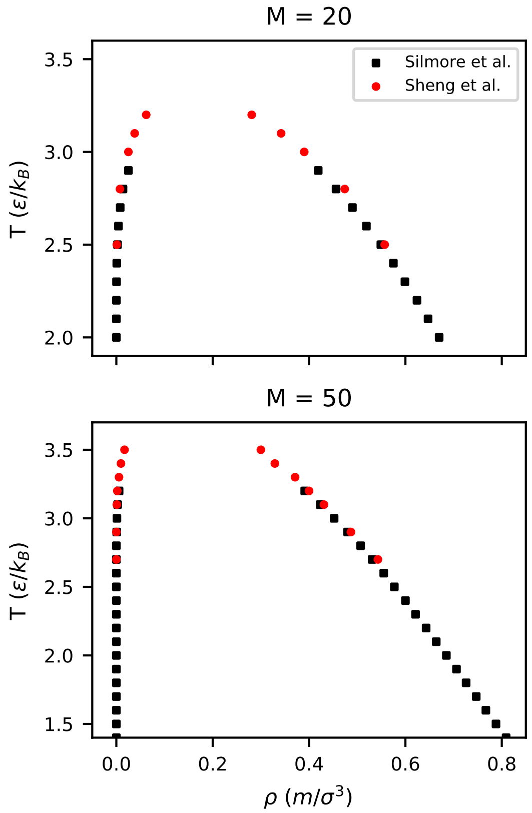

Figure 4.3: Phase diagrams calculated using different methods for fully flexible Lennard Jones chains give very similar results. Sheng et al.[287] use grand canonical Monte Carlo simulations of a small assembly of polymers and calculate chemical potential using chain increment method as an iterative approach to determine phase coexistence densities, while Silmore et al.[267] utilize molecular dynamics simulations with slab geometry similar our procedure in this work.

The use of slab method can also be justified against other methods of sampling phase coexistence, such as the agreement between results of LJ liquids from Sheng et al.[287] who use an iterative approach involving Monte Carlo simulations of flexible polymers and calculation of chemical potentials in the two phases, and from Silmore et al.[267], who utilize molecular dynamics simulations using slab method. We have plotted their data together to show they are in good agreement (Fig. 4.3).

4.3 Results

4.3.1 Obtaining phase diagrams from molecular simulations

In order to obtain the phase coexistence envelope for an IDP sequence, we utilize a recently proposed coarse-grained modeling framework [144], which has already been applied to understand the sequence determinants of specific IDPs [33, 35]. The sequences we consider include: the low complexity (LC) domain of FUS and its four different phosphomimetic variants [11, 29, 33], the multivalent repetitive FUS sequences ([FUS40]$_n$) [144], the LC domain of hnRNPA2 and two disease-related mutants[23, 35, 288], the N-terminal intrinsically disordered region (IDR) of LAF-1 [110] and five designed variants having similar or identical sequence composition, and the disordered C-terminal domain of TDP-43 [55]. We obtain the phase coexistence using two different CG potentials for several of the sequences to check whether the results presented here are model independent.

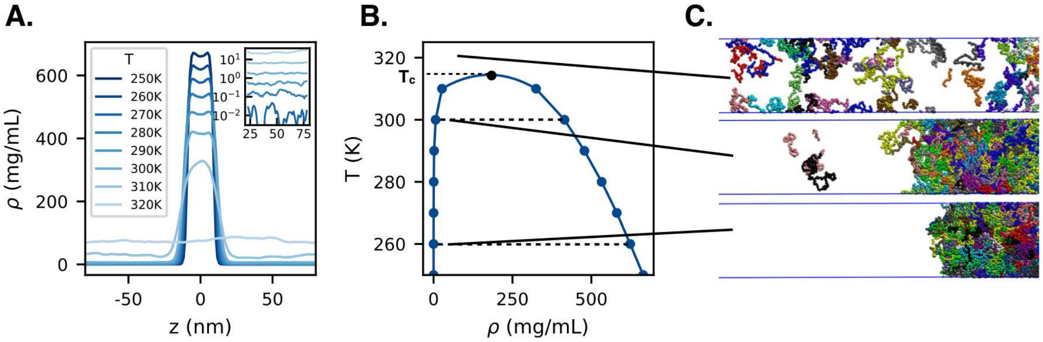

Figure 4.4: Determining phase diagrams via slab simulations. A) Density profiles, which become flat at temperatures higher than T_c. Inset shows low density phase in log scale to highlight that the density converges well and is nonzero. B) From the density profile we generate the coexistence curve of the sequence. C) Snapshots of slab simulations at super-critical (top), near-critical (middle), and below critical (bottom) conditions.

For efficient sampling of the phase diagram, we employ the slab method [144, 267], in which the simulations are initiated from a high-density protein slab which is continuous in the x, y-plane of simulation cell with elongated z-axis, allowing for efficient equilibration between the low- and high-density phases (Fig. 4.4C). This geometry has been widely used to study liquid-liquid[289], solid-liquid[290], and liquid-vapor[291] phase coexistence, and the results are highly consistent with conventional grand canonical approaches[267, 287] (Fig. 4.3). The standard chemical potential routes are computationally prohibitive for longer chains that we study here. The computational challenge can be overcome by the use of lattice models[155] or by simplifying the protein chains as patchy particles as done recently by Zhou and co-workers [150]. We instead prefer an approach that can faithfully capture the polymeric nature of IDPs and their interactions leading to the formation of a dense phase whose concentration is dependent on the protein sequence.

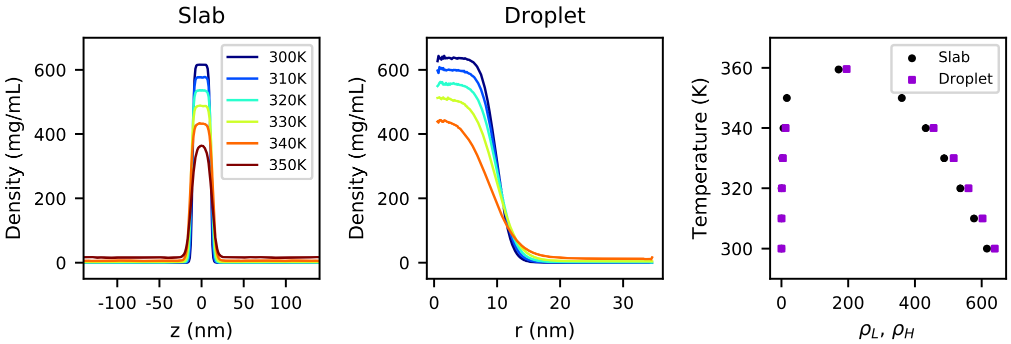

Figure 4.5: Density profiles of simulations of 100 chains of FUS using Slab and Droplet geometry, and comparison of their phase diagrams.

We show in Fig. 4.5 that simulations of phase coexistence using a cubic box with a spherical “droplet” of chains and using slab geometry produce very similar phase diagrams. By simulating the droplet geometry at different system sizes, we also find that the coexistence densities approach the values determined using slab geometry as the system size increases, even past the number of chains used for the slab simulations themselves (Fig. 4.6). We conclude that the slab geometry provides an advantage over simulations using “droplet” geometry in that it reduces finite-size effects.

For wild type hnRNPA2 at temperatures where we observe phase coexistence, we calculate concentrations in the low density phase to be in the range of 0.017-23.8 mg/mL (1.2-1870 μM), and the high density phase to be within 324.0-664.3 mg/mL (22.6-43.4 mM). These densities are in excellent agreement with experimental measurements by Ryan et al.[35], who observed μM concentration of the aqueous phase at low temperatures and estimate the concentration in the condensed phase to be between 30 and 40 mM. Brady et al.[27] also find the coexistence concentrations to be within the range of 0-50mg/mL and 150-400 mg/mL for the low-density, and high-density phases respectively for Ddx4, a sequence similar to LAF-1. We see good agreement

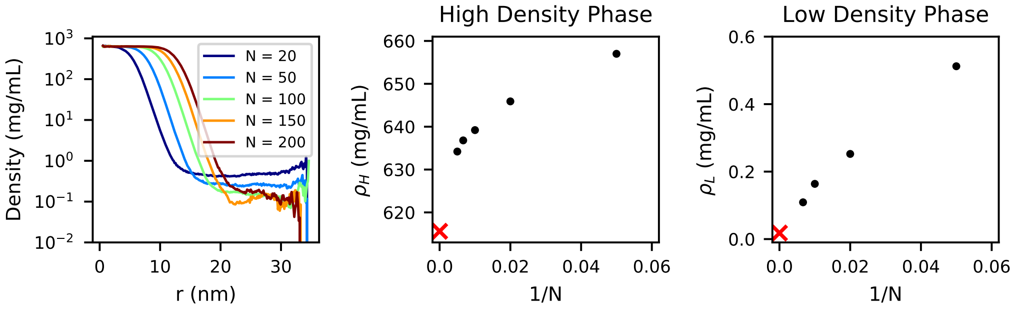

Figure 4.6: Density profiles of FUS chains at 300K using different system sizes. Slab densities are indicated by red “X”, and have been placed at 1/N = 0 as the slab geometry is expected to minimize finite-size effects. Densities in both phases tend toward the results of the slab simulation with increasing system size. For N=200, the low density phase does not converge due to the small box size, so that point has been omitted in the third subplot.

of coexistence densities with many other experimental studies as well which show proteins having saturation concentrations in the ∼1 mg/mL range and high-density concentration in hundreds of mg/mL range [29, 33, 55, 73].

As the nonbonded interactions between amino acid residues in our models are temperature-independent, we only observe the upper critical solution temperature (UCST), hereafter referred to as T_c. We calculate T_c from the coexistence densities as detailed in Methods. We note that the T_c estimated in this way is rather insensitive to the range of temperatures fitted (Fig. 4.1). We have also directly verified T_c estimated by this method in several cases by running simulations near T_c: just 5K above T_c, the slab is observed to disintegrate, while below T_c it is stable, and will spontaneously form even when the simulation is initiated from a fully dispersed initial condition (Fig. 4.2). Representative coexistence density profiles and the phase diagram are shown in Fig. 4.4A and B for wild-type hnRNPA2 along with snapshots of simulations at different temperatures (Fig. 4.4C). Critical temperatures for each protein tested can be found in Table 4.1.

4.3.2 The relation between single-chain properties and protein LLPS

We further determine R_g as a function of temperature from replica exchange molecular dynamics (REMD) simulations of a single coarse-grained protein chain. We note that the models used here provide chain dimensions and asphericity values similar to an all-atom implicit solvent model (ABSINTH) [139], which has been applied extensively to IDPs involved in LLPS [80, 292–294]

Table 4.1: List of intrinsically disordered or unfolded proteins and simulation model combinations used in this study where simulation models are hydrophobicity scaling (HPS) and Kim-Hummer (KH). Average hydropathy is calculated from the Kyte-Doolittle scale[251], q_{tot} is the net charge, and FCR is the fraction of charged residues.

Protein	Model	Length	T_c	T_θ	T_B	Hydropathy	q_{tot}	FCR
			(K)	(K)	(K)	(Kyte-Doolittle)		
FUS WT	HPS	163	359.1	332.7	464.6	-1.5030	-2	0.0123
FUS 6E	HPS	163	338.9	317.2	452.3	-1.6029	-8	0.0491
FUS 6E'	HPS	163	339.5	320.2	456.7	-1.6029	-8	0.0491
FUS 6E*	HPS	163	338.6	316.6	464.3	-1.6029	-8	0.0491
FUS12E	HPS	163	290.0	292.8	400.5	-1.7019	-14	0.0859
$[\text{FUS40}]_1$	HPS	40	316.1	275.9	422.4	-1.4247	-1	0.0250
$[\text{FUS40}]_2$	HPS	80	348.8	307.3	480.0	-1.4247	-2	0.0250
$[\text{FUS40}]_3$	HPS	120	361.5	324.9	500.7	-1.4247	-3	0.0250
$[\text{FUS40}]_4$	HPS	160	369.1	340.2	484.9	-1.4247	-4	0.0250
$[\text{FUS40}]_5$	HPS	200	374.5	349.6	485.7	-1.4247	-5	0.0250
FUS YtoF	HPS	163	372.0	350.0	505.3	-0.9000	-2	0.0123
hnRNPA2 WT	HPS	152	315.2	310.1	448.0	-1.1313	+4	0.0921
hnRNPA2 D290V	HPS	152	311.5	308.2	431.1	-1.0800	+5	0.0987
hnRNPA2 P298L	HPS	152	315.4	308.8	429.8	-1.0953	+4	0.0921
LAF-1 IDR WT	HPS	168	246.1	235.5	332.6	-1.7055	+4.5	0.2648
LAF-1 IDR P24G/P25G	HPS	168	243.1	231.5	321.3	-1.6911	+4.5	0.2648
LAF-1 IDR scramble(21-28)	HPS	168	242.7	236.3	341.4	-1.7055	+4.5	0.2648
TDP-43 CTD	HPS	141	340.4	318.4	460.7	-0.6066	+2	0.0426
FUS WT	KH	163	260.3	243.4	345.6	-1.5030	-2	0.0123
hnRNPA2 WT	KH	152	380.8	379.6	542.1	-1.1313	+4	0.0921
hnRNPA2 D290V	KH	152	384.2	390.9	534.5	-1.0800	+5	0.0987
hnRNPA2 P298L	KH	152	396.8	404.7	559.1	-1.0953	+4	0.0921
LAF-1 IDR WT	KH	168	223.6	240.4	320.9	-1.7055	+4.5	0.2648
LAF-1 IDR P24G/P25G	KH	168	213.8	236.6	316.9	-1.6911	+4.5	0.2648
LAF-1 IDR scramble(21-28)	KH	168	216.3	233.1	314.5	-1.7055	+4.5	0.2648
LAF-1 Shuffle	KH	163	265.7	287.1	381.6	-1.7055	+4.5	0.2648
TDP-43 CTD	KH	141	482.9	497.1	714.4	-0.6066	+2	0.0426

(Fig. 4.7). We also note that using the KH model, the R_g of the two sequences, hnRNPA2 and FUS are 2.2 and 3.4 nm, respectively, at 300K, which is comparable to experimental measurements also by Ryan et al. of 2.89 and 3.32 nm[35]. When comparing WT and phosphomimetic variants of FUS, we see the expected increase of chain dimensions as also reported by Ryan et al[35]. This increase in chain dimensions being accompanied by lower propensity for phase separation has been observed by several other studies[13, 33, 37], and should work well for sequences of similar length and composition.

For IDPs of different lengths, R_g scales with respect to chain length as follows [283], $R_g \propto N^\nu$, where N is the number of peptide bonds and ν is the Flory scaling exponent. As R_g by itself

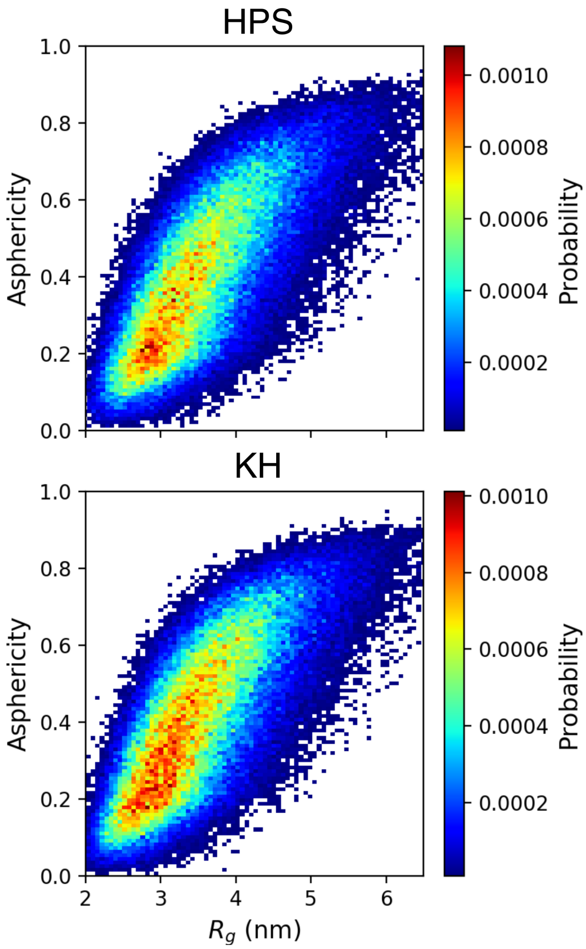

Figure 4.7: Comparison of chain dimensions of LAF-1 IDR WT at 300K between our models (i.e. HPS at the top and KH at the bottom) and ABSINTH model[139] shown by Fig. 3a in Wei et al.[14]. This shows LAF-1 is sampling both collapsed and extended conformations.

is not expected to correlate well with T_c for IDPs of different length, we use ν which we believe should scale more similarly to T_c with changes in chain length. ν can be determined directly by fitting average inter-residue distance R_{ij} between the i-th and j-th residues as a function of chain separation $|i-j|$ [266] as follows, $R_{ij} = b|i-j|^{\nu}$, where b is the Kuhn length, which is set to be 0.55 nm as suggested for disordered proteins [38, 40, 262] (Fig. 4.8A). It should be noted that for collapsed proteins R_{ij} can deviate from the power-law scaling (see low temperature data in Fig. 4.8A) due to intramolecular interactions resulting in the formation of non-specific structures. We do observe good fits for data near and above T_θ for most sequences as shown in Fig. 4.8A. For each protein, ν is estimated as a function of the temperature to determine T_θ as the temperature where $\nu = 1/2$. An example is shown in Fig. 4.8B for WT hnRNPA2 [23, 288]. At T_θ, the protein

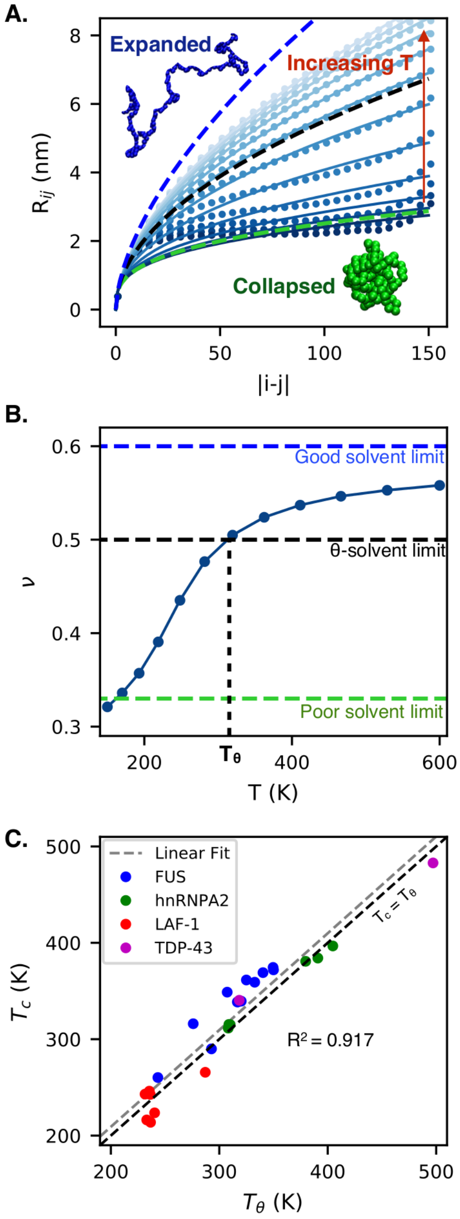

Figure 4.8: Comparison between T_θ obtained from single-chain simulations and T_c obtained from slab simulations. A) Scaling of average intramolecular distances, R_ij, with sequence separation $|i - j|$, as a function of temperature. The dashed green, black and blue curves correspond to scaling exponents of 1/3, 1/2 and 3/5, respectively. Inserted snapshots show an expanded conformation under Θ-solvent conditions, and a collapsed configuration under poor solvent conditions. B) ν crosses 0.5 at the Θ point and defines T_θ C) Correlation between T_c and T_θ. The error bars of both quantities are smaller than the symbols.

chain behaves like an ideal chain as attractive intra-protein interactions are canceled by repulsive excluded volume interactions [283]. T_θ values for all proteins tested can be found in Table 4.1.

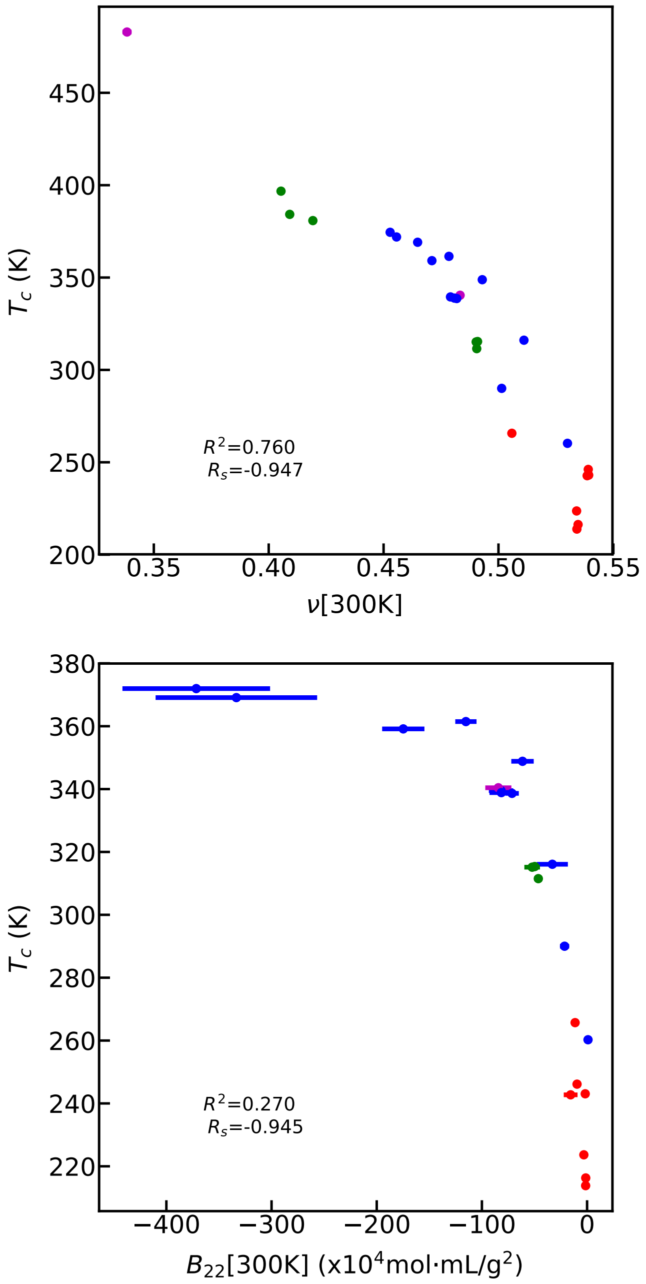

Figure 4.9: Comparison between ν or B_{22} at 300K, and T_c. R^2 shows the square of Pearson correlation coefficient, and R_s the Spearman correlation coefficient.

Previous studies have shown that T_θ of a homopolymer coincides with T_c in the limit of infinite chain length [155, 282]. We are interested in determining whether finite-length chains of heteropolymers such as IDPs may still show such a correspondence between the two temperatures. First, we test if solvent quality characterized by ν at a single temperature (300K) can provide useful information about T_c. The overall correlation between T_c and ν (Fig. 4.9) is quite reasonable ($R^2 = 0.760$). We observe an even stronger correlation (R^2=0.925) between T_θ and T_c as shown in Fig. 4.8C for all sequences with temperatures varying over a broad range, from $\approx$ 200K to 500K. The correlation between T_θ and T_c is found to be approximately linear with a slope of 1, and intercept at origin, which suggests that the temperature at which intramolecular attraction and repulsion are balanced is the same as the temperature at which intermolecular interactions in a protein solution are also canceled out. Thus, we expect that this correlation should not change significantly for experimental data on similar protein sequences, and hope that our results will motivate future experimental work in this direction. It is also notable that the correlation is independent of the computational model used, as we include data from two very different CG models of inter-residue interactions (Fig. 4.8C, 4.11C, 4.13) , further suggesting these correlations could be universal, regardless of technique or protein.

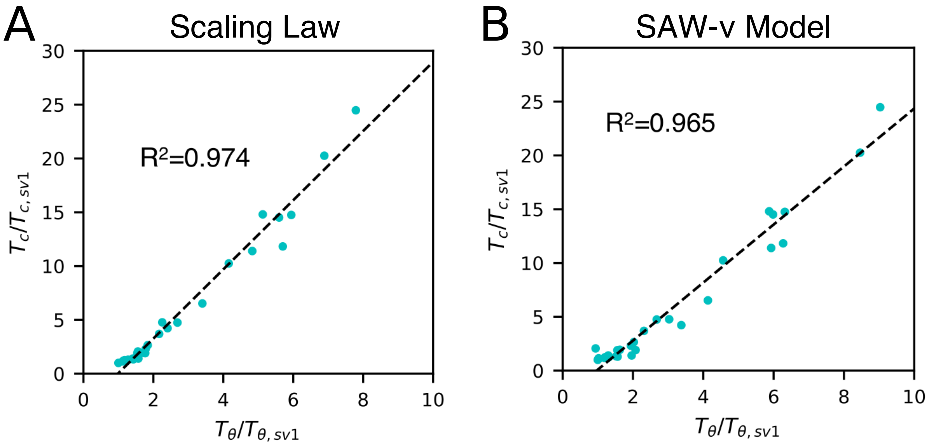

Figure 4.10: Test on designed sequences with different charge patterns[79]. T_c, as calculated by Lin and Chan[37], and T_θ, as determined from our monomer simulations obtained by A) fitting intrachain distances to the polymer scaling law or B) from the method of Zheng et al.[295], and normalized to the sequence with the lowest temperature (sv1), are plotted against each other and show good correlation.

The protein sequences considered so far do not account for additional factors which are important for the IDP behavior such as charge patterning [266]. To check if the correlation between T_c and T_θ can also be useful for capturing changes in the protein LLPS due to sequence patterning, we consider a set of 30 polyampholytic sequences with the same amino acid composition but differently arranged, ranging from perfectly mixed to block copolymer [79, 82]. Lin and

Chan recently investigated the LLPS behavior of these protein sequences (termed SV series) via Flory-Huggins theory combined with a random phase approximation method[37, 243]. We take advantage of the available T_c data from this study by comparing with T_θ from our model using REMD simulations. We find the T_c-T_θ correlation to be quite strong (R^2=0.974) for the SV series data as shown in Fig. 4.10A. We note that, for sequences having a high degree of charge segregation, the intra-chain distances do not fit the polymer scaling law well. Therefore, we have also applied an alternative method to obtain ν[40], and shown that T_θ, when calculated by this method, also correlates very well with the T_c results from Lin and Chan (Fig. 4.10B). The correlation is better than those seen for the previous 20 IDP sequences mainly due to the balanced charge and identical sequence composition. These results suggest that well-established polymer physics principles can be used to understand and predict the LLPS of IDPs. In general, when considering intramolecular interactions, as in the case of T_θ, each residue's self-interactions are ignored. Such interactions may be quite relevant for intermolecular interactions when estimating T_c, especially for sequences which are not well-mixed or charged sequences with long-range electrostatic interactions. In the case of LAF-1, we had previously identified a short 8-residue region which has high propensity of interactions with itself in other chains. Moreover, LAF-1 also has a higher fraction of charged residues which will further make it difficult to capture information on such interactions purely based on a single-chain property such as T_θ. We expect that such interactions will be better captured with measures such as second virial coefficient.

4.3.3 The relation between protein affinities in dilute solutions and LLPS

The observed correlation between T_θ and T_c suggests a strong similarity between the intrachain interactions of an isolated protein and the interchain interactions among proteins in a liquid-like phase. Interchain interactions can also be quantified by the osmotic second virial coefficient (B_{22}). Previously B_{22} has been shown to be related to the growth rates of protein crystals[296] and the temperature at which protein solutions become turbid[297]. As discussed in the introduction, T_θ of a homopolymer chain coincides with T_B in the limit of infinite chain length[155, 282]. Therefore, as in the previous section, we would like to test if such a correlation applies to finite-length IDPs which are involved in LLPS.

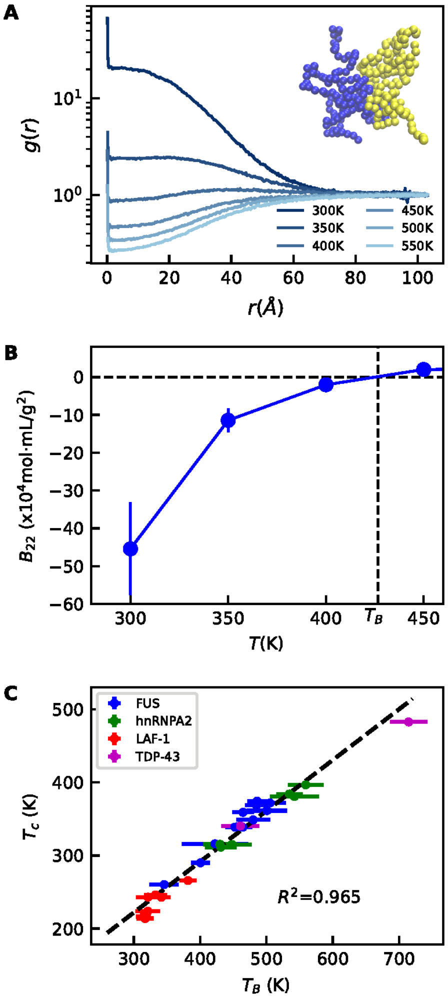

Figure 4.11: Comparison between Boyle temperature T_B, obtained from the temperature dependence of B_{22}, and T_c. A) Radial distribution function between the two isolated chains with increasing temperature. Inserted snapshots show typical configurations in the simulation. B) The temperature at which $B_{22} = 0$ defines T_B. C) Correlation between T_c and T_B. The errors of T_B are shown as bars in the figure whereas the errors of T_c are smaller than the symbols.

In molecular simulations, B_{22} can be calculated from the radial distribution function, $g(r)$, between two protein chains[298] as follows,

$$B_{22} = 2\pi \int_0^\infty [1 - g(r)]\, r^2 dr \tag{4.8}$$

It is somewhat cumbersome to estimate B_{22} from simulation for disordered proteins as the $g(r)$ sampling involves conformational changes as well as distance and orientational degrees of freedom.

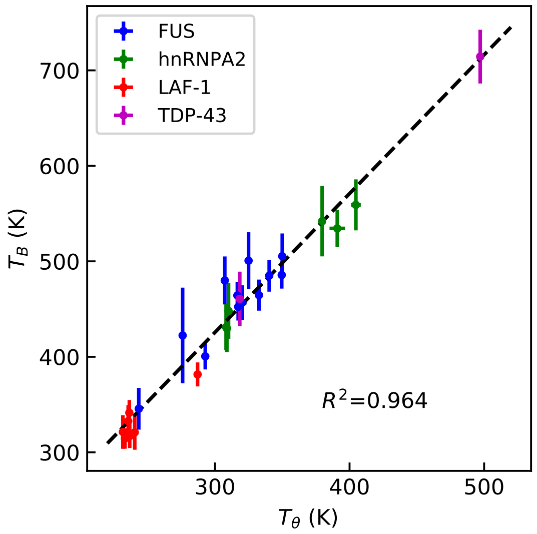

Figure 4.12: Comparison between T_B and T_θ.

However, it is still considerably less computationally demanding than sampling phase coexistence using slab geometry. Following previous work[151], we are able to determine B_{22} with excellent statistical certainty (especially near the T_B) using a combination of Monte Carlo and umbrella sampling methods. The $g(r)$ for WT hnRNPA2 is shown in Fig. 4.11A. We determine T_B by interpolating the temperature at which $B_{22} = 0$ in Fig. 4.11B for WT hnRNPA2. T_B values for all proteins tested can be found in Table 4.1. We observe a strong correlation between T_B and T_c (R^2=0.965) as shown in Fig. 4.11C (correlation between T_B and T_θ is shown in Fig. 4.12). The correlation between T_c and T_B appears to be statistically stronger than between T_c and T_θ, which suggests that for finite-length proteins, intermolecular interactions captured by B_{22} may be a better metric than ν that is based on intrachain interactions. Of course, the two temperatures differ from each other for finite-length proteins considered here as compared to the infinitely long homopolymers. The slope of the correlation is not 1 and its intercept is not zero as with T_c and T_θ. Considering the results from different models fall on the same correlation, it would be tempting to conclude that the slope and intercept of the fit to the data will remain unchanged if experimental data were also included here. Given the lack of suitable theories that can be used to assess this issue further, we reserve such judgment for future work.

4.4 Conclusions

We have shown in this work that theories originally intended for use with homopolymers of infinite chain length[155, 299] can also be applied to heteropolymeric, finite-length IDPs to relate the characteristics of dilute and condensed phase interactions. Specifically, we have observed strong correlations between T_c, T_θ, and T_B. This general correlation is encouraging for several reasons. First, the correlation suggests that our knowledge of homopolymers and the variety of tools from polymer physics are applicable to the study of LLPS of IDPs. Second, the correlation can advance the simulation prediction of T_c by using either more detailed or faster model. All-atom explicit-solvent simulations which predict single chain dimensions with high accuracy[137, 138, 300], but are also more computationally demanding, can be applied to the study of phase behavior by taking advantage of the correlation between T_c and T_θ. Rapid screening of IDP sequences for their phase behavior might be also possible if T_θ can be predicted by sequence composition (e.g. mean hydrophobicity or net charge) or by efficient coarse-grained simulations. More work is clearly needed to reach this ambitious goal. Third, this relation allows a variety of alternative experimental techniques such as single-molecular Förster resonance energy transfer (FRET)[301, 302] and small-angle X-ray scattering (SAXS) [13, 303] to be used to infer the relative propensity of different protein sequences to undergo LLPS. It is possible to estimate ν using single-molecule FRET, as previously done by the Schuler group[38] with multiple FRET dye labeling positions. In addition, recently developed methods for analysis of SAXS [39, 40] and FRET[295] allow a value of ν to be estimated more easily from a single variant of any protein with the same experimental burden as would be needed to obtain R_g. In Fig. 4.9 and 4.13, we present correlations of T_c with alternative experimental observables that may be easier to determine, depending on the context.

One can, in principle, also extend this beyond critical temperature and estimate critical salt concentration, critical pH, and other solution conditions just from the response of ν and B_{22} to perturbations. An additional advantage of using T_B as an indicator of phase separation is that it has the potential to be extended to characterize co-assembly of mixtures of different proteins and possibly of protein and RNA mixtures, which is of great biological interest[97, 148]. However, there are going to be examples where interactions are more complicated than what can be captured at the level of one or two chains using T_θ or T_B, such as systems driven by

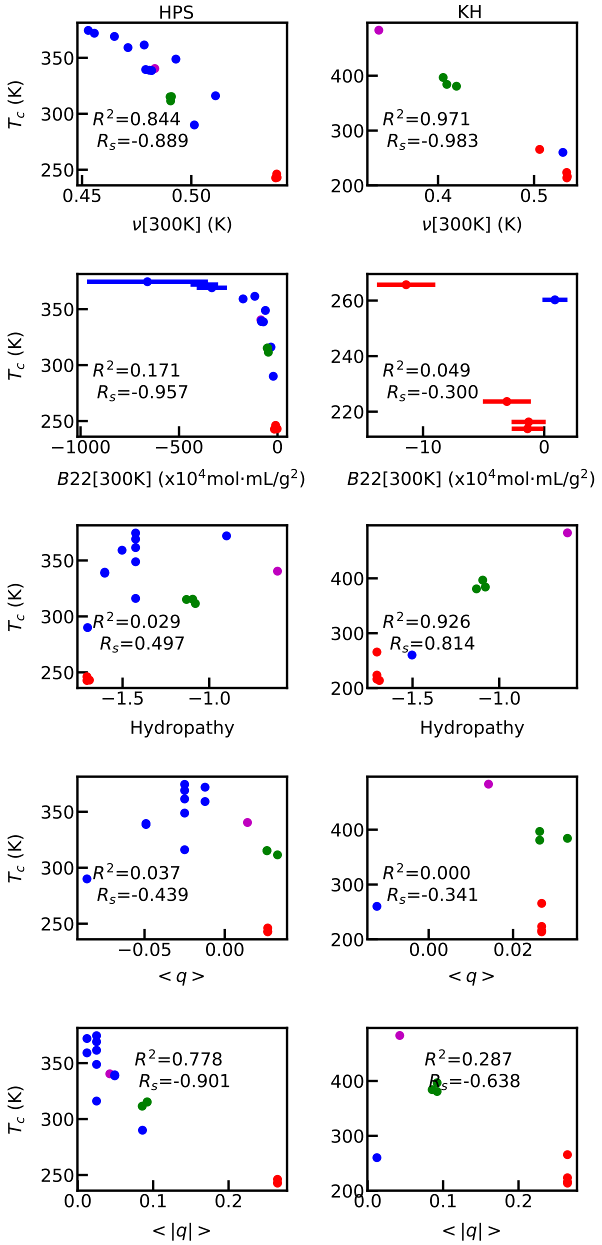

Figure 4.13: Comparison between different metrics shown in Table 4.1 and T_c using two different coarse-grained potentials. R^2 shows the square of Pearson correlation coefficient, and R_s the Spearman correlation coefficient. Hydropathy is the mean Kyte-Doolittle score [251] of residues within the sequence, $\langle q \rangle$ is the mean net charge per residue, and $\langle |q| \rangle$ is the mean absolute charge per residue.

higher order oligomerization of folded domains[43, 148]. Despite these caveats, we trust that our results here will promote rigorous characterization of IDPs at different conditions, and allow the field to progress toward solving the problem of sequence encoded phase behavior. Such a high throughput method would allow for rapid design of sequences such as those designed by Quiroz et al.[26], and would aid in the development of protein models and techniques which can be used to more accurately predict the properties of IDP LLPS.[103, 144, 148].

Chapter 5

Temperature-dependent solvent-mediated interactions enable both UCST and LCST phase transitions[1]

5.1 Introduction

It is now well recognized that cellular compartments may form in the absence of lipid membranes through liquid-liquid phase separation (LLPS), driven by proteins, nucleic acids and other biomolecules[2, 304]. These “membraneless organelles” or “biomolecular condensates” have since been shown to be highly diverse and ubiquitous within biological systems, and constitute organelles such as the nucleolus[97, 148], ribonucleoprotein (RNP) granules[29, 35], stress granules[13, 42], and many others[2, 46, 229, 305]. Protein LLPS is commonly associated with proteins containing regions that are intrinsically disordered[25, 228] and is mediated by a myriad of interactions types such as electrostatic attraction, cation-π, π-π, hydrogen bonding, and hydrophobic interactions[72, 78, 80, 306, 307]. External stimuli such as changes in salt concentration, pH, other biomolecules such as RNA or ATP, and temperature are all factors that may be used to

[1] reproduced from ref.[102]

modulate protein LLPS [29, 42, 55, 110, 117, 306].

Intrinsically disordered protein-based polymers have been used for decades in the design of functional polymeric materials for applications in biomaterials and drug delivery[308–315]. The advantages of protein-based polymers include direct control over the sequence and length by using recombinant expression[310] and the ability to directly encode functional domains such as enzyme-cleavage sites[15], light-activated domains[50, 316], cross-linking motifs[280], and substrate-specific binding motifs[15]. The high degree of control over the protein sequence also allows one to finely tune the protein LLPS in response to solution conditions and external stimuli[56]. As temperature is a factor that is easily controlled in vitro, there is a large interest in thermoresponsive LLPS[101, 317]. Thermoresponsive protein-based polymers can be designed such that they are miscible at high temperatures and demix at low temperatures, showing an upper critical solution temperature (UCST), or to demix at high temperatures and be miscible at lower temperatures, displaying a lower critical solution temperature (LCST) [26]. Tropoelastin and resilin are two proteins which are commonly used as templates to design elastin-like and resilin-like peptides (ELPs and RLPs) exhibiting LCST and UCST phase behavior respectively[26]. Some variants of RLPs have also been shown to exhibit a dual-response phase separation and will condense upon both heating and cooling, with a region of miscibility in between[100, 318]. The amino acid composition and sequence have been implicated as being responsible for the differences in phase behavior[26].

Designing IDPs with controllable LLPS is a nontrivial task due to the near-infinite selection of possible IDP sequences. Computational modelling can be an effective approach to inform experimental design and to gain insights about the sequence determinants of temperature-dependent LLPS, and the underlying physics[133, 319]. Temperature-dependent amino acid properties have previously been used in understanding properties of both folded and unfolded/disordered proteins including cold denaturation [129] and temperature-induced collapse[320]. All-atom explicit-solvent simulations can in principle provide an atomically detailed view of the interactions driving phase separation [76, 78], but are still computationally demanding and prohibitive to directly simulate protein LLPS. To overcome this obstacle, coarse-grained (CG) models, in which amino acids are simplified into CG beads and solvent is accounted for implicitly, can be used to handle sufficiently large systems and to compute phase behavior of a large number of protein sequences [85, 92]. In these cases, the solvent-mediated interactions are indirectly captured via interactions between the CG beads composing the protein molecules[85, 144, 151, 321]. Most existing CG models were built at a specific temperature (e.g., room temperature)[144, 151, 253] without taking into account the temperature-dependence of such solvent-mediated interactions[104, 322–325]. These models are not able to capture properties like disordered protein collapse with increasing temperature [103, 320] and the emergence of LCST behavior [26, 317]. Therefore, there is an urgent need for a temperature-dependent CG model, given the compelling prospect of using IDP LLPS in designing thermoresponsive materials.

In this paper, we take advantage of our recently developed CG model in which amino acid hydrophobicity was used in modelling the pairwise interactions between different amino acids [144] and introduce an amino-acid-type-specific temperature dependence into the hydrophobicity scale. We then tune the model parameters using knowledge from both single molecule Förster Resonance Energy Transfer (smFRET) experiment and all-atom simulations on the dimensions of disordered proteins across a wide range of temperatures. The optimized model successfully predicts the experimentally-known phase behavior of a large library of ELPs and RLPs qualitatively by distinguishing between UCST and LCST. This strongly suggests that the difference in the thermoresponsive behavior of a protein sequence is encoded in its amino-acid-specific solvent-mediated interactions and how these change with temperature. Using this newfound knowledge, we apply the new model to propose sequence-determinants of the protein LLPS in terms of their UCST or LCST characteristics, which should allow for the design of protein-based polymers with controllable thermoresponsive phase behavior.

5.2 Simulation and Analysis Methods

5.2.1 HPS model

Our original framework represents proteins as flexible chains of amino acids with harmonic bonds, screened electrostatics, and a non-bonded pairwise interaction potential to account for different amino acid types[144]. The full energy function of the system is:

$$\Phi(\mathbf{r}) = \sum_{bonds} \Phi_{\text{bond}}(r_{ij}) + \sum_{i<j} [\Phi_{\text{elec}}(r_{ij}) + \Phi_{\text{nb}}(r_{ij})] \tag{5.1}$$

where Φ_{bond} is a standard harmonic spring:

$$\Phi_{\text{bond}}(r) = k_{\text{spring}}(r - r_0)^2 \tag{5.2}$$

with k_{spring} = 10 kcal/(mol·$\AA$) and r_0 = 3.8 $\AA$. The electrostatic term is represented using Debye-Hückel electrostatic screening[111]:

$$\Phi_{\text{elec}}(r) = \frac{q_i q_j}{4\pi D r} e^{-r/\kappa} \tag{5.3}$$

where q_i and q_j are the net charges of formally charged amino acids, (D, E = -1; K, R = 1; H = 0.5), D is the dielectric constant, which is set to 80 for water, and κ is the screening length, which we set to 10 $\AA$ to represent a salt concentration of 100 mM. For the non-bonded pairwise interactions, we utilize a Lennard-Jones-like functional form with a tunable well depth as used by by Ashbaugh and Hatch[253, 259],

$$\Phi_{\text{nb}}(r) = \begin{cases} \Phi_{\text{LJ}}(r) + (1 - \lambda(T))\epsilon, & \text{if } r \leq 2^{1/6}\sigma \\ \lambda(T)\Phi_{\text{LJ}}(r), & \text{otherwise} \end{cases} \tag{5.4}$$

where Φ_{LJ} is the standard Lennard-Jones potential, and λ is the temperature-independent interaction strength from a hydrophobicity scale presented by Kapcha and Rossky[252]. We originally optimized the free parameter ϵ to 0.2 kcal/mol based on the agreement between the model and experimentally-determined radius of gyration (R_g) of a set of IDPs[144] (3).

We account for protein-solvent interactions implicitly through the protein-protein interaction term as more hydrophobic amino acids will have a stronger attraction, and hydrophilic will be more repulsive. The use of Debye-Hückel screened electrostatics, in addition to a standard non-bonded potential based on the amino acid contact probability, is justified by the expectation that charge-charge interactions are not fully captured in data based on folded protein structures, and that attractive and repulsive electrostatic interactions would be averaged over for the charged amino acids. Similar approaches have been used extensively in the protein CG modeling literature and have provided accurate information on protein binding thermodynamics and structure[151].

5.2.2 Molecular dynamics simulations

We conduct single chain simulations using the LAMMPS software package[268], and two-chain simulations using an in-house Monte Carlo code[151] with umbrella sampling to enhance sampling of binding and unbinding events[326]. Results from umbrella sampling were reweigted using a weighted histogram analysis method (WHAM) [286]. To efficiently sample phase coexistence, we conduct coexistence simulations in slab geometry[144, 157, 267] using the HOOMD-blue v2.1.5 software package[269]. Single chain simulations were conducted for 1 μs at each temperature, and two-chain simulations were conducted for 5×10^7 Monte Carlo steps, and coexistence simulations are carried out for up to 5 μs.

5.2.3 Empirical R_g and ν predictions

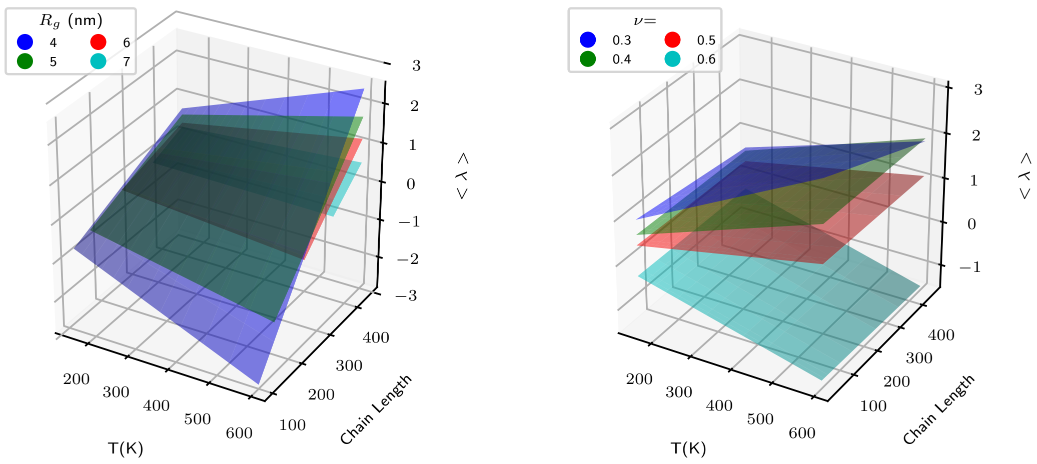

Figure 5.1: 3D contours show constant R_g (left) and ν values in the Temperature-hydrophobicity-length space. Predictions of R_g and ν can be interpolated from the known temperature, mean hydrophobicity $< \lambda >$ and the chain length of a sequence.

To empirically predict R_g of an IDP sequence based on its average sequence properties, we conducted simulations on a large number of homopolymers using the HPS model, varying two important sequence descriptors, the chain length and the average hydropathy. We use 8 chain lengths from 25 to 450 residues, and 16 average interaction strengths ($< \lambda >$) from -3 to 3, and simulated each at 12 temperatures ranging from 150 to 600K. For each of the 1536 systems, we calculated both R_g and the Flory scaling exponent (ν) [283] , to determine the dependence of chain dimensions on each of these three factors. Using this data set, we are able to use 3D linear grid interpolation approximate R_g and ν for any sequence of a given $< \lambda >$ and chain length at any temperature within the range of the data set. The dimensions of the homopolymers are visualized as a function of T, λ and chain length in Fig. 5.1. To validate the accuracy of predictions from this method, we tested 2000 randomly generated sequences with a chain length of 80 and measured R_g and ν from simulation to compare with estimates from the predictor (Fig. 5.2). We find the largest source of error to be sequences with a high net charge, which is expected since our predictor only takes into account average hydropathy of sequence. However the predictor gives less than 10% error on R_g for most sequences with a low net charge.

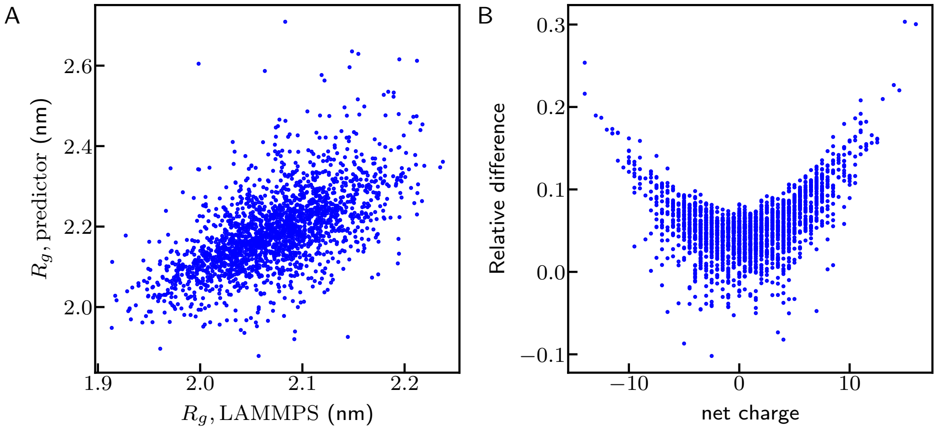

Figure 5.2: R_g of 2000 random sequences generated using the amino acid propensity of IDPs[327]. A) The comparison between the simulated R_g using LAMMPS and the R_g from the predictor (see Methods). B) The relative difference as a function of the net charge.

5.2.4 Fitting all-atom data to experimental data

The FRET experimental data for the five reference proteins[320] is first analyzed using the new SAW-ν model[40]. Since the experimental data only covers a small range of temperatures (285K-

350K), we opted to use all-atom data which gave many more data points in a larger range of temperatures (275K-500K). The simulations were conducted using the Amber ff03w[213], which has been shown to give overly collapsed configurations for proteins[138]. Toward this end, we applied a scaling and shift to the data set:

$$R'_g(T) = R_g(T) * n_1 + n_2 \tag{5.5}$$

where n_1 and n_2 are fitting parameters used to minimize the difference from the experimental data from Wuttke et al.[320] on the same protein sequences (Fig. 5.3).

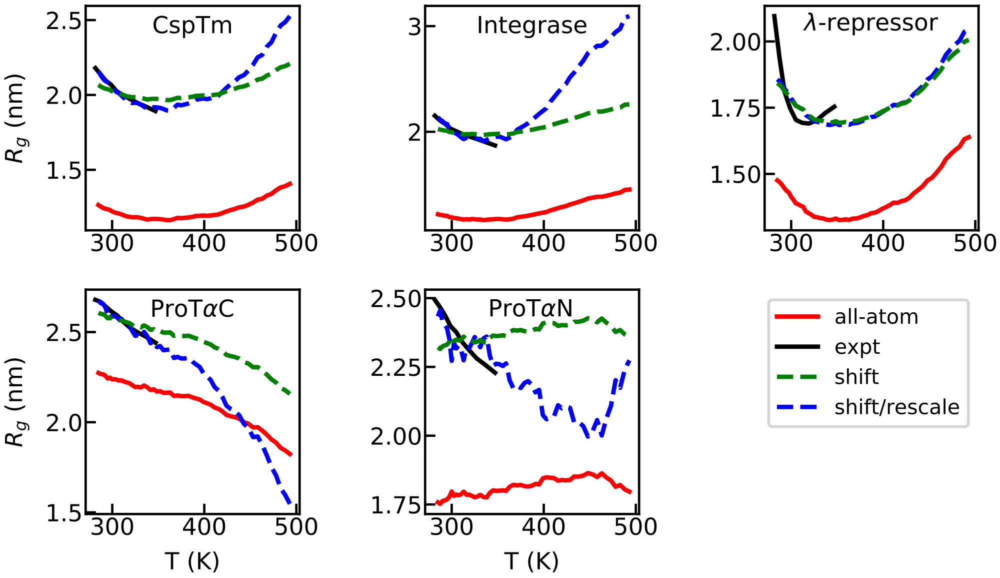

Figure 5.3: Reference data set containing temperature-dependent R_g of five sequences. The experimental data [320] analyzed with the newly proposed method [40] is shown in black; the all-atom simulation [103] in red; the fitted data set by shifting the simulation data to minimize its difference from experiment in green; and the fitted data set by shifting and rescaling the simulation data to minimize its difference from experiment in blue. For each model we parameterize to reference data in this work, we use the blue curve, except for two specific cases in which either the blue or black curve is used as a reference data set as specified.

5.2.5 Implementation of temperature-dependent models

Dill-Alonso-Hutchinson Model

To apply the thermodynamics-based model from Dill, Alonso and Hutchinson[108], we use the following functional form from their work:

$$\chi(T) = \frac{1.4}{kT}(\Delta H^\circ + \Delta C_p(T - T_0) - T[\Delta S^\circ + \Delta C_p ln(T/T_0)]) \tag{5.6}$$

where the factor of 1.4 accounts for the number of amino acids per lattice segment from the original derivation of this equation [328], T_0 = 298K, and ΔH^o, ΔS^o and ΔCp^o are obtained from experimental measurements of hydrophobic amino acid side chain analogues [329] and are equal to 0.0 cal, -6.7 cal/(mol·K), and 55 cal/(mol·K) respectively. We adopt the same set of solvophobic amino acids from that work A, F, I, L, M, V, W, Y as hydrophobic. We then apply this functional form to the hydrophobicity model as:

$$\lambda(T) = \lambda_{HPS} + \frac{kT}{\epsilon}(\chi(T) - \chi(T_{\text{ref}})) \tag{5.7}$$

where T_{ref} is set to 300K where the original HPS model was parameterized, and ϵ is 0.2 kcal/mol as in the original HPS model. Multiplying by kT results in the χkT functional form presented in that work[108]. For amino acids not considered hydrophobic, we do not impose any temperature dependence on λ.

Optimized Quadratic Models

An alternate approach is to simply use quadratic functional forms based on the bioinformatics potential derived by van Dijk et al.[330]. The potentials they derive are amino-acid type specific, which can all be represented by quadratic functional forms:

$$E_H(T) = -22.657 + 0.15379T - 0.00025597T^2 \tag{5.8a}$$

$$E_A(T) = -23.364 + 0.15876T - 0.00026696T^2 \tag{5.8b}$$

$$E_O(T) = 2.1607 - 0.015064T + 0.000026T^2 \tag{5.8c}$$

$$E_P(T) = 10.475 + 0.071482T + 0.0001201T^2 \tag{5.8d}$$

$$E_C(T) = 8.5997 - 0.057676T + 0.000093317T^2 \tag{5.8e}$$

The result from these temperature-dependent interaction terms is a temperature-dependent hydrophobicity. The average hydrophobicity of the reference protein sequences will change with temperature, but may still be predicted by the emperical R_g predictor, simply as a function of T, λ(T), and L rather than T, λ and L. These functional forms are then scaled and transposed by introducing free parameters, which are optimized to most closely fit the reference data set. Use of the bioinformatics-based data as a starting point is advantageous as it provides information into the relative interaction strengths and temperature dependences of the different amino acid types.

5.2.6 Method for sampling sequences.

To sample sequences with amino acid compositions , we start from the CspTm sequence with a chain length of 66. We then pick 10 positions randomly along the sequence and mutate each to other residues, with probabilities adjusted according to the relative abundances of different amino acids in IDP sequences[327]. We repeat this procedure 1 million times, allowing mutations to accumulate, thus generating 1 million sequences, corresponding to the average amino acid composition of IDPs. The errors of the probabilities of the phase-diagram shape and the abundance of amino acids in one phase-diagram shape are estimated using a block average with five blocks. Results are listed in Table 5.1.

We further test whether differing overall amino acid abundance has an appreciable effect on the abundance of amino acids in each phase-diagram shape. We have generated additional sequences (1 million each) using uniformly weighted amino acid probability (Table 5.2) and probability based on amino acid composition of structures from the protein data bank (Table 5.3) [327]. The

Table 5.1: Possible states in sequence space with different phase behaviors using sequences generated with amino acid compositions based on their relative abundances in IDPs [327]. Numbers in brackets show the errors of the last digit.

Phase-diagram shape	P(shape) (%)	Abundance H	A	O	P	C
IDP		0.289	0.068	0.160	0.241	0.242
none	31.42(4)	0.2631(1)	0.0627(1)	0.1533(1)	0.2501(1)	0.2707(1)
single-UCST	8.48(5)	0.20887(7)	0.04863(7)	0.1712(1)	0.2892(2)	0.2821(2)
closed-loop	59.63(9)	0.3154(2)	0.0736(1)	0.1618(2)	0.2282(1)	0.2208(2)
hourglass	0.468(8)	0.1607(2)	0.0355(4)	0.1774(9)	0.3293(9)	0.2971(6)

probabilities of sampling different phase-diagram shapes however differ considerably from those tested using amino acid abundances from IDPs. For both cases, the closed-loop shape is with a much higher probability whereas single-UCST is almost completely absent. The abundance of amino acids with different phase-diagram shape does qualitatively agree among all the three cases using different amino acid composition for generating the sequences.

Table 5.2: Possible states in sequence space with different phase behaviors using sequences generated with unbiased amino acid compositions. Numbers in brackets show the errors of the last digit.

Phase-diagram shape	P(shape) (%)	Abundance H	A	O	P	C
none	7.145(5)	0.1903(3)	0.1567(3)	0.1541(6)	0.2395(4)	0.2594(2)
single-UCST	0.166(8)	0.146(1)	0.114(2)	0.1762(8)	0.292(2)	0.272(2)
closed-loop	92.69(6)	0.2547(1)	0.2034(1)	0.1497(2)	0.1969(2)	0.1952(2)
hourglass	0.0022(5)	0.124(4)	0.072(4)	0.17(1)	0.34(1)	0.29(1)

5.3 Results

5.3.1 Use of knowledge-based temperature-dependence to scale CG interactions

Given our recent success using amino acid resolution CG models to study IDP LLPS[92, 144], we apply a similar philosophy to build a model based on amino acid hydrophobicity with temperature-dependent interactions. We previously used a top-down approach to parameterize the CG model

Table 5.3: Possible states in sequence space with different phase behaviors using sequences generated with amino acid compositions based on their relative abundances in the protein data bank [327]. Numbers in brackets show the errors of the last digit.

Phase-diagram shape	P(shape) (%)	Abundance H	 A	 O	 P	 C
PDB		0.318	0.114	0.139	0.199	0.230
none	10.25(7)	0.2501(1)	0.0912(1)	0.1405(2)	0.2302(3)	0.2879(2)
single-UCST	0.49(1)	0.1955(6)	0.0677(6)	0.1595(5)	0.274(1)	0.303(1)
closed-loop	89.26(8)	0.3265(1)	0.11679(8)	0.13884(8)	0.19500(8)	0.2229(1)
hourglass	0.0055(9)	0.151(5)	0.051(3)	0.158(6)	0.315(9)	0.325(9)

to reproduce experimental measurables using the radius of gyration (R_g) of a large number of IDPs[144] at a single temperature (300K). Building a model that accurately captures temperature-dependent IDP dimensions, and therefore, LLPS, necessitates a set of data for IDPs spanning a range of temperatures. The temperature-induced expansion and collapse of a diverse set of IDPs and denatured proteins has been observed previously in smFRET experiments[320] and all-atom simulations[103]. The proteins examined in these studies include: the cold-shock protein from *Thermotoga maritima* (CspTm), the N-terminal domain of HIV integrase, the DNA-binding domain of λ-repressor, and the N- and C-terminal segments of human Prothymosin-α (ProTαN and ProTαC). For protein sequences, please refer to Supporting Methods 1.1. These two data sets are complementary as the experimental study is limited to a smaller temperature range, whereas the all-atom results were obtained using an older force field which results in protein dimensions smaller than expected [138]. Here, we merge the two data sets to a single reference data set to take into account the desirable features of each, i.e. the wider temperature range from simulation and the quantitative accuracy of experiment (Fig. 5.3).

The amino acid composition of a protein is largely responsible for the differences in the observed phase behavior[26]. To account for the sequence-dependent behavior of proteins, we aim to develop a physics-based CG model which can capture amino acid residue-specific changes induced by temperature. Van Dijk et al. used a library of solved protein structures at different temperatures to build a knowledge-based contact potential as a function of temperature between protein residues [330]. They used a reduced classification by lumping the temperature-dependence of all 20 amino acids into five different types, generally having similar responses to temperature within each group (Table 5.4). We note that the temperature-dependence from this knowledge-based potential is also consistent with the changes in solvation free energy of amino acid sidechain analogs as a function of temperature [104, 107, 322, 323]. The resulting potential was successfully used to obtain the estimates of protein thermal stability [331] and to explain protein cold denaturation by invoking the changes in solvent-mediated interactions with temperature [129]. For further use, we fit the temperature-dependent contact potential to a parabolic function for each amino acid type (Supporting Methods 1.3). At low temperatures, the interactions between hydrophobic groups will strengthen with increasing temperature, whereas these interactions will weaken with further increase in temperature after a point of maximum strength. This behavior arises from the dominance of the enthalpic component of the free energy at low temperatures and its entropic part at higher temperatures [104, 106, 107]. The parabolic functional forms fit the bioinformatic contact potential[330] within a small temperature range, and allow us to extrapolate to a wider range of temperatures relevant to the experimental studies of thermoresponsive phase separation. We anticipate this is a reasonable assumption to capture the qualitative changes in single chain compaction and phase behavior.

Table 5.4: List of amino acids and type classifications.

Hydrophobic (H)	Ala, Ile, Leu, Met, Val
Aromatic (A)	His, Phe, Trp, Tyr
Other (O)	Cys, Gly, Pro
Polar (P)	Asn, Gln, Ser, Thr
Charged (C)	Arg, Asp, Glu, Lys

The non-bonded interactions in our original CG model is based on the hydrophobicity (λ) of each amino acid (see Methods) [144]. Therefore the contact potential described above can be used to introduce temperature dependence on λ in the model by an appropriate scale as

$$\lambda_i(T) = \lambda_{i,HPS} + \frac{1}{\epsilon}[E_X(T) - E_X(T_{\mathrm{ref}})], \tag{5.9}$$

where i is the amino acid, X is the amino acid type, E_X is the corresponding parabolic function from Eq. 5.8, ϵ is 0.337 k_BT (0.2 kcal/mol), as in the original HPS model[144] (See Methods) to convert to the correct units for use with the LJ-like functional form, $\mathrm{T_{ref}}$ is the reference

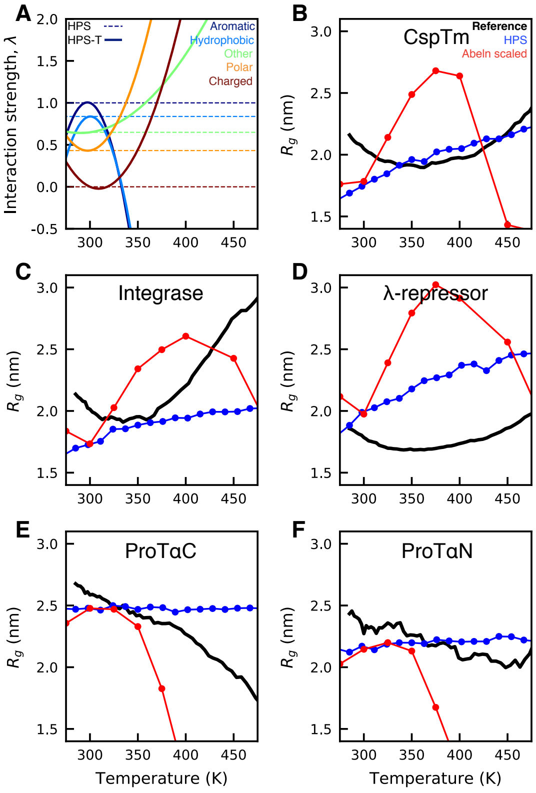

Figure 5.4: A) Model resulting from directly using functional forms from van Dijk et al.[330] rescaled to units of 0.2 kcal/mol so that well depth follows the same temperature dependence. Wellness of fit to reference data for proteins B) CspTm, C) Integrase, D) λ-repressor, E) ProTαC, and F) ProTαN.

temperature (300K) for which the model will be equal to the original HPS model, and $\lambda_{i,HPS}$ is the hydrophobicity value for residue i in the original HPS model (Table 3.1). To test the new model, we simulate the five proteins for which radius of gyration is available as a function of temperature from experiment and all-atom simulations[103, 320]. In contrast to the original HPS model, we are able to observe a non-monotonic trend in R_g as seen in experiment (Fig. 5.4), although it is in poor qualitative agreement. Specifically, CspTm, integrase and λ-respressor do initially collapse and then expand, however the turning point of R_g is at about 300K instead of about 350K seen in the reference data set, which suggests allowing for the shifting of the contact potential extrema is necessary for further refinements.

To better capture the reference data, we modified our approach to empirically define a temperature-dependent model which quantitatively agrees with the reference data by making

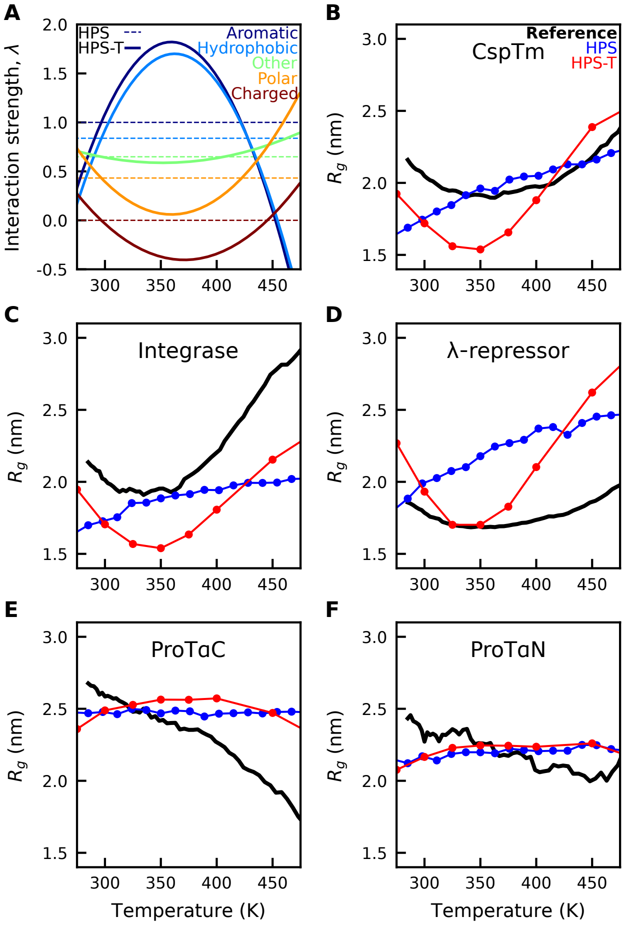

Figure 5.5: Temperature-dependent interaction potential and protein dimensions. A) Original λ values from HPS potential shown as dashed lines, and new temperature-dependent model is shown as solid lines. Example HPS values are shown for phenylalanine (Aromatic), methionine (Hydrophobic), glycine (Other), asparagine (Polar), and arginine (Charged). B-F) Experimental/all-atom radius of gyration data for 5 proteins used to fit temperature-dependent (HPS-T) model.

$\mathrm{T_{ref}}$ a free parameter, and introducing two additional free parameters as the prefactor (α) and a shift along the temperature axis ($\mathrm{T_{shift}}$) into the function:

$$\lambda_i(T) = \lambda_{i,HPS} + \alpha \cdot [E_X(T - T_{\mathrm{shift}}) - E_X(T_{\mathrm{ref}} - T_{\mathrm{shift}})] \tag{5.10}$$

To find the optimal parameters for Eq. 5.10 with minimized deviation from the reference data, we needed a way to estimate the R_g from our CG model in an efficient manner. Toward this goal, we use a homopolymer-based predictor which can be used to quickly calculate the R_g for a specific protein sequence based on its chain length and average hydrophobicity (See Methods). Using a standard global optimization method [332], we minimize the difference between the predicted R_g and that from the reference data set. We optimize two different models, one where the free parameters are allowed to vary for all five amino acid types, yielding a 15-dimensional optimization, and the other where the free parameters are made universal for the different amino acid types, yielding a 3-dimensional optimization problem. Optimized parameters for both models are listed in Table 5.5. We find that keeping the 3-parameter model is sufficiently accurate to achieve good agreement with the reference set (Fig. 5.5). By allowing parameters for each amino acid type to vary independently in the 15-parameter model, we are able to match the empirical predictions to the reference data very well, while imposing the following constraints on the parameters: 0< α <2; 250<$\mathrm{T}_{\mathrm{ref}}$ <350; -100<$\mathrm{T}_{mathrmshift}$ <100, in order to search through the physically meaningful parameter space. While the empirical R_g predictions become more similar to the reference data, we find that results from simulations are actually less accurate than the 3-parameter model (Fig. 5.6). We believe this is due to the homopolymer-based predictor not accounting for the greater heterogeneity of interaction strengths within this model.

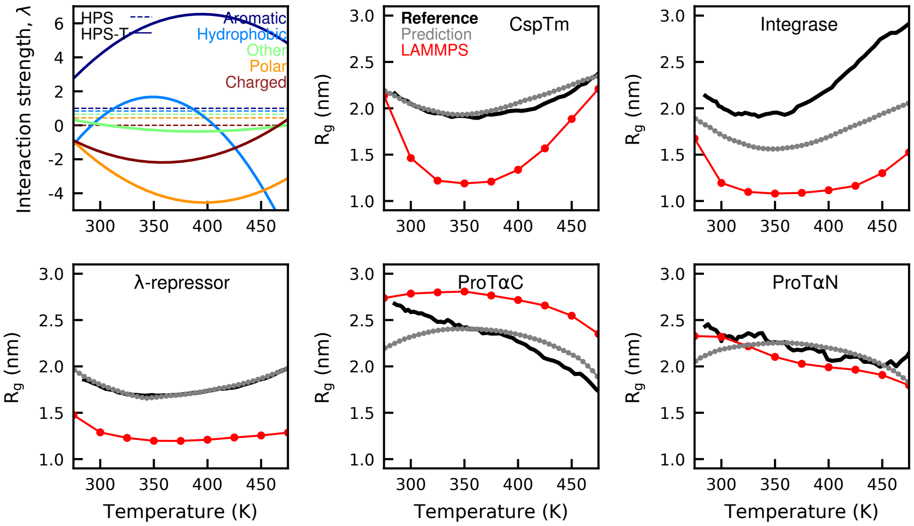

Figure 5.6: Model optimized to training data using independent free parameters for each of the five amino acid types, for a total of 15 free parameters as listed in Table 5.5

We also consider the use of a physics-based model from Dill, Alonso and Hutchinson[108] as suggested recently by Lin, Forman-Kay, and Chan[109]. Upon implementing this temperature-dependence into the HPS model, we find reasonable agreement with the training set from predictions and simulations for the first three sequences, but the collapse of ProTαC and ProTαN is not observed (Fig. 5.7), because this model does not capture the temperature-dependence of hydrophilic amino acid interactions. Thus we conclude that the use of the bioinformatics-based temperature dependence from van Dijk is advantageous in its ability to capture the temperature dependence of all types of amino acids. Future work, however, may focus on a more fundamental approach to estimate temperature dependence, and even pressure dependence of interactions from all-atom explicit solvent simulations[131].

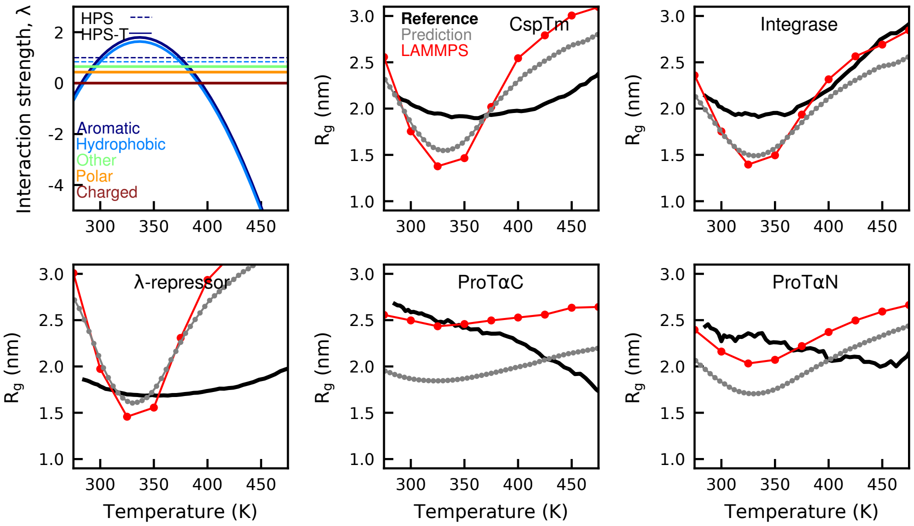

Figure 5.7: Application of Dill-Alonso-Hutchinson temperature-dependent model[108] into the hydrophobic scaling.

To demonstrate that the use of temperatures outside the realm of the experimentally realizable range in the reference data is not negatively affecting the model, we also conducted the optimization just using temperature points below 400K. We find that using this truncated temperature range results in a nearly identical set of parameters as the full reference data set. The use of a scaling parameter (eq. 5.5) to create the reference set may also modify the results due to the magnitude of R_g variation in response to change in temperature. To assess the effect of this, we created a separate reference data set for the 5 test proteins, by setting n_1 equal to 1. Optimizing to this reference set results in a similar model to the HPS-T model, with a somewhat weaker temperature dependence. We additionally attempt to fit directly to the experimental data to avoid having to combine with simulation data, but find that the even more limited range of temperatures does not account for the full shape of the temperature dependence of R_g. Thus, using only the limited experimental data available is likely not suitable for describing the thermoresponsive behavior of phase separating IDPs.

Table 5.5: Optimized parameters for 3-parameter and 15-parameter model.

Parameter	Hydrophobic	Aromatic	Other	Polar	Charged	3-parameter model
α (k_BT^{-1})	0.995	2.0	2.0	2.0	2.0	0.7836
$\mathrm{T_{ref}}$ (K)	250.0	308.8	250.0	253.6	250.0	296.7
$\mathrm{T_{shift}}$ (K)	97.07	48.72	100.0	100.0	49.24	61.97

The resulting temperature-dependent interaction strengths from the 3-parameter CG model are shown in Fig. 5.5A. The functional forms may then be simplified to

$$\lambda_{i,H}(T) = \lambda_{i,HPS} - 25.475 + 0.14537T - 0.00020059T^2, \tag{5.11a}$$

$$\lambda_{i,A}(T) = \lambda_{i,HPS} - 26.189 + 0.15034T - 0.00020920T^2, \tag{5.11b}$$

$$\lambda_{i,O}(T) = \lambda_{i,HPS} + 2.4580 - 0.014330T + 0.000020374T^2, \tag{5.11c}$$

$$\lambda_{i,P}(T) = \lambda_{i,HPS} + 11.795 + 0.067679T + 0.000094114T^2, \tag{5.11d}$$

$$\lambda_{i,C}(T) = \lambda_{i,HPS} + 9.6614 - 0.054260T + 0.000073126T^2, \tag{5.11e}$$

for the 3-parameter model, and

$$\lambda_{i,H}(T) = \lambda_{i,HPS} - 34.690 + 0.20242T - 0.000254463T^2, \tag{5.12a}$$

$$\lambda_{i,A}(T) = \lambda_{i,HPS} - 63.201 + 0.36955T - 0.000533392T^2, \tag{5.12b}$$

$$\lambda_{i,O}(T) = \lambda_{i,HPS} + 6.8820 - 0.040528T + 0.000052T^2, \tag{5.12c}$$

$$\lambda_{i,P}(T) = \lambda_{i,HPS} + 32.994 - 0.19100T + 0.0002402T^2, \tag{5.12d}$$

$$\lambda_{i,C}(T) = \lambda_{i,HPS} + 21.768 - 0.13373T + 0.00018663T^2, \tag{5.12e}$$

for the 15-parameter model. Hereafter, we refer to this temperature-dependent hydrophobicity-based model as the HPS-T model. Given our previous work, intramolecular interactions driving single chain collapse and intermolecular interactions driving LLPS are fundamentally related[92], thus we expect this model should be sufficient to capture the thermoresponsive phase behavior of

IDPs.

5.3.2 Temperature-dependent solvent-mediated interactions can help distinguish between UCST and LCST proteins

Garcia-Quiroz and Chilkoti synthesized a large number of low-complexity IDPs mimicking the short, repetitive amino-acid motif characteristic of tropoelastins and resilins, with a highly diverse range of amino acid compositions[26]. Through this, they provided an excellent characterization of how amino acid composition can influence the thermoresponsive protein phase behavior [26]. They found that RLPs are generally composed of charged and polar amino acids, and show UCST behavior, while ELPs tend to contain more hydrophobic amino acids, and exhibit LCST behavior. Due to the large number of sequences spanning a wide range of amino acid compositions, and the direct observance of thermoresponsive phase behavior, this data set is ideal for testing the applicability of the HPS-T model. We classified the 39 sequences, termed QC sequences, into three groups: LCST, UCST, and no phase separation (Table A.1).

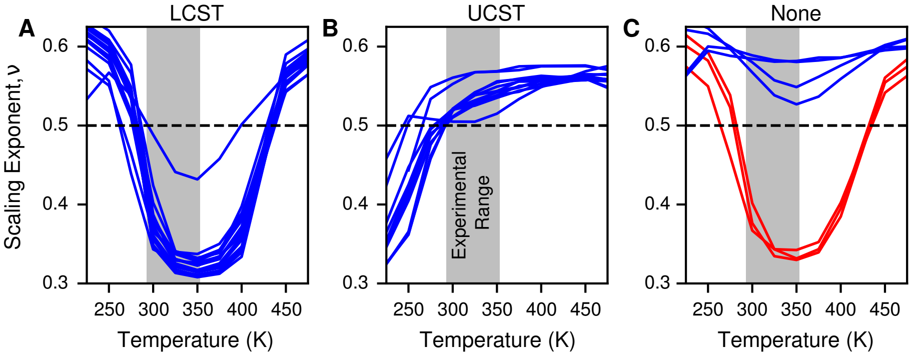

Figure 5.8: Experimental verification of the model with QC sequences. A) The simulated ν as a function of T for QC sequences with experimental LCST behavior. B) QC sequences with experimental UCST behavior. C) QC sequences without phase behavior in experiment. The red lines show QC3, 6 and 7 with crossing points to $\nu = 0.5$ in simulations. The grey bar shows the range of temperature scanned by the experiment[26].

Since it is impractical to conduct coexistence simulations for all 39 sequences in the QC data set, we take advantage of a recently suggested correlation between the critical temperature T_c (which separates the two-phase region from the single-phase region in the phase diagram) and the Θ-temperature (T_θ) (the temperature at which the polymer scaling exponent, ν, is equal to

0.5 [92]). For conditions where $\nu < 0.5$, the effective intrachain or interchain interactions are attractive causing chain collapse or phase separation, whereas ν values larger than 0.5 imply repulsive interactions are dominant, causing chain expansion and inability to phase separate. One can also approximately calculate T_c from the Boyle temperature (T_B), the temperature at which the osmotic second virial coefficient (B_{22}) goes to zero. We found these relationships to be non-model-specific as was applied to two different potential energy functions [92], and therefore should, in principle, be able to predict both UCST and LCST behavior from T_θ or T_B in the new HPS-T model.

In Fig. 5.8, we present the polymer scaling exponent (ν) for each of the QC sequences for a wide temperature range to determine what conditions will allow for phase separation, and to predict whether these will display UCST or LCST behavior. For the first set of QC sequences (LCST), we first observe chain collapse (decreasing ν) with increasing temperature, and a subsequent expansion (increasing ν) from our simulation, with the initial collapse occurring at the range of temperatures tested in experiment (Fig. 5.8A). For the second set (UCST), an initial chain expansion is followed by collapse highlighting UCST behavior within the experimental temperature range (Fig. 5.8B). Most of these sequences show a dual-response phase behavior with two crossing points, which is not observed in experiment. In general, the two T_θ values are quite far apart which would be difficult to observe in experiment without making other perturbations to the system, and thus would only observe the single phase transition we see near the experimental range.

For the third group of QC sequences which were shown to not phase separate in experiment, we find that the ν values for four of the seven proteins never decrease below 0.5, suggesting that these particular proteins are not expected to phase separate within the broad temperature range tested (Fig 5.8C). The three dissonant sequences are all in the set of proteins mimicking the content of elastin, for which simulation results predict LCST behavior at experimental conditions as predicted for the majority of the other proteins in that set (Fig. 5.8A). The simulation results are at odds with the experimentally documented behavior for these three proteins, which raises questions about the general validity of the HPS-T model despite its strong predictive capabilities for 36 out of 39 proteins. A careful look at these protein sequences highlights similarities with other sequences in the QC data set, some of which are nearly identical to QC3, QC6, and QC7 in composition. As these analogous sequences (QC2 $\sim$ QC3, QC4/5 $\equiv$ QC6, and QC9 $\sim$ QC7) show LCST behavior, the predictions of the model are not entirely unreasonable. Another possibility is that the experimental temperature range is not sufficiently broad to induce phase separation at relatively low concentrations, and our results may provide impetus to revisit these experiments in the future. We also note that use of the Dill-Alonso-Hutchinson model captures the sequences which undergo LCST, but not those with UCST, supporting our assertion that temperature-dependent interactions between polar amino acids must also be accounted for (Fig. 5.9).

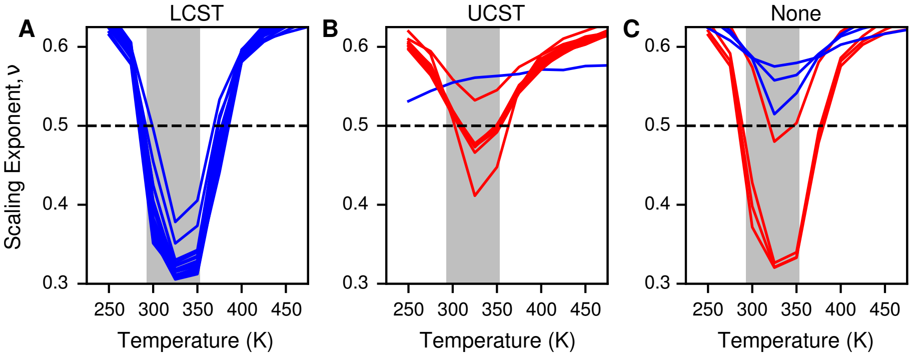

Figure 5.9: Simulation results of QC sequences using Dill-Alonso-Hutchinson model[108] for sequences showing A) LCST, B) UCST, and C) no phase separation from experimental results[26].

5.3.3 Reentrant phase behavior as a function of temperature

The qualitative agreement of the HPS-T model and experimental results indicates it is promising approach to directly study the LLPS of proteins undergoing UCST and LCST. It is therefore instructive to ask if the UCST versus LCST phase behavior predicted by changes in ν as a function of temperature due to solvent-mediated interactions is present in the thermodynamic phase diagram. The appearance of different phases as a function of temperature in a multi-protein simulation will also allow one to probe the differences in the molecular structure and dynamics directly from the simulated trajectories. We select a QC sequence from each of the first two groups (QC21 and QC37) and conduct slab coexistence simulations to obtain the thermodynamic phase diagram as well as two-chain simulations to determine B_{22} at a range of temperatures, and to estimate T_B following the same protocol as in previous work[92, 144].

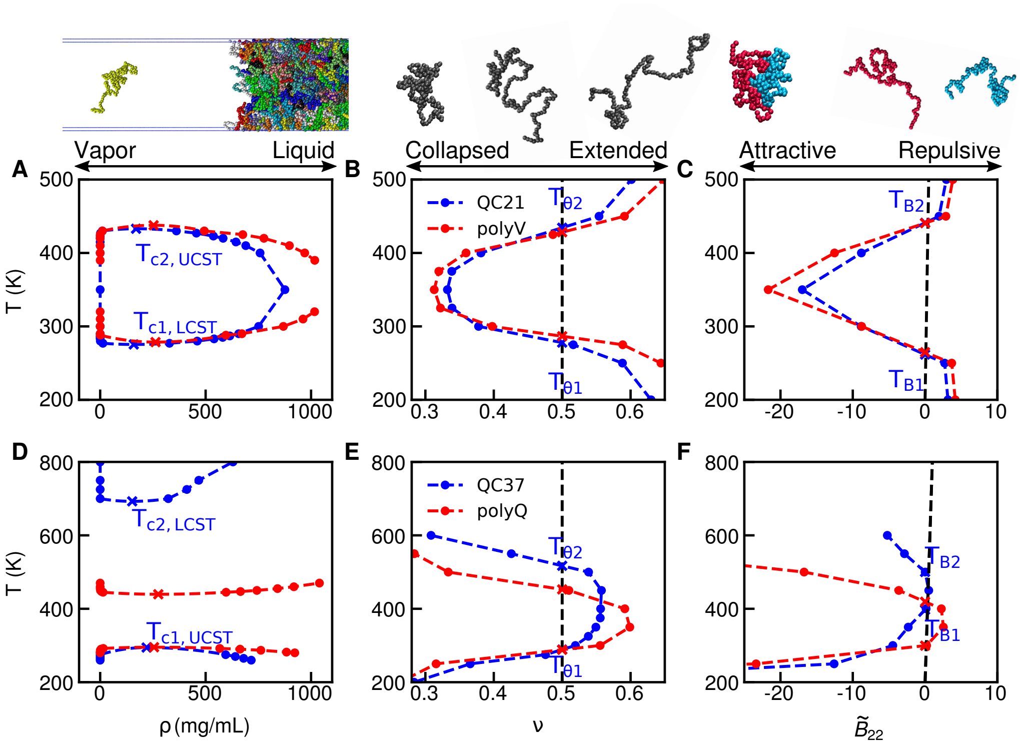

Figure 5.10: Dual-phase behavior of IDP sequences. A) Temperature-dependent ν, B) B_{22} and C) phase coexistence of a hydrophobic homopolymer (V_{50}) and an elastin-like LCST sequence from Garcia-Quiroz et al.[26]. D) Temperature-dependent ν, E) B_{22} and F) phase coexistence of a hydrophilic homopolymer (Q_{50}) and a resilin-like UCST sequence from Garcia-Quiroz et al.[26].

QC21 exhibits a dual-response phase behavior described by an LCST at 275K, and a UCST at 432K, with a region in between where LLPS is observed, having the shape of a closed-loop phase diagram (Fig. 5.10A)[101, 333, 334]. The closed-loop phase diagram is analogous to the predicted cold-denaturation of folded proteins which unfold at both extreme high and low temperatures[129]. This observed phase behavior is also qualitatively consistent with the collapse and expansion of a single protein chain with increasing temperature (Fig. 5.10B), as well as with the preference for intermolecular attraction ($B_{22} < 0$) and repulsion ($B_{22} > 0$) between two proteins as a function of temperature (Fig. 5.10C). Moreover, there is a good correspondence between the different transition temperatures that can be identified from Fig. 5.10A-C. This suggests that the previously proposed correlations [92] as well as the homopolymer predictor model used to fine-tune the CG model parameters can be accurate enough for future use.

QC37, on the other hand, phase separates at low temperatures and is miscible at temperatures above 294K. With further increase in temperature to a very high value (692K), which is not physically meaningful, the system shows a reentrant behavior by phase separating again into two phases (Fig. 5.10D). Such dual-responsive phase behavior has been observed experimentally for various RLPs such as Rec1[318] and An16 resilin[100] within temperature ranges accessible to experiment. The qualitative behavior observed from the other two transition temperatures based on protein collapse (Fig. 5.10E) and intermolecular interactions between a pair of protein molecules (Fig. 5.10F) is also consistent with this phase diagram. However, only the lower transition temperatures (T_{c1}, $T_{\theta 1}$, and T_{B1}) are in a quantitative agreement with each other, while the LCST (T_{c2}), is significantly higher than $T_{\theta 2}$ or T_{B2} (Fig. 5.10D-F).

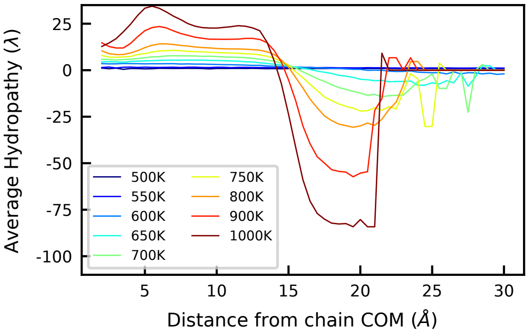

Figure 5.11: Average hydrophobicity of amino acids as a function of distance from the COM of the QC37 chain, calculated from single chain simulation.

A closer examination of the QC37 multi-protein system between temperatures $T_{\theta 2}$ and T_{c2} suggests that these proteins prefer to form intramolecular contacts (leading to collapsed globular conformations) as opposed to the intermolecular contacts required to stabilize a condensed protein-rich phase. A possible explanation for this behavior is the relative importance of enthalpy and entropy in the free energy of the system. We hypothesize that the entropic cost of incorporating protein chains into a condensed phase, which is not appropriately accounted for in a single chain simulation to estimate T_θ, becomes more important at higher temperatures. The system free energy is thus minimized through maximizing intramolecular contacts by forming collapsed globules, and maximizing the system entropy by keeping the proteins dispersed in a larger volume. If this is indeed the case, one would expect the proteins to adopt conformations such that hydrophobic residues are deeply buried inside and the protein surface is more hydrophilic and therefore less likely to form favorable contacts with other proteins. Indeed, we find that a single QC37 chain will isolate the more hydrophobic amino acids toward the center of the globule, while the more repulsive/hydrophilic residues occupy the surrounding region at high temperatures (Fig. 5.11). Considering the average and standard deviation of λ values for the amino acids in the QC sequences, we see that the variation between different amino acids is much higher for QC37 at $T_{\theta 2}$ than it is for $T_{\theta 1}$ or either T_θ of QC21 (Fig. 5.12), thus facilitating the collapse of more attractive amino acids to the center with repulsive residues at the exterior.

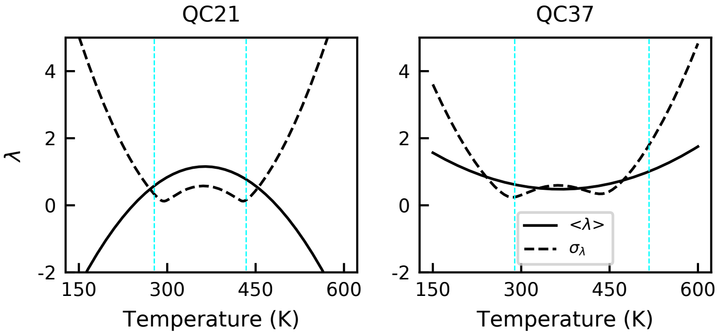

Figure 5.12: Average (solid line) and standard deviation (dashed line) of λ for all residues of QC21 and QC37 sequences over the range of temperatures tested. Blue dashed lines indicate single-chain T_θ values.

A simple test to determine whether the variation of attraction and repulsion within an IDP sequence is causing the unfavorability of the LCST phase transition is to simulate a simple homopolymeric protein expected to display a similarly-shaped phase diagram. Therefore, we conduct simulations of a poly-glutamine (Q_{50}) protein sequence to compute the phase diagram as well as the single-chain and two-chain properties as a function of temperature as shown in Fig. 5.10D-F. Our observed ν value for Q_{50} at room temperature is consistent with the expectation from the work of Singh and Lapidus on polyglutamine peptides[335], though we do not expect our model to be in perfect agreement with all available experimental data[336]. In this case, we find that all the transition temperatures are in quantitative agreement with each other and the LCST is also much lower than the heteropolymer QC37 sequence, no longer showing a large mismatch between T_{c2} and the other transition temperatures. This suggests that heterogeneity of the sequence, having large variance in attraction and repulsion within the sequence contribute to the breakdown of the general correlations between T_c, T_θ and T_B[92, 155, 282]. We also note that the phase behavior of a poly-valine (V_{50}) sequence expected to have a closed-loop phase diagram is quite similar to the heterogeneous sequence QC21 within this model (Fig. 5.10A-C).

The protein assemblies formed by QC37 at extremely high temperatures resemble solid aggregates. Interestingly, we find that the diffusion of protein chains within the condensed phase formed above the LCST are significantly slower than within the condensed phases formed below the UCST (Fig. 5.13). This behavior is reminiscent of experimental findings on RLPs which undergo similar reentrant phase transitions upon cooling and heating, and having slower recovery from the high-temperature LCST[100, 318]. It stands to reason that having few strong interaction sites within a sequence would lead to slower dynamics than having many weaker interaction sites. Thus, we postulate that the variation of attraction and repulsion within a sequence can be used to manipulate the dynamics within the condensed phase, which may be tuned by sequence, and temperature-dependent hydrophobicity.

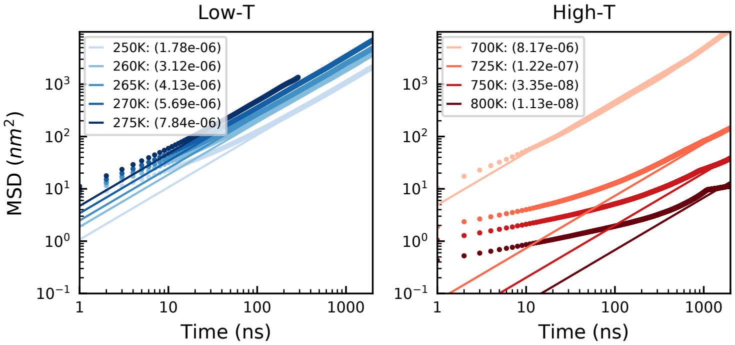

Figure 5.13: Mean squared displacement of QC37 within the condensed phase below UCST (left) and above LCST (right). Diffusion coefficients are reported in parentheses in units of cm^2/s.

5.3.4 Role of amino acid composition in the thermoresponsive behavior of disordered proteins

Given the success of our new HPS-T model in distinguishing UCST versus LCST sequences with the help of a simple predictor, we have a unique opportunity to identify the molecular determinants of the temperature-dependent phase behavior of IDPs. We scan a large number of sequences ($\approx$ 1 million) with the chain length the same as CspTm (66 amino acids) based on the relative abundance of each amino acid in the intrinsically disordered proteome[327] and compute ν for these proteins as a function of temperature (See Supporting Methods 1.6). We can use this information to infer the shape of the phase diagram regarding their transition temperatures, number of such transitions, and their type (UCST or LCST). Based on this analysis, we can classify IDP sequences into four groups: none ($\nu > 0.5$ always) without phase behaviors like QC sequences in Fig. 5.8C; single UCST with monotonically decreasing ν when increasing temperature; closed-loop with UCST higher than LCST (Fig. 5.10A); and hourglass with UCST lower than LCST (Fig. 5.10D).

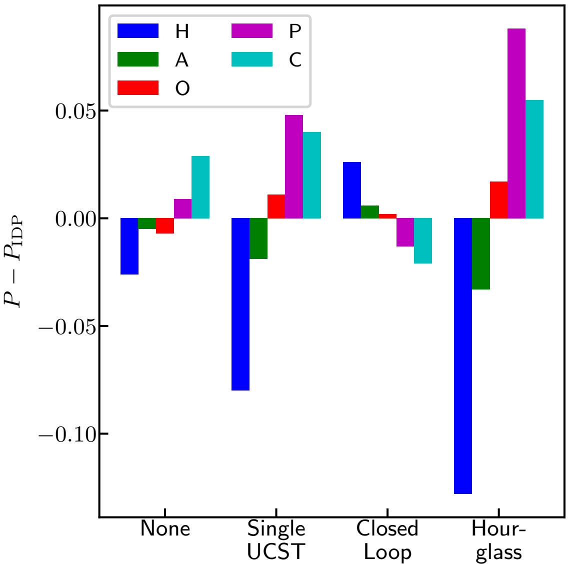

Figure 5.14: The difference between probabilities of H/A/O/P/C type amino acids (see Table 5.4) in sequences with a specific phase-diagram shape (labelled in x-axis) and the probabilities of those in a typical IDP sequence from a bioinformatics study[327]. The definition of the phase-diagram shape is shown in Table 5.1. Errors are shown in Table 5.1 and not noticeable in the figure.

To understand the role of specific amino acids in the marked preference for a given type of phase behavior, we compute the probability of their occurrence with respect to the probability of those amino acids for a typical IDP sequence from a bioinformatics study. As shown in Fig. 5.14 and Table S3, the amino acid probabilities in the types “closed-loop” and “none” are most similar to a typical IDP sequence. Whereas, an enhanced polar and charged amino acids content would be needed to observe single UCST or hourglass type behavior. These results present a path forward for the design of thermoresponsive materials with tunable properties by changing their amino acid content. However, we caution that the use of empirical predictions may not be directly applicable to all IDPs due to sequence-specific effects such as patterning of charges or hydrophobic regions. Rather, we hope this analysis serves as a demonstration of the possibilities, with more work to follow using direct MD simulations.

5.4 Conclusions

In this study, we provide a direct interrogation of the thermoresponsive phase behavior of IDPs through use of a novel coarse-grained model which explicitly represents the amino acid sequence, and accounts for the temperature-dependent solvent-mediated interactions of each type of amino acid. We validate the model using experimental and all-atom data on the R_g of several disordered proteins, as well as the thermoresponsive phase behavior of a large library of designed RLP and ELP sequences. The qualitative capture of the sequence-encoded phase behavior shows promise for the model to extend to the furthest reaches of the IDP sequence space when coupled with an empirical homopolymer-based predictor. From this, we learn that a typical IDP sequence will undergo phase separation with a closed-loop phase diagram, having LCST at the more physiological conditions. Sequences with an hourglass-shaped phase diagram generally contain more polar or charged residues than a typical IDP sequence.

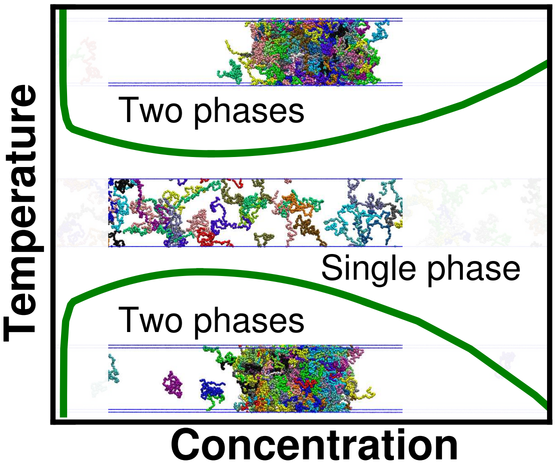

Figure 5.15: Houglass-shaped phase diagram and representative slab configurations.

Chapter 6

Atomistic Simulations of IDPs Involved in LLPS

6.1 Introduction

Up to this point, many of the simulations and models presented in this thesis have been informed or motivated by experimental results. In this chapter, I will discuss several cases where the opposite is true, and the model has provided unique insights, aiding in interpretation of experimental results, as well as motivating particular experiments to be conducted.

A minimalist system that which has been widely used to represent biomolecular phase separation in vitro would be a single purified protein in solution which is capable of phase separating under some condition[15, 27, 29, 34, 46, 56, 110, 337]. Such a simplified system is ideal for characterizing biophysical properties of LLPS as it allows for direct characterization of the changes to overall interaction strengths within a particular protein in response to stimuli[26, 101]. In addition, direct evidence may be collected on the presence (or lack thereof) of secondary structure[27, 29, 55, 338–340], microscopic dynamics of particular amino acids within the single protein[29, 33], and contacts occurring between different regions of the protein, or particular amino acids[33–35, 55].

To date, several explanations have been asserted to explain the driving forces behind LLPS of disordered proteins. Evidence suggests that LLPS of low-complexity, disorderd proteins is driven solely by weak multivalent interactions evenly distributed throughout the amino acid sequence, making it similar to a homopolymer[10, 29, 33]. Such an assumption is useful as it allows for use of well-established polymer physics models such as Flory-Huggins[271, 272] in order to model phase behavior of IDPs[27, 95, 144]. This also allows for phase separation to be predicted through use of mean-field theory[95] or through determination of the Flory scaling exponent[283] and θ-temperature [92] as done previously for homopolymers of infinite length[155, 282].

To observe whether or not this model is valid for phase separation of disordered proteins, it is imperative to obtain high-resolution information on the structural organization, and intermolecular contacts occurring within a condensate. Such a task is daunting as experimental methods generally lack either sufficient spatiotemporal resolution, or the dynamic nature of the liquid-like proteinaceous assemblies[341, 342]. NMR spectroscopy, in principle, may overcome both of these obstacles, but will return ensemble-averaged information, and not give information on individual microstates observed[343, 344]. Simulations may also overcome these obstacles by yielding atomic-resolution information on individual microstates that occur[53, 55], but suffer from imperfect potential energy functions, and difficulty sampling long timescales and the large configurational space of an IDP[345]. However, with the use of state-of-the-art potential energy functions[136] and enhanced sampling methods[218, 219, 346, 347], one may accurately reproduce structural information of many disordered and unfolded proteins using conventional supercomputing resources.

Here we demonstrate how atomic-resolution simulations of IDPs involved in biomolecular phase separation can highlight the different modes of interaction that occur withing biomolecular condensates and membraneless organelles. Using simulations of two chains, we observe the intermolecular interactions that occur between two proteins, and determine how amino acid composition may alter the modes of interaction that drive self-association and phase separation.

6.2 Simulation and Analysis Methods

6.2.1 PTWTE and PTWTE-MetaD Simulations

Atomic resolution simulations were conducted using the GROMACS software package[212], with the PLUMED software plugin[348, 349] for simulations involving metadynamics (MetaD) and well-tempered ensemble (WTE) simulations[219, 347]. All proteins were simulated as truncated peptide sequences of 40 residues in order to ensure sufficient sampling of the conformational space and that the simulations is computationally tractable. The use of a truncated sequence may still provide relevant structural information in the case of TDP-43 where the transiently structured region is only 25 residues[55], and useful information on contacts for sequences such as FUS and hnRNPA2 which are low complexity (LC), having a small selection of amino acid types[120, 228, 238].

hnRNPA2 variants were simulated as a single chain in water, while FUS variants were simulated as two chains in water. Simulations were carried out using state-of-the-art protein force fields, Amber ff03ws and Amber ff99SBws[136] both with tip4p/2005 water[214] (See Table 6.1). All simulations were conducted in a truncated octahedral box with periodic boundary conditions, and sufficiently large that protein chains are unlikely to interact with their periodic images. Single chain simulations were conducted using replica exchange molecular dynamics (REMD) coupled with a well-tempered ensemble (WTE) or parallel tempering in the well-tempered ensemble (PT-WTE) in order to accelerate the convergence of the equilibrium conformational ensemble[220, 350]. Two-chain simulations were also run using PT-WTE with an additional MetaD bias on the number of intermolecular contacts to enhance sampling of binding and unbinding events. The temperature list was selected based on a geometric function to obtain even exchange probabilities between adjacent replicas[351]. Simulations were run in the NVT ensemble with temperature kept constant using a Langevin thermostat. For all simulations, initial configurations were randomly selected from a short equilibrium simulation at 300K, and then further equilibrated at different temperatures, those used for replica exchange.

Table 6.1: List of simulations conducted for work presented in this chapter.

Protein	Force Field	Method	Simulation Time (ns)
hnRNPA2$_{190-233}$	Amber ff99SBws	PT-WTE	150
hnRNPA2$_{190-233}$ (ADMA)	Amber ff99SBws	PT-WTE	150
FUS$_{120-163}$ WT (x2)	Amber ff03ws	PT-WTE-MetaD	250

Table 6.2: Heavy atoms from each amino acid with absolute net charge less than 0.25, which are considered as contributing to hydrophobic contacts.

Ala	Cα, Cβ	Cys	Cα, Cβ
Asp	Cα, Cβ	Glu	Cα, Cβ, Cγ
Phe	Cα, Cβ, Cγ, C$\delta_{1,2}$, C$\epsilon_{1,2}$, Cζ	Gly	Cα
His	Cα, Cβ, Cγ, Cδ_2	Ile	Cα, Cβ, C$\gamma_{1,2}$, Cδ_1
Lys	Cα, Cβ, Cγ, Cδ, Cϵ	Leu	Cα, Cβ, Cγ, C$\delta_{1,2}$
Met	Cα, Cβ, Cγ, Cϵ	Asn	Cα, Cβ
Pro	N, Cα, Cβ, Cγ, Cδ	Gln	Cα, Cβ, Cγ
Arg	Cα, Cβ, Cγ, Cδ	Ser	Cα, Cβ
Thr	Cα, Cβ, Cγ_2	Val	Cα, Cβ, C$\gamma_{1,2}$
Trp	Cα, Cβ, Cγ, C$\delta_{1,2}$, C$\epsilon_{2,3}$, C$\zeta_{2,3}$, Cη_2	Tyr	Cα, Cβ, Cγ, C$\delta_{1,2}$, C$\epsilon_{1,2}$, Cζ

6.2.2 Contact Mode Definitions

Coarsely-defined contacts Similar to the coarse-grained contact definition, this is simply based on distance between atoms. To determine whether a pair of residues are in contact, the distance between all atom pairs are calculated, and if any is less than 6 $\AA$, the two residues are considered to be in contact. This definition differs from the other definitions in that it is more general, and largely meant to identify what residues are nearby, similar to a distance matrix.

Hydrophobic contacts For hydrophobic contacts, we calculate contacts between any two heavy atoms (non-hydrogen) having an absolute net charge greater than 0.25 according to the atomistic OPLS force field[352] as done for the Kapcha and Rossky hydrophobicity scale[252] (Table 6.2). A contact is then defined as any two such atoms being within 6 $\AA$ of each other.

Hydrogen Bonds Hydrogen bonds are simply defined by donors and acceptors being within a cutoff radius of 0.35 nm and having an angle less than 30°[353].

π-π and sp^2-hybridized interactions The definition used for π-π interactions is adapted from the work of Vernon et al.[72] who consider all atoms in sp^2-hybridized groups as being capable of forming planar π-π interactions (Table 6.3). The definition follows three particular criteria. First, at least two atoms from each of two sp^2-hybridized groups must be within 4.9 $\AA$ of each other.

Then the π faces must be calculated, which are parallel to the sp^2 group, and 1.7 $\mathring{A}$ away along the normal vector in either direction. Two of the points on the π face from each residue must then be within 1.5 $\mathring{A}$ of those on the other residue. Finally, if the first two criteria are satisfied, the angle between the normal vectors of the two planar sp^2 groups is calculated, and if the dot product of the unit vectors is ≥ 0.8 ($\theta \leq 36.9°$), it is considered as a planar π-π contact.

Table 6.3: Atoms involved in π-π interactions from backbone and amino acid side chains. *backbone π group contains nitrogen atom from the next residue (res i+1) in the sequence.

BB	$C\alpha$, C, O, N*	Asp	$C\beta$, $C\gamma$, $O\delta_{1,2}$
Glu	$C\gamma$, $C\delta$, $O\epsilon_{1,2}$	Phe	$C\beta$, $C\gamma$, $C\delta_{1,2}$, $C\epsilon_{1,2}$, $C\zeta$
His	$C\beta$, $C\gamma$, $N\delta_1$, $C\delta_2$, $C\epsilon_1$, $N\epsilon_2$	Asn	$C\beta$, $C\gamma$, $O\delta_1$, $N\delta_2$
Gln	$C\gamma$, $C\delta$, $O\epsilon_1$, $N\epsilon_2$	Arg	$N\epsilon$, $C\zeta$, $N\eta_{1,2}$
Trp	$C\beta$, $C\gamma$, $C\delta_{1,2}$, $N\epsilon_1$, $C\epsilon_{2,3}$, $C\zeta_{2,3}$, $C\eta_2$	Tyr	$C\beta$, $C\gamma$, $C\delta_{1,2}$, $C\epsilon_{1,2}$, $C\zeta$

6.3 Results

6.3.1 hnRNPA2 Compaction is Disrupted by Arginine Methylation[1]

The RNA-binding protein, heteronuclear ribonucleoprotein A2 (hnRNPA2) is a multidomain protein containing a disordered, low-complexity (LC) domain. This LC domain is the site of several disease-related mutations which promote aggregation into pathological aggregates[10, 23, 288]. In addition, the LC domain contains several RGG motifs, which are known to be targeted by protein arginine methyltransferase 1 (PRMT1) to reversibly convert arginine into asymmetric dimethylarginine (ADMA)[354]. This change to sidechain chemistry is known to alter interactions of the amino acid side chain, though the overall charge state of the protein is unchanged[87].

Experiments show that dimethylation of arginine residues greatly reduces the LLPS propensity of hnRNPA2 LC[35]. In this work, we use simulations of a truncated region of hnRNPA2 LC from residue 190-233 (hnRNPA2$_{190-233}$) to observe the changes to behavior resulting from dimethylated arginine residues. The region we simulate contains four arginine residues, three of which are in RGG motifs, and shown to be dimethylated.

From simulations of the 44-residue fragment, hnRNPA2$_{190-233}$, we find that the variant with

[1] Adapted from ref.[35]

three dimethylated arginine residues is considerably less collapsed than the unmodified variant (Fig. 6.1). Since the chain dimensions of an IDP are related to the intramolecular interactions, and can be considered to yield evidence into the strength of interactions driving LLPs of that protein[92] (See Chapter 4), this result is consistent with experimental observations that dimethylation of arginine residues reduces LLPS propensity. It is curious, however, the reason why dimethylation of arginine residues would result in weaker intramolecular interactions, as the hydrophobicity of arginine would likely increase with the addition of two methyl groups.

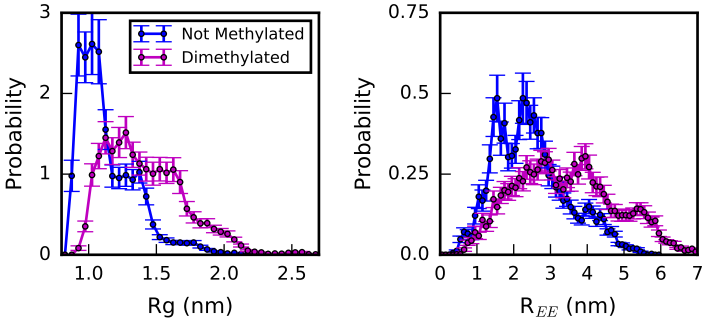

Figure 6.1: Single chain dimensions of hnRNPA2$_{190-233}$ from all-atom explicit solvent simulation. Radius of gyration (R_g) and end-to-end distance (R$_{EE}$) of hnRNPA2$_{190-233}$ with and without three arginine residues within RGG motifs dimethylated.

To directly interogate the effects of ADMA on single chain dimensions, we look at the particular interactions occurring between different residues within the sequence. Most contacts occur when the chain is collapsed, so we isolated frames where R_g is less than 1.2 nm, and visualize the intramolecular contact probabilities for each residue pair (Fig. 6.2). Another advantage of isolating only collapsed frames is that the dimethylated variant of hnRNPA2$_{190-233}$ would have far lower contact probability, simply due to the fact that is is much less collapsed than the unmethylated variant.

Interestingly, the contact probability goes down considerably for many amino acids upon dimethylation of the three arginine residues (Fig. 6.2). Indeed, when considering contacts between arginine and all other residues, it becomes clear that dimethylation of arginine considerably disrupts contacts with other residues, particularly aromatic residues as well as oppositely-charged anionic residues (Fig. 6.3). Upon dimethylation, there is a clear reduction of contacts between the

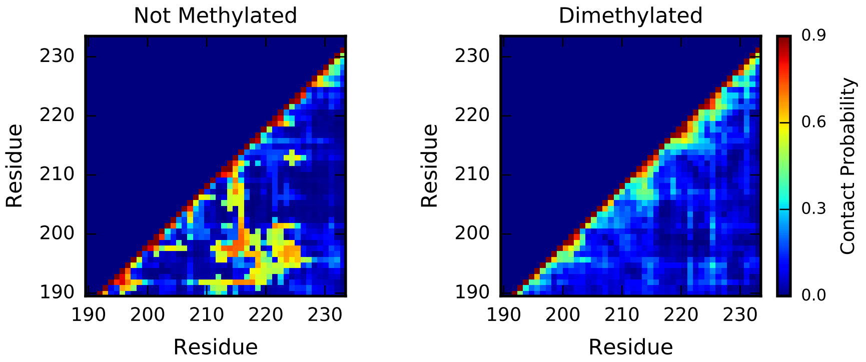

Figure 6.2: Intramolecular contacts within hnRNPA2$_{190-233}$ chains with and without dimethylated arginine residues, considering only frames where $R_g \leq 1.2$nm.

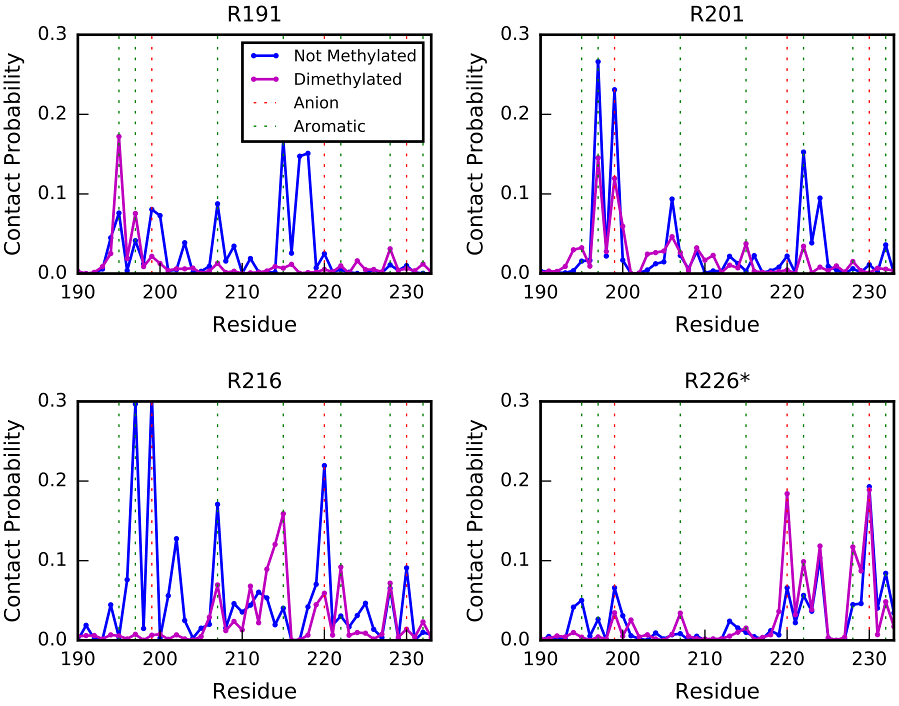

Figure 6.3: Intramolecular contacts between each arginine residue and all other residues within the hnRNPA2$_{190-233}$ sequence. *Only the first three arginine residues shown are dimethylated in the dimethylated protein, and R226 is the one that is not.

arginine residues and most other amino acids within the sequence. It has also been suggested that cation-π interactions between arginine and tyrosine are important to LLPS[71]. We also see evidence of this in that interactions with aromatic residues seem to be considerably less favorable for

ADMA than for arginine (Fig. 6.3). However, since ADMA is still charged, it is likely still capable of cation-π interactions, though the reduced contact probability could still be due to disruption of π-stacking by the added methyl groups. These results together suggest that dimethylation of arginine, while increasing hydrophobicity of the amino acid, reduces the ability of the side chains to form interactions with many of the other amino acids within the hnRNPA2$_{190-233}$ sequence.

Interactions differ the least between the methylated and unmethylated variants for arginine 226, which is the one arginine residue within the sequence that was not dimethylated as it is not within an RGG motif. This lack of change in the contacts of residues 226 suggests that overall configurations and contacts within the protein may not change appreciably due to the overall change in compaction, and that changes to contact probability may only occur locally at the post-translational modification sites.

6.3.2 FUS Self-Association is Driven by Diverse Interactions[2]

FUS is another RNA-binding protein containing a disordered, low-complexity domain that is necessary and sufficient to undergo LLPS in vitro. FUS LC has been comprehensively characterized in vitro using sophisticated NMR techniques to look at its secondary structure in certain conditions[338] or lack thereof in other conditions[29, 30], short-timescale dynamics[29], long-range intermolecular contacts[30, 33], and interactions between particular amino acid types[30] all within condensates composed of FUS LC.

This detailed characterization, however, is still insufficient to directly observe the modes of interaction that occur within a condensed phase of FUS. By using all-atom explicit solvent simulations of two chains of residues 120-163 of FUS (FUS$_{120-163}$), we highlight the different interaction modes that contribute to its self-association. The use of a truncated fragment of FUS LC is approprtiate here as it is a low-complexity sequence, making a fragment of the sequence very similar in composition to the full LC domain. Two chain simulations allow for us to observe the intermolecular interactions occurring between two chains without having to consider the local effects accompanying intramolecular interactions.

We find that intermolecular hydrogen bonds, nonpolar interactions, and π-π interactions are all

[2] Adapted from ref.[30]

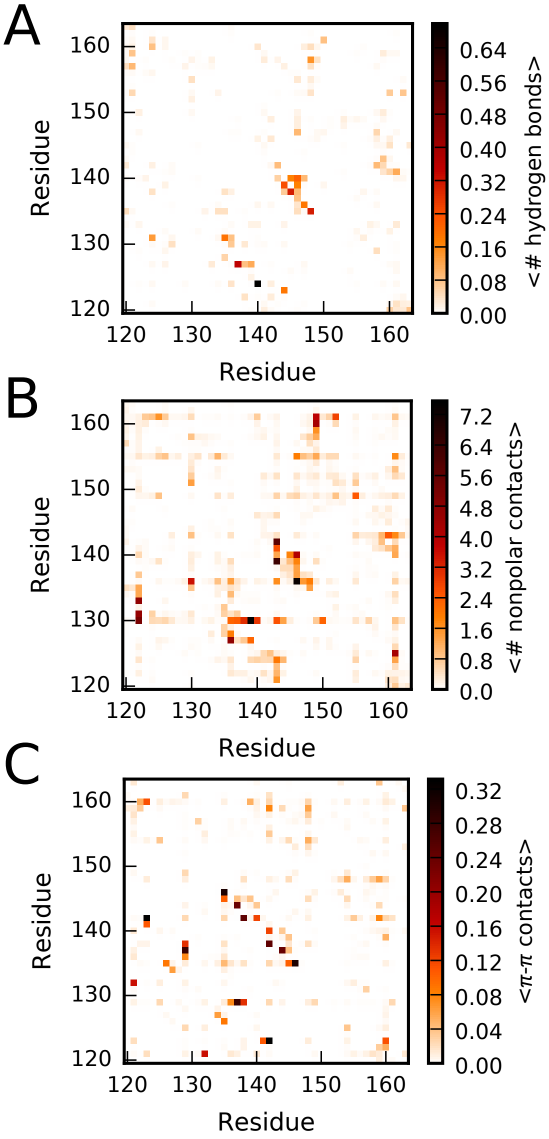

Figure 6.4: Per-residue intermolecular contact maps of 2-chain FUS simulation showing different interaction modes. Average number of A) hydrogen bonds B) nonpolar atom-atom interactions, and C) π-π interactions between all residue pairs within the $\mathrm{FUS}_{120-163}$ sequence.

distributed throughout the sequence relatively evenly (Fig. 6.4) as has been suggested previously for the FUS LC sequence[33]. By collapsing the contact maps to a single dimension, one can also compare the contributions of backbone or sidechain groups to the different modes of interaction (Fig. 6.5). It is clear that in addition to most amino acids contributing to interactions through at least one interaction mode, each interaction mode may be contributed to by either side chain, or backbone atoms (Fig. 6.5).

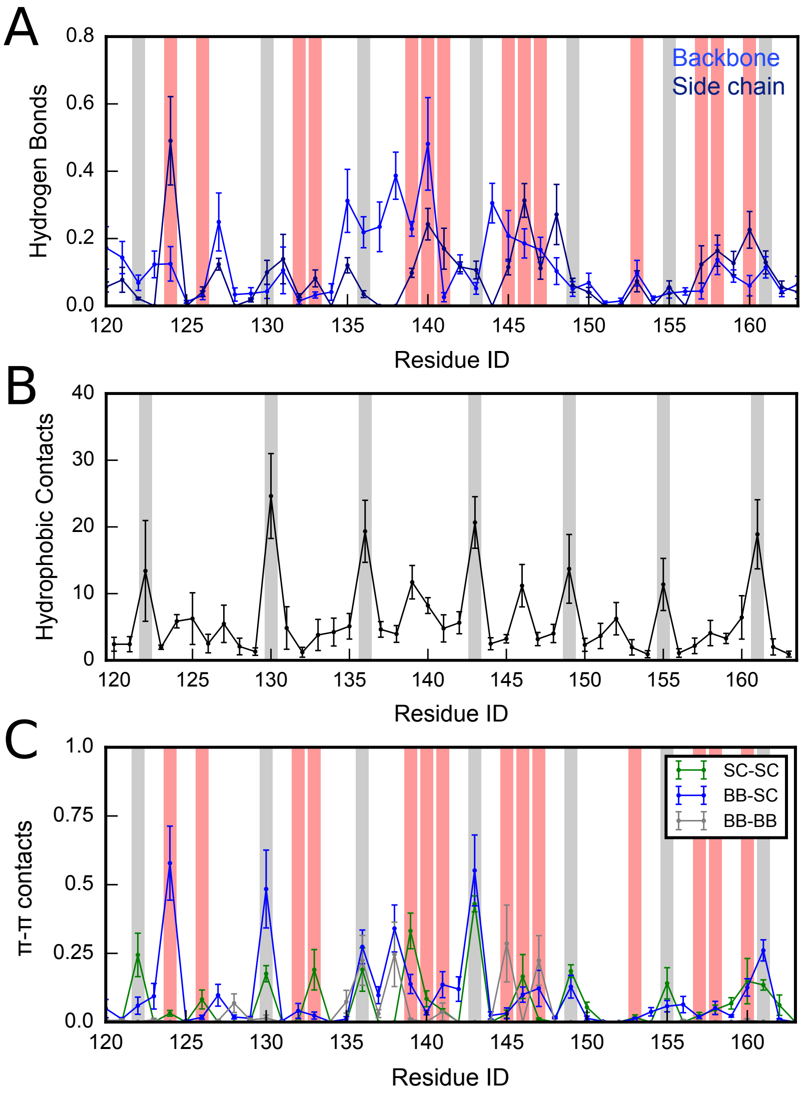

Figure 6.5: Per-residue intermolecular contacts between two $FUS_{120-163}$ chains. A) Average number of hydrogen bonds formed by each amino acid along the sequence. Number of hydrogen bonds has been decoupled to show contributions of backbone and sidechain to interactions. B) Average number of intermolecular nonpolar interactions between atoms within each residue of the protein. C) Average number of intermolecular π-π interactions formed by each amino acid along the sequence. Backbone-backbone, sidechain-sidechain and backbone-sidechain interactions are all shown separately to highlight contribution of different groups to formation of contacts. Tyrosine residues are hihilghted in gray, and glutamine in red.

Hydrogen bonds are highly prevalent among glutamine residues, but occur within most amino acid types. It is clear that the number of nonpolar contacts is by far the greatest at the tyrosine residues, which is due to the larger number of nonpolar atoms within the amino acid compared to the other amino acids within the sequence. π-π interactions are also very prevalent within most of the amino acid types, particularly tyrosine and glutamine. It is also important to note that the use of simulations is capable of reasonably capturing π-π interactions, as the nubmers observed here are comparable to recent studies using bioinformatics to quantify the probability of π-π contact formation between different amino acid types[72].

6.4 Conclusions

From these studies we have looked at simplified systems of two short IDP chains and characterized their equilibrium ensembles. We have found that in the case of the hnRNPA2$_{190-233}$ fragment, dimethylated arginine has a considerably lower propensity to form contacts with most aromatic residues, particularly tyrosines, resulting in more extended configurations in bulk solution. It is likely that modifications to proteins which interrupt interactions, not only result in increased chain dimensions, but also a reduction in phase separation propensity[35]. Such a relationship will be discussed further in chapter 4. In simulations of the truncated FUS$_{120-163}$ fragment, we provide a comprehensive characterization of the different interactions occurring and driving self-association (Fig. 6.4 and 6.5). It is likely that such analysis may be useful in determining the interactions contributing to LLPS, as they are intermolecular interactions. We discuss three major modes of interaction that particularly contribute to FUS$_{120-163}$ self-association, namely hydrogen bonding, hydrophobic interactions, and planar π-π interactions. Particular residues of interest are the glutamine, and tyrosine residues are both highly prominent within the sequence, and are enabled in all three interaction modes. We finally proposed a model of phase separation that is driven by weak multivalent interactions between many amino acid types, all of which interact promiscuously with many other partners[30]. Fig. 6.6 shows a schematic of an atomic-resolution condensate and its interface with a surrounding aqueous environment. The atomic configuration was generated from an atomic-resolution simulations of a “slab” of FUS LC chains with explicit solvent and ions. Results from such simulations are preliminary, and thus are not discussed in this thesis, however, the visualization is useful for demonstrating the presence of many different weak interactions that occur simultaneously within a FUS droplet.

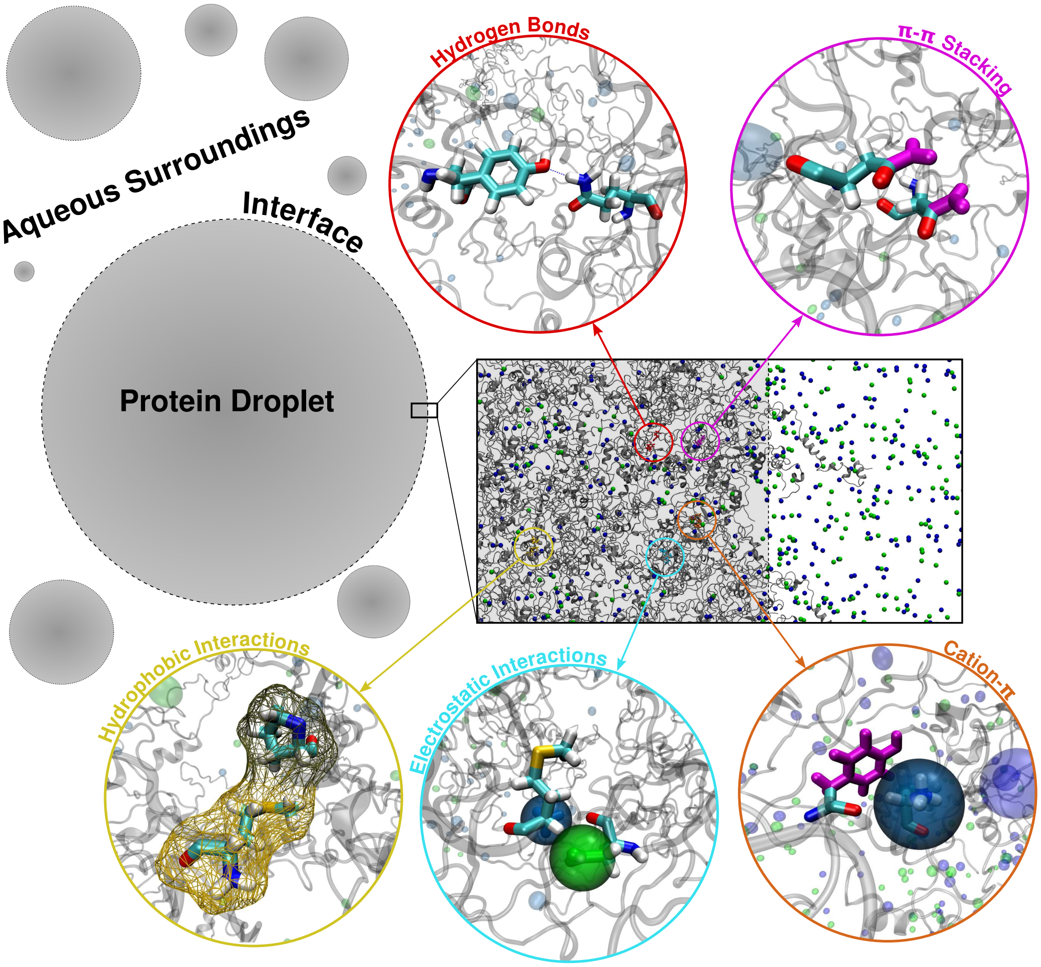

Figure 6.6: Many different interaction modes occur within a condensed phase of FUS LC, and contribute to the self-association of protein chains, stabilizing the highly concentrated, yet liquid-like phase. Such interactions may be present in condensates of other proteins at different levels based on differing amino acid composition and arrangement. Reproduced from article currently under review.

Bibliography

[1] Wagner, R. *Müllers Archiv Anat Physiol Wissenschaft Med* **1835**, *268*, 373–7.

[2] Brangwynne, C. P.; Eckmann, C. R.; Courson, D. S.; Rybarska, A.; Hoege, C.; Gharakhani, J.; Jülicher, F.; Hyman, A. A. *Science* **2009**, *324*, 1729–1732.

[3] Hyman, A. A.; Weber, C. A.; Jülicher, F. *Annual Review of Cell and Developmental Biology* **2014**, *31*, 39–58.

[4] Brangwynne, C. P.; Mitchison, T. J.; Hyman, A. A. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 4334–4339.

[5] Weber, S. C.; Brangwynne, C. P. *Cell* **2012**, *149*, 1188–1191.

[6] Shin, Y.; Brangwynne, C. P. *Science* **2017**, *357*, eaaf4382.

[7] Holehouse, A. S.; Pappu, R. V. *Biochemistry* **2018**, *57*, 2415–2423.

[8] Alberti, S.; Gladfelter, A.; Mittag, T. *Cell* **2019**, *176*, 419–434.

[9] Li, Y. R.; King, O. D.; Shorter, J.; Gitler, A. D. *J Cell Biol* **2013**, *201*, 361–372.

[10] Malinovska, L.; Kroschwald, S.; Alberti, S. *Biochimica et Biophysica Acta - Proteins and Proteomics* **2013**, *1834*, 918–931.

[11] Patel, A. et al. *Cell* **2015**, *162*, 1066–1077.

[12] Wegmann, S. et al. *EMBO J.* **2018**, e98049.

[13] Riback, J. A.; Katanski, C. D.; Kear-Scott, J. L.; Pilipenko, E. V.; Rojek, A. E.; Sosnick, T. R.; Drummond, D. A. *Cell* **2017**, *168*, 1028–1040.

[14] Wei, M.-T.; Elbaum-Garfinkle, S.; Holehouse, A. S.; Chen, C. C.-H.; Feric, M.; Arnold, C. B.; Priestley, R. D.; Pappu, R. V.; Brangwynne, C. P. *Nat. Chem.* **2017**, *9*, 1118.

[15] Schuster, B. S.; Reed, E. H.; Parthasarathy, R.; Jahnke, C. N.; Caldwell, R. M.; Bermudez, J. G.; Ramage, H.; Good, M. C.; Hammer, D. A. *Nat. Commun.* **2018**, *9*, 2985.

[16] Darling, A. L.; Liu, Y.; Oldfield, C. J.; Uversky, V. N. *Proteomics* **2018**, *18*, 1700193.

[17] Van Der Lee, R. et al. *Chemical reviews* **2014**, *114*, 6589–6631.

[18] Kriwacki, R. W.; Hengst, L.; Tennant, L.; Reed, S. I.; Wright, P. E. *Proc. Natl. Acad. Sci. U. S. A.* **1996**, *93*, 11504–11509.

[19] Uversky, V. N.; Gillespie, J. R.; Fink, A. L. *Proteins* **2000**, *41*, 415–427.

[20] Dunker, A. K.; Romero, P.; Obradovic, Z.; Garner, E. C.; Brown, C. J. *Genome Inf.* **2000**, *11*, 161–171.

[21] Tompa, P. *Trends Biochem. Sci.* **2002**, *27*, 527–533.

[22] Muiznieks, L. D.; Sharpe, S.; Pomès, R.; Keeley, F. W. *Journal of molecular biology* **2018**, *430*, 4741–4753.

[23] Kim, H. J. et al. *Nature* **2013**, *495*, 467.

[24] Mittag, T.; Parker, R. *Journal of Molecular Biology* **2018**, *430*, 4636–4649.

[25] Uversky, V. N.; Kuznetsova, I. M.; Turoverov, K. K.; Zaslavsky, B. *FEBS Lett.* **2015**, *589*, 15–22.

[26] Quiroz, F. G.; Chilkoti, A. *Nat. Mater.* **2015**, *14*, 1164.

[27] Brady, J. P.; Farber, P. J.; Sekhar, A.; Lin, Y.-H.; Huang, R.; Bah, A.; Nott, T. J.; Chan, H. S.; Baldwin, A. J.; Forman-Kay, J. D.; Kay, L. E. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, E8194–E8203.

[28] Majumdar, A.; Dogra, P.; Maity, S.; Mukhopadhyay, S. *The Journal of Physical Chemistry Letters* **2019**,

[29] Burke, K. A.; Janke, A. M.; Rhine, C. L.; Fawzi, N. L. *Mol. Cell* **2015**, *60*, 231–241.

[30] Murthy, A. C.; Dignon, G. L.; Kan, Y.; Zerze, G. H.; Parekh, S. H.; Mittal, J.; Fawzi, N. L. *Nature Structural and Molecular Biology* **2019**,

[31] Galea, C. A.; Wang, Y.; Sivakolundu, S. G.; Kriwacki, R. W. *Biochemistry* **2008**, *47*, 7598–7609.

[32] Wang, J. T.; Smith, J.; Chen, B.-C.; Schmidt, H.; Rasoloson, D.; Paix, A.; Lambrus, B. G.; Calidas, D.; Betzig, E.; Seydoux, G. *eLife* **2014**, *3*, e04591.

[33] Monahan, Z.; Ryan, V. H.; Janke, A. M.; Burke, K. A.; Zerze, G. H.; O'Meally, R.; Dignon, G. L.; Conicella, A. E.; Zheng, W.; Best, R. B.; Cole, R. N.; Mittal, J.; Shewmaker, F. *EMBO J.* **2017**, e201696394.

[34] Larson, A. G.; Elnatan, D.; Keenen, M. M.; Trnka, M. J.; Johnston, J. B.; Burlingame, A. L.; Agard, D. A.; Redding, S.; Narlikar, G. J. *Nature* **2017**, *547*, 236.

[35] Ryan, V. H.; Dignon, G. L.; Zerze, G. H.; Chabata, C. V.; Silva, R.; Conicella, A. E.; Amaya, J.; Burke, K. A.; Mittal, J.; Fawzi, N. L. *Mol. Cell* **2018**, *39*, 465–479.

[36] Saito, M.; Hess, D.; Eglinger, J.; Fritsch, A. W.; Kreysing, M.; Weinert, B. T.; Choudhary, C.; Matthias, P. *Nature chemical biology* **2019**, *15*, 51.

[37] Lin, Y.-H.; Chan, H. S. *Biophys. J.* **2017**, *112*, 2043–2046.

[38] Hofmann, H.; Soranno, A.; Borgia, A.; Gast, K.; Nettels, D.; Schuler, B. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 16155–16160.

[39] Riback, J. A.; Bowman, M. A.; Zmyslowski, A. M.; Knoverek, C. R.; Jumper, J. M.; Hinshaw, J. R.; Kaye, E. B.; Freed, K. F.; Clark, P. L.; Sosnick, T. R. *Science* **2017**, *358*, 238–241.

[40] Zheng, W.; Zerze, G. H.; Borgia, A.; Mittal, J.; Schuler, B.; Best, R. B. *J. Chem. Phys.* **2018**, *148*, 123329.

[41] Marsh, J. A.; Forman-Kay, J. D. *Biophysical journal* **2010**, *98*, 2383–2390.

[42] Kroschwald, S.; Munder, M. C.; Maharana, S.; Franzmann, T. M.; Richter, D.; Ruer, M.; Hyman, A. A.; Alberti, S. *Cell Rep.* **2018**, *23*, 3327–3339.

[43] Wang, A.; Conicella, A. E.; Schmidt, H. B.; Martin, E. W.; Rhoads, S. N.; Reeb, A. N.; Nourse, A.; Montero, D. R.; Ryan, V. H.; Rohatgi, R.; Shewmaker, F.; Naik, M. T.; Mittag, T.; Ayala, Y. M.; Fawzi, N. L. *The EMBO journal* **2018**, e97452.

[44] Mitrea, D. M.; Cika, J. A.; Stanley, C. B.; Nourse, A.; Onuchic, P. L.; Banerjee, P. R.; Phillips, A. H.; Park, C.-G.; Deniz, A. A.; Kriwacki, R. W. *Nature communications* **2018**, *9*, 842.

[45] An, S.; Kumar, R.; Sheets, E. D.; Benkovic, S. J. *Science* **2008**, *320*, 103–106.

[46] Sabari, B. R. et al. *Science* **2018**, *361*, eaar3958.

[47] Asherie, N. *Methods* **2004**, *34*, 266–272.

[48] Braun, M. K.; Wolf, M.; Matsarskaia, O.; Da Vela, S.; Roosen-Runge, F.; Sztucki, M.; Roth, R.; Zhang, F.; Schreiber, F. *J. Phys. Chem. B* **2017**, *121*, 1731–1739.

[49] Jiang, L.-L.; Che, M.-X.; Zhao, J.; Zhou, C.-J.; Xie, M.-Y.; Li, H.-Y.; He, J.-H.; Hu, H.-Y. *Journal of Biological Chemistry* **2013**, *288*, 19614–19624.

[50] Shin, Y.; Berry, J.; Pannucci, N.; Haataja, M. P.; Toettcher, J. E.; Brangwynne, C. P. *Cell* **2017**, *168*, 159–171.

[51] Bracha, D.; Walls, M. T.; Wei, M.-T.; Zhu, L.; Kurian, M.; Avalos, J. L.; Toettcher, J. E.; Brangwynne, C. P. *Cell* **2018**, *175*, 1467–1480.

[52] Loughlin, F. E.; Lukavsky, P. J.; Kazeeva, T.; Reber, S.; Hock, E.-M.; Colombo, M.; Von Schroetter, C.; Pauli, P.; Cléry, A.; Mühlemann, O.; Polymenidou, M.; Ruepp, M.-D.; Allain, F. H.-T. *Molecular cell* **2019**, *73*, 490–504.

[53] Knott, M.; Best, R. B. *J. Chem. Phys.* **2014**, *140*, 175102.

[54] Bah, A.; Vernon, R. M.; Siddiqui, Z.; Krzeminski, M.; Muhandiram, R.; Zhao, C.; Sonenberg, N.; Kay, L. E.; Forman-Kay, J. D. *Nature* **2015**, *519*, 106.

[55] Conicella, A. E.; Zerze, G. H.; Mittal, J.; Fawzi, N. L. *Structure* **2016**, *24*, 1537–1549.

[56] Roberts, S.; Harmon, T. S.; Schaal, J.; Miao, V.; Li, K. J.; Hunt, A.; Wen, Y.; Oas, T. G.; Collier, J. H.; Pappu, R. V.; Chilkoti, A. *Nat. Mater.* **2018**, *17*, 1154–1163.

[57] Prouteau, M.; Loewith, R. *Biomolecules* **2018**, *8*, 160.

[58] Jankowsky, E. *Trends in biochemical sciences* **2011**, *36*, 19–29.

[59] Rai, A. K.; Chen, J.-X.; Selbach, M.; Pelkmans, L. *Nature* **2018**, *559*, 211.

[60] Dunker, A. K.; Cortese, M. S.; Romero, P.; Iakoucheva, L. M.; Uversky, V. N. *FEBS J.* **2005**, *272*, 5129–5148.

[61] Markmiller, S. et al. *Cell* **2018**, *172*, 590–604.

[62] Borgia, A.; Borgia, M. B.; Bugge, K.; Kissling, V. M.; Heidarsson, P. O.; Fernandes, C. B.; Sottini, A.; Soranno, A.; Buholzer, K. J.; Nettels, D.; Kragelund, B. B.; Best, R. B.; Schuler, B. *Nature* **2018**, *555*, 61.

[63] Chakraborty, A. K. *Physics Reports* **2001**, *342*, 1–61.

[64] Amaya, J.; Ryan, V. H.; Fawzi, N. L. *Journal of Biological Chemistry* **2018**, *293*, 19522–19531.

[65] Zhang, H.; Elbaum-Garfinkle, S.; Langdon, E. M.; Taylor, N.; Occhipinti, P.; Bridges, A. A.; Brangwynne, C. P.; Gladfelter, A. S. *Mol. Cell* **2015**, *60*, 220–230.

[66] Langdon, E. M.; Qiu, Y.; Niaki, A. G.; McLaughlin, G. A.; Weidmann, C. A.; Gerbich, T. M.; Smith, J. A.; Crutchley, J. M.; Termini, C. M.; Weeks, K. M.; Myong, S.; Gladfelter, A. S. *Science* **2018**, *360*, 922–927.

[67] Khoury, G. A.; Baliban, R. C.; Floudas, C. A. *Scientific reports* **2011**, *1*, 90.

[68] Gomes, E.; Shorter, J. *Journal of Biological Chemistry* **2018**, *294*, 7115–7127.

[69] Martin, E. W.; Mittag, T. *Biochemistry* **2018**, *57*, 2478–2487.

[70] Bentley, E. P.; Frey, B. B.; Deniz, A. A. *Chemistry–A European Journal* **2019**, 5600–5610.

[71] Wang, J.; Choi, J.-M.; Holehouse, A. S.; Lee, H. O.; Zhang, X.; Jahnel, M.; Maharana, S.; Lemaitre, R.; Pozniakovsky, A.; Drechsel, D.; Poser, I.; Pappu, R. V.; Alberti, S.; Hyman, A. A. *Cell* **2018**, *174*, 688–699.

[72] Vernon, R. M.; Chong, P. A.; Tsang, B.; Kim, T. H.; Bah, A.; Farber, P.; Lin, H.; Forman-Kay, J. D. *eLife* **2018**, *7*, e31486.

[73] Li, H.-R.; Chiang, W.-C.; Chou, P.-C.; Wang, W.-J.; Huang, J.-r. *JBC* **2018**, *293*, 6090–6098.

[74] Shakya, A.; King, J. T. *ACS Macro Letters* **2018**, *7*, 1220–1225.

[75] Ackermann, B. E.; Debelouchina, G. T. *Angewandte Chemie* **2019**,

[76] Zhao, B.; Li, N. K.; Yingling, Y. G.; Hall, C. K. *Biomacromolecules* **2015**, *17*, 111–118.

[77] Martin, E. W.; Holehouse, A. S.; Grace, C. R.; Hughes, A.; Pappu, R. V.; Mittag, T. *Journal of the American Chemical Society* **2016**, *138*, 15323–15335.

[78] Rauscher, S.; Pomès, R. *eLife* **2017**, *6*, e26526.

[79] Das, R. K.; Pappu, R. V. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 13392–13397.

[80] Pak, C. W.; Kosno, M.; Holehouse, A. S.; Padrick, S. B.; Mittal, A.; Ali, R.; Yunus, A. A.; Liu, D. R.; Pappu, R. V.; Rosen, M. K. *Mol. Cell* **2016**, *63*, 72–85.

[81] Zagrovic, B.; Bartonek, L.; Polyansky, A. A. *FEBS letters* **2018**, *592*, 2901–2916.

[82] Sawle, L.; Ghosh, K. *J. Chem. Phys.* **2015**, *143*, 085101.

[83] Lin, Y.-H.; Song, J.; Forman-Kay, J. D.; Chan, H. S. *J. Mol. Liq.* **2017**, *228*, 176–193.

[84] Lin, Y.; Currie, S. L.; Rosen, M. K. *J. Biol. Chem.* **2017**, *292*, 19110–19120.

[85] Das, S.; Amin, A. N.; Lin, Y.-H.; Chan, H. S. *Phys. Chem. Chem. Phys.* **2018**, *20*, 28558–28574.

[86] Das, S.; Eisen, A.; Lin, Y.-H.; Chan, H. S. *The Journal of Physical Chemistry B* **2018**, *122*, 5418–5431.

[87] Nott, T. J.; Petsalaki, E.; Farber, P.; Jervis, D.; Fussner, E.; Plochowietz, A.; Craggs, T. D.; Bazett-Jones, D. P.; Pawson, T.; Forman-Kay, J. D.; Baldwin, A. J. *Mol. Cell* **2015**, *57*, 936–947.

[88] Samanta, H. S.; Chakraborty, D.; Thirumalai, D. *J. Chem. Phys.* **2018**, *149*, 163323.

[89] Lytle, T. K.; Chang, L.-W.; Markiewicz, N.; Perry, S. L.; Sing, C. E. *ACS Central Science* **2019**, *5*, 709–718.

[90] Khokhlov, A. R.; Khalatur, P. G. *Physical review letters* **1999**, *82*, 3456.

[91] Ashbaugh, H. S. *The Journal of Physical Chemistry B* **2009**, *113*, 14043–14046.

[92] Dignon, G. L.; Zheng, W.; Best, R. B.; Kim, Y. C.; Mittal, J. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, *115*, 9929–9934.

[93] Brangwynne, C. P.; Tompa, P.; Pappu, R. V. *Nat. Phys.* **2015**, *11*, 899.

[94] Cuylen, S.; Blaukopf, C.; Politi, A. Z.; Müller-Reichert, T.; Neumann, B.; Poser, I.; Ellenberg, J.; Hyman, A. A.; Gerlich, D. W. *Nature* **2016**, *535*, 308.

[95] Lin, Y.-H.; Brady, J. P.; Forman-Kay, J. D.; Chan, H. S. *New J. Phys.* **2017**, *19*, 115003.

[96] Golumbfskie, A. J.; Pande, V. S.; Chakraborty, A. K. *Proceedings of the National Academy of Sciences* **1999**, *96*, 11707–11712.

[97] Falahati, H.; Wieschaus, E. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, 1335–1340.

[98] Nakashima, K. K.; Baaij, J. F.; Spruijt, E. *Soft Matter* **2018**, *14*, 361–367.

[99] Yoo, H.; Triandafillou, C.; Drummond, D. A. *Journal of Biological Chemistry* **2019**, *294*, 7151–7159.

[100] Balu, R.; Dutta, N. K.; Choudhury, N. R.; Elvin, C. M.; Lyons, R. E.; Knott, R.; Hill, A. J. *Acta Biomater.* **2014**, *10*, 4768–4777.

[101] Ruff, K. M.; Roberts, S.; Chilkoti, A.; Pappu, R. V. *J. Mol. Biol.* **2018**, *430*, 4619–4635.

[102] Dignon, G. L.; Zheng, W.; Kim, Y. C.; Mittal, J. *ACS Central Science* **2019**,

[103] Zerze, G. H.; Best, R. B.; Mittal, J. *J. Phys. Chem. B* **2015**, *119*, 14622–14630.

[104] Privalov, P. L.; Makhatadze, G. I. *J. Mol. Biol.* **1993**, *232*, 660–679.

[105] Garde, S.; Garcıa, A. E.; Pratt, L. R.; Hummer, G. *Biophys. Chem.* **1999**, *78*, 21–32.

[106] Huang, D. M.; Chandler, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 8324–8327.

[107] Chandler, D. *Nature* **2005**, *437*, 640–647.

[108] Dill, K. A.; Alonso, D. O.; Hutchinson, K. *Biochemistry* **1989**, *28*, 5439–5449.

[109] Lin, Y.-H.; Forman-Kay, J. D.; Chan, H. S. *Biochemistry* **2018**, *57*, 2499–2508.

[110] Elbaum-Garfinkle, S.; Kim, Y.; Szczepaniak, K.; Chen, C. C.-H.; Eckmann, C. R.; Myong, S.; Brangwynne, C. P. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, 7189–7194.

[111] Debye, P.; Hückel, E. *Physikalische Zeitschrift* **1923**, *24*, 185–206.

[112] Zhang, Y.; Cremer, P. S. *Annual review of physical chemistry* **2010**, *61*, 63–83.

[113] Cho, Y.; Zhang, Y.; Christensen, T.; Sagle, L. B.; Chilkoti, A.; Cremer, P. S. *The Journal of Physical Chemistry B* **2008**, *112*, 13765–13771.

[114] Onuchic, P. L.; Milin, A. N.; Alshareedah, I.; Deniz, A. A.; Banerjee, P. R. *Scientific reports* **2019**, *9*, 1–10.

[115] Altmeyer, M.; Neelsen, K. J.; Teloni, F.; Pozdnyakova, I.; Pellegrino, S.; Grøfte, M.; Rask, M.-B. D.; Streicher, W.; Jungmichel, S.; Nielsen, M. L.; Lukas, J. *Nat. Commun.* **2015**, *6*, 8088.

[116] McGurk, L.; Gomes, E.; Guo, L.; Mojsilovic-Petrovic, J.; Tran, V.; Kalb, R. G.; Shorter, J.; Bonini, N. M. *Molecular cell* **2018**, *71*, 703–717.

[117] Patel, A.; Malinovska, L.; Saha, S.; Wang, J.; Alberti, S.; Krishnan, Y.; Hyman, A. A. *Science* **2017**, *356*, 753–756.

[118] Wurtz, J. D.; Lee, C. F. *Physical review letters* **2018**, *120*, 078102.

[119] Wurtz, J. D.; Lee, C. F. *New Journal of Physics* **2018**, *20*, 045008.

[120] Molliex, A.; Temirov, J.; Lee, J.; Coughlin, M.; Kanagaraj, A. P.; Kim, H. J.; Mittag, T.; Taylor, J. P. *Cell* **2015**, *163*, 123–133.

[121] Putnam, A.; Cassani, M.; Smith, J.; Seydoux, G. *Nature structural & molecular biology* **2019**, 1.

[122] Choi, K.-J.; Tsoi, P. S.; Moosa, M. M.; Paulucci-Holthauzen, A.; Liao, S.-C. J.; Ferreon, J. C.; Ferreon, A. C. M. *Biochemistry* **2018**, *57*, 6822–6826.

[123] Lin, Y.; Protter, D. S.; Rosen, M. K.; Parker, R. *Mol. Cell* **2015**, *60*, 208–219.

[124] Lau, H. K.; Paul, A.; Sidhu, I.; Li, L.; Sabanayagam, C. R.; Parekh, S. H.; Kiick, K. L. *Advanced Science* **2018**, *5*, 1701010.

[125] Kaur, T.; Alshareedah, I.; Wang, W.; Ngo, J.; Moosa, M. M.; Banerjee, P. R. *Biomolecules* **2019**, *9*, 71.

[126] Reed, E. H.; Hammer, D. A. *Soft matter* **2018**, *14*, 6506–6513.

[127] Kato, M.; Yang, Y.-S.; Sutter, B. M.; Wang, Y.; McKnight, S. L.; Tu, B. P. *Cell* **2019**, *177*, 711–721.

[128] Miyazawa, S.; Jernigan, R. L. *J. Mol. Biol.* **1996**, *256*, 623–644.

[129] van Dijk, E.; Varilly, P.; Knowles, T. P.; Frenkel, D.; Abeln, S. *Phys. Rev. Lett.* **2016**, *116*, 078101.

[130] Du, H. et al. *ACS Central Science* **2019**, *5*, 97–108.

[131] Dias, C. L.; Chan, H. S. *The Journal of Physical Chemistry B* **2014**, *118*, 7488–7509.

[132] Winter, R. H. A.; Cinar, H.; Fetahaj, Z.; Cinar, S.; Vernon, R. M.; Chan, H. S. *Chemistry–A European Journal* **2019**,

[133] Dignon, G. L.; Zheng, W.; Mittal, J. *Curr. Opin. Chem. Eng.* **2019**, *24*.

[134] Friesner, R. A.; Guallar, V. *Annu. Rev. Phys. Chem.* **2005**, *56*, 389–427.

[135] Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. A. *J. Comput. Chem.* **2003**, *24*, 1999–2012.

[136] Best, R. B. *Biophys. J.* **2014**, *107*, 1040–1041.

[137] Robustelli, P.; Piana, S.; Shaw, D. E. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, doi:10.1073/pnas.1800690115.

[138] Best, R. B.; Zheng, W.; Mittal, J. *J. Chem. Theor. Comput.* **2014**, *10*, 5113–5124.

[139] Vitalis, A.; Pappu, R. V. *Journal of computational chemistry* **2009**, *30*, 673–699.

[140] Pak, A. J.; Voth, G. A. *Curr. Opin. Struct. Biol.* **2018**, *52*, 119–126.

[141] Ruff, K. M.; Pappu, R. V.; Holehouse, A. S. *Current opinion in structural biology* **2019**, *56*, 1–10.

[142] Carmichael, S. P.; Shell, M. S. *J. Phys. Chem. B* **2012**, *116*, 8383–8393.

[143] Ruff, K. M.; Harmon, T. S.; Pappu, R. V. *J. Chem. Phys.* **2015**, *143*, 12B607_1.

[144] Dignon, G. L.; Zheng, W.; Kim, Y. C.; Best, R. B.; Mittal, J. *PLoS Comput. Biol.* **2018**, *14*, e1005941.

[145] Ghavami, A.; van der Giessen, E.; Onck, P. R. *J. Chem. Theory Comput.* **2012**, *9*, 432–440.

[146] Wu, H.; Wolynes, P. G.; Papoian, G. A. *The Journal of Physical Chemistry B* **2018**, *122*, 11115–11125.

[147] Ando, D.; Zandi, R.; Kim, Y. W.; Colvin, M.; Rexach, M.; Gopinathan, A. *Biophys. J.* **2014**, *106*, 1997–2007.

[148] Feric, M.; Vaidya, N.; Harmon, T. S.; Mitrea, D. M.; Zhu, L.; Richardson, T. M.; Kriwacki, R. W.; Pappu, R. V.; Brangwynne, C. P. *Cell* **2016**, *165*, 1686–1697.

[149] Fei, J.; Jadaliha, M.; Harmon, T. S.; Li, I. T.; Hua, B.; Hao, Q.; Holehouse, A. S.; Reyer, M.; Sun, Q.; Freier, S. M.; Pappu, R. V.; Prasanth, K. V.; Ha, T. *J Cell Sci* **2017**, *130*, 4180–4192.

[150] Nguemaha, V.; Zhou, H.-X. *Sci. Rep.* **2018**, *8*.

[151] Kim, Y. C.; Hummer, G. *J. Mol. Biol.* **2008**, *375*, 1416–1433.

[152] Harmon, T. S.; Holehouse, A. S.; Pappu, R. V. *New J. Phys.* **2018**, *20*, 045002.

[153] Mahynski, N. A.; Errington, J. R.; Shen, V. K. *J. Chem. Phys.* **2017**, *147*, 054105.

[154] Mooij, G.; Frenkel, D.; Smit, B. *J. Phys. Condens. Matter* **1992**, *4*, L255.

[155] Panagiotopoulos, A. Z.; Wong, V.; Floriano, M. A. *Macromolecules* **1998**, *31*, 912–918.

[156] Jacobs, W. M.; Frenkel, D. *Biophys. J.* **2017**, *112*, 683–691.

[157] Jung, H.; Yethiraj, A. *J. Chem. Phys.* **2018**, *148*, 244903.

[158] Dignon, G. L.; Zerze, G. H.; Mittal, J. *The Journal of Physical Chemistry B* **2017**, *121*, 8661–8668.

[159] Uversky, V. N.; Fink, A. L. *Biochim. Biophys. Acta, Proteins Proteomics* **2004**, *1698*, 131–153.

[160] Lucato, C. M.; Lupton, C. C.; Halls, M. L.; Ellisdon, A. M. *J. Mol. Biol.* **2017**, *429*, 1289–1304.

[161] Uversky, V. N.; Oldfield, C. J.; Dunker, A. K. *Annu. Rev. Biophys.* **2008**, *37*, 215–246.

[162] Uversky, V. N. *J. Biol. Chem.* **2016**, *291*, 6681–6688.

[163] Uversky, V. N.; Oldfield, C. J.; Dunker, A. K. *J. Mol. Recognit.* **2005**, *18*, 343–384.

[164] Friedman, R.; Pellarin, R.; Caflisch, A. *J. Mol. Biol.* **2009**, *387*, 407–415.

[165] Laghaei, R.; Mousseau, N.; Wei, G. *J. Phys. Chem. B* **2010**, *114*, 7071–7077.

[166] Murphy, R. D.; Conlon, J.; Mansoor, T.; Luca, S.; Vaiana, S. M.; Buchete, N.-V. *Biophys. Chem.* **2012**, *167*, 1–7.

[167] Knott, M.; Best, R. B. *PLoS Comput. Biol.* **2012**, *8*, e1002605.

[168] Mittal, J.; Yoo, T. H.; Georgiou, G.; Truskett, T. M. *J. Phys. Chem. B* **2013**, *117*, 118–124.

[169] Liang, G.; Zhao, J.; Yu, X.; Zheng, J. *Biochemistry* **2013**, *52*, 1089–1100.

[170] Pannuzzo, M.; Raudino, A.; Milardi, D.; La Rosa, C.; Karttunen, M. *Sci. Rep.* **2013**, *3*, 2781.

[171] Rosenman, D. J.; Connors, C. R.; Chen, W.; Wang, C.; García, A. E. *J. Mol. Biol.* **2013**, *425*, 3338–3359.

[172] Rudzinski, J. F.; Noid, W. G. *J. Chem. Theory Comput.* **2015**, *11*, 1278–1291.

[173] Levine, Z. A.; Shea, J.-E. *Curr. Opin. Struct. Biol.* **2017**, *43*, 95–103.

[174] Sunde, M.; Serpell, L. C.; Bartlam, M.; Fraser, P. E.; Pepys, M. B.; Blake, C. C. *J. Mol. Biol.* **1997**, *273*, 729–739.

[175] Hull, R. L.; Westermark, G. T.; Westermark, P.; Kahn, S. E. *J. Clin. Endocrinol. Metab.* **2004**, *89*, 3629–3643.

[176] Lorenzo, A.; Razzaboni, B.; Weir, G. C.; Yankner, B. A. *Nature* **1994**, *368*, 756–760.

[177] Mirzabekov, T. A.; Lin, M.-c.; Kagan, B. L. *J. Biol. Chem.* **1996**, *271*, 1988–1992.

[178] Janson, J.; Ashley, R. H.; Harrison, D.; McIntyre, S.; Butler, P. C. *Diabetes* **1999**, *48*, 491–498.

[179] Vaiana, S. M.; Ghirlando, R.; Yau, W.-M.; Eaton, W. A.; Hofrichter, J. *Biophys. J.* **2008**, *94*, L45–L47.

[180] Brender, J. R.; Salamekh, S.; Ramamoorthy, A. *Acc. Chem. Res.* **2011**, *45*, 454–462.

[181] Westermark, P.; Engström, U.; Johnson, K. H.; Westermark, G. T.; Betsholtz, C. *Proc. Natl. Acad. Sci. U. S. A.* **1990**, *87*, 5036–5040.

[182] Brender, J. R.; Lee, E. L.; Cavitt, M. A.; Gafni, A.; Steel, D. G.; Ramamoorthy, A. *J. Am. Chem. Soc.* **2008**, *130*, 6424–6429.

[183] Cayla Miller and Gül H. Zerze and Jeetain Mittal, *J. Phys. Chem. B* **2013**, *117*, 16066–16075.

[184] Williamson, J. A.; Miranker, A. D. *Protein Sci.* **2007**, *16*, 110–117.

[185] Williamson, J. A.; Loria, J. P.; Miranker, A. D. *J. Mol. Biol.* **2009**, *393*, 383–396.

[186] Jayasinghe, S. A.; Langen, R. *Biochemistry* **2005**, *44*, 12113–12119.

[187] Knight, J. D.; Hebda, J. A.; Miranker, A. D. *Biochemistry* **2006**, *45*, 9496–9508.

[188] Lopes, D.; Meister, A.; Gohlke, A.; Hauser, A.; Blume, A.; Winter, R. *Biophys. J.* **2007**, *93*, 3132–3141.

[189] Mukherjee, S.; Chowdhury, P.; Gai, F. *J. Phys. Chem. B* **2008**, *113*, 531–535.

[190] Young, L. M.; Tu, L.-H.; Raleigh, D. P.; Ashcroft, A. E.; Radford, S. E. *Chem. Sci.* **2017**, *8*, 5030–5040.

[191] Engel, M. F.; Yigittop, H.; Elgersma, R. C.; Rijkers, D. T.; Liskamp, R. M.; de Kruijff, B.; Höppener, J. W.; Killian, J. A. *J. Mol. Biol.* **2006**, *356*, 783–789.

[192] Apostolidou, M.; Jayasinghe, S. A.; Langen, R. *J. Biol. Chem.* **2008**, *283*, 17205–17210.

[193] Patil, S. M.; Xu, S.; Sheftic, S. R.; Alexandrescu, A. T. *J. Biol. Chem.* **2009**, *284*, 11982–11991.

[194] Nanga, R. P. R.; Brender, J. R.; Vivekanandan, S.; Ramamoorthy, A. *Biochim. Biophys. Acta, Biomembr.* **2011**, *1808*, 2337–2342.

[195] Skeby, K. K.; Andersen, O. J.; Pogorelov, T. V.; Tajkhorshid, E.; Schiøtt, B. *Biochemistry* **2016**, *55*, 2031–2042.

[196] Kagan, B.; Azimov, R.; Azimova, R. *J. Membr. Biol.* **2004**, *202*, 1–10.

[197] Martel, A.; Antony, L.; Gerelli, Y.; Porcar, L.; Fluitt, A.; Hoffmann, K. Q.; Kiesel, I.; Vivaudou, M.; Fragneto, G.; de Pablo, J. J. *J. Am. Chem. Soc.* **2016**, *139*, 137–148.

[198] Fu, L.; Ma, G.; Yan, E. C. *J. Am. Chem. Soc.* **2010**, *132*, 5405–5412.

[199] Rustenbeck, I.; Matthies, A.; Lenzen, S. *Lipids* **1994**, *29*, 685–692.

[200] Jayasinghe, S. A.; Langen, R. *Biochim. Biophys. Acta, Biomembr.* **2007**, *1768*, 2002–2009.

[201] Ling, Y. L.; Strasfeld, D. B.; Shim, S.-H.; Raleigh, D. P.; Zanni, M. T. *J. Phys. Chem. B* **2009**, *113*, 2498–2505.

[202] Zhang, X.; St Clair, J. R.; London, E.; Raleigh, D. P. *Biochemistry* **2017**, *56*, 376–390.

[203] Qian, Z.; Jia, Y.; Wei, G. *J. Diabetes Res.* **2015**, *2016*, 1749196.

[204] Sciacca, M. F.; Lolicato, F.; Di Mauro, G.; Milardi, D.; DUrso, L.; Satriano, C.; Ramamoorthy, A.; La Rosa, C. *Biophys. J.* **2016**, *111*, 140–151.

[205] Cooper, G.; Willis, A.; Clark, A.; Turner, R.; Sim, R.; Reid, K. *Proc. Natl. Acad. Sci. U. S. A.* **1987**, *84*, 8628–8632.

[206] Balasubramaniam, A.; Renugopalakrishnan, V.; Stein, M.; Fischer, J.; Chance, W. *Peptides* **1991**, *12*, 919–924.

[207] Turk, J.; Wolf, B. A.; Lefkowith, J. B.; Stump, W. T.; McDaniel, M. L. *Biochim. Biophys. Acta, Lipids Lipid Metab.* **1986**, *879*, 399–409.

[208] Jämbeck, J. P.; Lyubartsev, A. P. *J. Phys. Chem. B* **2012**, *116*, 3164–3179.

[209] Jämbeck, J. P.; Lyubartsev, A. P. *J. Chem. Theory Comput.* **2012**, *8*, 2938–2948.

[210] Jämbeck, J. P.; Lyubartsev, A. P. *J. Chem. Theory Comput.* **2012**, *9*, 774–784.

[211] Jo, S.; Kim, T.; Iyer, V. G.; Im, W. *J. Comput. Chem.* **2008**, *29*, 1859–1865.

[212] Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.

[213] Best, R. B.; Mittal, J. *J. Phys. Chem. B* **2010**, *114*, 14916–14923.

[214] Abascal, J. L. F.; Vega, C. *J. Chem. Phys.* **2005**, *123*, 234505.

[215] Palazzesi, F.; Prakash, M. K.; Bonomi, M.; Barducci, A. *J. Chem. Theory Comput.* **2014**, *11*, 2–7.

[216] Parrinello, M.; Rahman, A. *J. Appl. Phys.* **1981**, *52*, 7182–7190.

[217] Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

[218] Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

[219] Bonomi, M.; Parrinello, M. *Phys. Rev. Lett.* **2010**, *104*, 190601.

[220] Deighan, M.; Bonomi, M.; Pfaendtner, J. *J. Chem. Theory Comput.* **2012**, *8*, 2189–2192.

[221] Bussi, G.; Donadio, D.; Parrinello, M. *J. Chem. Phys.* **2007**, *126*, 014101.

[222] Wiltzius, J. J.; Sievers, S. A.; Sawaya, M. R.; Eisenberg, D. *Protein Sci.* **2009**, *18*, 1521–1530.

[223] Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577–2637.

[224] Iglesias, J.; Sanchez-Martínez, M.; Crehuet, R. *Intrinsically Disord. Proteins* **2013**, *1*, e25323.

[225] Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33–38.

[226] Wippich, F.; Bodenmiller, B.; Trajkovska, M. G.; Wanka, S.; Aebersold, R.; Pelkmans, L. *Cell* **2013**, *152*, 791–805.

[227] Fromm, S. A.; Kamenz, J.; Nöldeke, E. R.; Neu, A.; Zocher, G.; Sprangers, R. *Angew. Chem. Int. Edit.* **2014**, *53*, 7354–7359.

[228] Kato, M. et al. *Cell* **2012**, *149*, 753–767.

[229] Marzahn, M. R. et al. *EMBO J.* **2016**, e201593169.

[230] Uversky, V. N. *Adv. Colloid Interfac.* **2017**, *239*, 97–114.

[231] Biamonti, G.; Vourch, C. *Cold Spring Harb. Perspect. Biol.* **2010**, *2*, a000695.

[232] Morimoto, M.; Boerkoel, C. F. *Biology* **2013**, *2*, 976–1033.

[233] Hnisz, D.; Shrinivas, K.; Young, R. A.; Chakraborty, A. K.; Sharp, P. A. *Cell* **2017**, *169*, 13–23.

[234] Su, X.; Ditlev, J. A.; Hui, E.; Xing, W.; Banjade, S.; Okrut, J.; King, D. S.; Taunton, J.; Rosen, M. K.; Vale, R. D. *Science* **2016**, *352*, 595–599.

[235] Li, P.; Banjade, S.; Cheng, H.-C.; Kim, S.; Chen, B.; Guo, L.; Llaguno, M.; Hollingsworth, J. V.; King, D. S.; Banani, S. F.; Russo, P. S.; Jiang, Q.-X.; Nixon, B. T.; Rosen, M. K. *Nature* **2012**, *483*, 336.

[236] Jiang, H.; Wang, S.; Huang, Y.; He, X.; Cui, H.; Zhu, X.; Zheng, Y. *Cell* **2015**, *163*, 108–122.

[237] Nott, T. J.; Craggs, T. D.; Baldwin, A. J. *Nat. Chem.* **2016**, *8*, 569–575.

[238] Xiang, S.; Kato, M.; Wu, L. C.; Lin, Y.; Ding, M.; Zhang, Y.; Yu, Y.; McKnight, S. L. *Cell* **2015**, *163*, 829–839.

[239] Mateju, D.; Franzmann, T. M.; Patel, A.; Kopach, A.; Boczek, E. E.; Maharana, S.; Lee, H. O.; Carra, S.; Hyman, A. A.; Alberti, S. *EMBO J.* **2017**, e201695957.

[240] Berry, J.; Weber, S. C.; Vaidya, N.; Haataja, M.; Brangwynne, C. P. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, E5237–E5245.

[241] Kim, Y.; Myong, S. *Mol. Cell* **2016**, *63*, 865–876.

[242] Bates, F. S. *Science* **1991**, *251*, 898–905.

[243] Lin, Y.-H.; Forman-Kay, J. D.; Chan, H. S. *Phys. Rev. Lett.* **2016**, *117*, 178101.

[244] Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. *Science* **2010**, *330*, 341–346.

[245] Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. *Science* **2011**, *334*, 517–520.

[246] Piana, S.; Donchev, A. G.; Robustelli, P.; Shaw, D. E. *J. Phys. Chem. B* **2015**, *119*, 5113–5123.

[247] Blas, F. J.; MacDowell, L. G.; de Miguel, E.; Jackson, G. *J. Chem. Phys.* **2008**, *129*, 144703.

[248] Kim, J.; Keyes, T.; Straub, J. E. *J. Chem. Phys.* **2010**, *132*, 224107.

[249] Vance, C. et al. *Science* **2009**, *323*, 1208–1211.

[250] Kwiatkowski, T. J. et al. *Science* **2009**, *323*, 1205–1208.

[251] Kyte, J.; Doolittle, R. F. *J. Mol. Biol.* **1982**, *157*, 105–132.

[252] Kapcha, L. H.; Rossky, P. J. *J. Mol. Biol.* **2014**, *426*, 484–498.

[253] Ashbaugh, H. S.; Hatch, H. W. *J. Am. Chem. Soc.* **2008**, *130*, 9536–9542.

[254] Müller-Späth, S.; Sorrano, A.; Hirschfeld, V.; Hofmann, H.; Rüegger, S.; Reymond, L.; Nettels, D.; Schuler, B. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 14609–14614.

[255] Sherman, E.; Haran, G. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 11539–11543.

[256] Kjaergaard, M.; Norholm, A.-B.; Hendus-Altenburger, R.; Pedersen, S. F.; Poulsen, F. M.; Kragelund, B. B. *Protein Sci.* **2010**, *19*, 1555–1564.

[257] Flanagan, J. M.; Kataoka, M.; Shortle, D.; Engelman, D. M. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89*, 748–752.

[258] Nath, A.; Sammalkorpi, M.; DeWitt, D. C.; Trexler, A. J.; Elbaum-Garfinkle, S.; OHern, C. S.; Rhoades, E. *Biophys. J.* **2012**, *103*, 1940–1949.

[259] Miller, C. M.; Kim, Y. C.; Mittal, J. *Biophys. J.* **2016**, *111*, 28–37.

[260] Weeks, J. D.; Chandler, D.; Andersen, H. C. *J. Chem. Phys.* **1971**, *56*, 5237–5247.

[261] Borgia, A.; Zheng, W.; Buholzer, K.; Borgia, M. B.; Schuler, A.; Hofmann, H.; Soranno, A.; Nettels, D.; Gast, K.; Grishaev, A.; Best, R. B.; Schuler, B. *J. Am. Chem. Soc.* **2016**, *138*, 11714–11726.

[262] Zheng, W.; Borgia, A.; Buholzer, K.; Grishaev, A.; Schuler, B.; Best, R. B. *J. Am. Chem. Soc.* **2016**, *138*, 11702–11713.

[263] Fuertes, G. et al. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, E6342–E6351.

[264] Song, J.; Gomes, G.-N.; Shi, T.; Gradinaru, C. C.; Chan, H. S. *Biophys. J.* **2017**, *113*, 1012–1024.

[265] O'Brien, E. P.; Morrison, G.; Brooks, B. R.; Thirumalai, D. *J. Chem. Phys.* **2012**, *130*, 124903.

[266] Mao, A. H.; Crick, S. L.; Vitalis, A.; Chicoine, C.; Pappu, R. V. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 8183–8188.

[267] Silmore, K. S.; Howard, M. P.; Panagiotopoulos, A. Z. *Mol. Phys.* **2017**, *115*, 320–327.

[268] Plimpton, S. *J. Comput. Phys.* **1995**, *117*, 1–19.

[269] Anderson, J. A.; Lorenz, C. D.; Travesset, A. *J. Comput. Phys.* **2008**, *227*, 5342–5359.

[270] Rowlinson, J. S.; Widom, B. *Molecular theory of capillarity*; Courier Corporation, 2013.

[271] Flory, P. J. *J. Chem. Phys.* **1942**, *10*, 51–61.

[272] Huggins, M. L. *J. Phys. Chem.* **1942**, *46*, 151–158.

[273] Šali, A.; Blundell, T. L. *J. Mol. Biol.* **1993**, *234*, 779–815.

[274] Sengoku, T.; Nureki, O.; Nakamura, A.; Kobayashi, S.; Yokoyama, S. *Cell* **2006**, *125*, 287–300.

[275] Kwon, I.; Kato, M.; Xiang, S.; Wu, L.; Theodoropoulos, P.; Mirzaei, H.; Han, T.; Xie, S.; Corden, J. L.; McKnight, S. L. *Cell* **2013**, *155*, 1049–1060.

[276] Schrier, E. E.; Schrier, E. B. *J. Phys. Chem.* **1967**, *71*, 1851–1860.

[277] Nandi, P. K.; Robinson, D. R. *J. Am. Chem. Soc.* **1972**, *94*, 1299–1308.

[278] Nandi, P. K.; Robinson, D. R. *J. Am. Chem. Soc.* **1972**, *94*, 1308–1315.

[279] Baldwin, R. L. *Biophys. J.* **1996**, *71*, 2056–2063.

[280] Lau, H. K.; Li, L.; Jurusik, A. K.; Sabanayagam, C. R.; Kiick, K. L. *ACS Biomater. Sci. Eng.* **2016**, *3*, 757–766.

[281] Simon, J. R.; Carroll, N. J.; Rubinstein, M.; Chilkoti, A.; López, G. P. *Nat Chem* **2017**, *9*, 509.

[282] Wang, R.; Wang, Z.-G. *Macromolecules* **2014**, *47*, 4094–4102.

[283] Flory, P. J. *J. Chem. Phys.* **1949**, *17*, 303–310.

[284] Witten Jr, T.; Schafer, L. *J. Phys A-Math. Gen.* **1978**, *11*, 1843.

[285] Le Guillou, J.; Zinn-Justin, J. *Physical Review Letters* **1977**, *39*, 95.

[286] Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13*, 1011–1021.

[287] Sheng, Y. J.; Panagiotopoulos, A. Z.; Kumar, S. K.; Szleifer, I. *Macromolecules* **1994**, *27*, 400–406.

[288] Qi, X.; Pang, Q.; Wang, J.; Zhao, Z.; Wang, O.; Xu, L.; Mao, J.; Jiang, Y.; Li, M.; Xing, X.; Yu, W.; Asan,; Xia, W. *Calcified. Tissue Int.* **2017**, *101*, 159–169.

[289] Lynden-Bell, R.; Kohanoff, J.; Del Popolo, M. *Faraday discussions* **2005**, *129*, 57–67.

[290] Vega, C.; De Miguel, E. *J. Chem. Phys.* **2007**, *126*, 154707.

[291] Müller, E. A.; Mejía, A. *J. Phys. Chem. B* **2011**, *115*, 12822–12834.

[292] Harmon, T. S.; Holehouse, A. S.; Rosen, M. K.; Pappu, R. V. *eLife* **2017**, *6*.

[293] Warner IV, J. B.; Ruff, K. M.; Tan, P. S.; Lemke, E. A.; Pappu, R. V.; Lashuel, H. A. *J. Am. Chem. Soc.* **2017**, *139*, 14456–14469.

[294] Posey, A. E.; Ruff, K. M.; Harmon, T. S.; Crick, S. L.; Li, A.; Diamond, M. I.; Pappu, R. V. *J. Biol. Chem.* **2018**, http://www.jbc.org/content/early/2018/01/22/jbc.RA117.000357.

[295] Zheng, W.; Best, R. B. *J. Mol. Biol.* **2018**,

[296] George, A.; Chiang, Y.; Guo, B.; Arabshahi, A.; Cai, Z.; Wilson, W. W. *Methods Enzymol.*; Elsevier, 1997; Vol. 276; pp 100–110.

[297] Platten, F.; Hansen, J.; Wagner, D.; Egelhaaf, S. U. *J. Phys. Chem. Lett.* **2016**, *7*, 4008–4014.

[298] Neal, B.; Asthagiri, D.; Lenhoff, A. *Biophys. J.* **1998**, *75*, 2469–2477.

[299] Wang, L.-P.; Martinez, T. J.; Pande, V. S. *J. Phys. Chem. Lett.* **2014**, *5*, 1885–1891.

[300] Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell Jr, A. D. *Nat Methods* **2017**, *14*, 71.

[301] Michalet, X.; Weiss, S.; Jäger, M. *Chem. Rev.* **2006**, *106*, 1785–1813.

[302] Schuler, B.; Hofmann, H. *Curr. Opin. Struct. Biol.* **2013**, *23*, 36–47.

[303] Bernadó, P.; Svergun, D. I. *Mol. BioSyst.* **2012**, *8*, 151–167.

[304] Ditlev, J. A.; Case, L. B.; Rosen, M. K. *J. Mol. Biol.* **2018**, *430*, 4666–4684.

[305] Milovanovic, D.; Wu, Y.; Bian, X.; De Camilli, P. *Science* **2018**, *361*, 604–607.

[306] Shakya, A.; King, J. T. *Biophys. J.* **2018**, *115*, 1840–1847.

[307] Lu, H.; Yu, D.; Hansen, A. S.; Ganguly, S.; Liu, R.; Heckert, A.; Darzacq, X.; Zhou, Q. *Nature* **2018**, *558*, 318–323.

[308] Partridge, S.; Davis, H.; Adair, G. *Biochem. J.* **1955**, *61*, 11.

[309] Urry, D. W. *J. Prot. Chem.* **1988**, *7*, 1–34.

[310] Meyer, D. E.; Chilkoti, A. *Biomacromolecules* **2002**, *3*, 357–367.

[311] Bellingham, C. M.; Lillie, M. A.; Gosline, J. M.; Wright, G. M.; Starcher, B. C.; Bailey, A. J.; Woodhouse, K. A.; Keeley, F. W. *Biopolymers* **2003**, *70*, 445–455.

[312] Lyons, R. E.; Nairn, K. M.; Huson, M. G.; Kim, M.; Dumsday, G.; Elvin, C. M. *Biomacromolecules* **2009**, *10*, 3009–3014.

[313] Muiznieks, L. D.; Weiss, A. S.; Keeley, F. W. *Biochem. and Cell Biol.* **2010**, *88*, 239–250.

[314] Garcia, C. G.; Kiick, K. L. *Acta Biomater.* **2019**, *84*, 34–48.

[315] Panganiban, B.; Qiao, B.; Jiang, T.; DelRe, C.; Obadia, M. M.; Nguyen, T. D.; Smith, A. A.; Hall, A.; Sit, I.; Crosby, M. G.; Dennis, P. B.; Drockenmuller, E.; de la Cruz, M. O.; Xu, T. *Science* **2018**, *359*, 1239–1243.

[316] Zhang, H.; Aonbangkhen, C.; Tarasovetc, E. V.; Ballister, E. R.; Chenoweth, D. M.; Lampson, M. A. *Nat. Chem. Biol.* **2017**, *13*, 1096.

[317] Li, L.; Luo, T.; Kiick, K. L. *Macromol. Rapid Commun.* **2015**, *36*, 90–95.

[318] Dutta, N. K.; Truong, M. Y.; Mayavan, S.; Roy Choudhury, N.; Elvin, C. M.; Kim, M.; Knott, R.; Nairn, K. M.; Hill, A. J. *Angew. Chem.* **2011**, *123*, 4520–4523.

[319] Das, P.; Matysiak, S.; Mittal, J. *ACS Cent. Sci.* **2018**, *4*, 534–542.

[320] Wuttke, R.; Hofmann, H.; Nettels, D.; Borgia, M. B.; Mittal, J.; Best, R. B.; Schuler, B. *Proc. Natl. Acad. Sci. U.S.A.* **2014**, *111*, 5213–5218.

[321] Nguyen, T. D.; Qiao, B.; de la Cruz, M. O. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, *115*, 6578–6583.

[322] Makhatadze, G. I.; Privalov, P. L. *J. Mol. Biol.* **1993**, *232*, 639–659.

[323] Garde, S.; Hummer, G.; García, A. E.; Paulaitis, M. E.; Pratt, L. R. *Phys. Rev. Lett.* **1996**, *77*, 4966.

[324] Cheung, J. K.; Shah, P.; Truskett, T. M. *Biophys. J.* **2006**, *91*, 2427–2435.

[325] Patel, B. A.; Debenedetti, P. G.; Stillinger, F. H.; Rossky, P. J. *Biophys. J.* **2007**, *93*, 4116–4127.

[326] Torrie, G. M.; Valleau, J. P. *J. Comput. Phys.* **1977**, *23*, 187–199.

[327] Coeytaux, K.; Poupon, A. *Bioinformatics* **2005**, *21*, 1891–1900.

[328] Dill, K. A. *Biochemistry* **1997**, *24*, 1501–1509.

[329] Nozaki, Y.; Tanford, C. *Journal of Biological Chemistry* **1971**, *246*, 2211–2217.

[330] van Dijk, E.; Hoogeveen, A.; Abeln, S. *PLoS Comput. Biol.* **2015**, *11*, e1004277.

[331] Pucci, F.; Bourgeas, R.; Rooman, M. *Sci. Rep.* **2016**, *6*, 23257.

[332] Wales, D. J.; Doye, J. P. *J. Phys. Chem. A* **1997**, *101*, 5111–5116.

[333] Jung, J. G.; Bae, Y. C. *J. Polym. Sci. B.* **2010**, *48*, 162–167.

[334] Clark, E.; Lipson, J. *Polymer* **2012**, *53*, 536–545.

[335] Singh, V. R.; Lapidus, L. J. *The Journal of Physical Chemistry B* **2008**, *112*, 13172–13176.

[336] Crick, S. L.; Jayaraman, M.; Frieden, C.; Wetzel, R.; Pappu, R. V. *Proceedings of the National Academy of Sciences* **2006**, *103*, 16764–16769.

[337] Yang, Y.; Jones, H. B.; Dao, T. P.; Castañeda, C. A. *The Journal of Physical Chemistry B* **2019**, *123*, 3618–3629.

[338] Murray, D. T.; Kato, M.; Lin, Y.; Thurber, K. R.; Hung, I.; McKnight, S. L.; Tycko, R. *Cell* **2017**, *171*, 615–627.

[339] Hughes, M. P.; Sawaya, M. R.; Boyer, D. R.; Goldschmidt, L.; Rodriguez, J. A.; Cascio, D.; Chong, L.; Gonen, T.; Eisenberg, D. S. *Science* **2018**, *359*, 698–701.

[340] Guenther, E. L.; Cao, Q.; Trinh, H.; Lu, J.; Sawaya, M. R.; Cascio, D.; Boyer, D. R.; Rodriguez, J. A.; Hughes, M. P.; Eisenberg, D. S. *Nature structural & molecular biology* **2018**, *25*, 463.

[341] Kendrew, J. C.; Bodo, G.; Dintzis, H. M.; Parrish, R.; Wyckoff, H.; Phillips, D. C. *Nature* **1958**, *181*, 662–666.

[342] Huber, R.; Bennett Jr, W. S. *Biopolymers: Original Research on Biomolecules* **1983**, *22*, 261–279.

[343] Wüthrich, K. *Journal of Biological Chemistry* **1990**, *265*, 22059–22062.

[344] Kosol, S.; Contreras-Martos, S.; Cedeño, C.; Tompa, P. *Molecules* **2013**, *18*, 10802–10828.

[345] Pietrek, L. M.; Stelzl, L. S.; Hummer, G. *bioRxiv* **2019**, 731133.

[346] Laio, A.; Gervasio, F. L. *Reports on Progress in Physics* **2008**, *71*, 126601.

[347] Barducci, A.; Bussi, G.; Parrinello, M. *Phys. Rev. Lett.* **2008**, *100*, 020603.

[348] Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R. A.; Parrinello, M. *Computer Physics Communications* **2009**, *180*, 1961–1972.

[349] Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. *Computer Physics Communications* **2014**, *185*, 604–613.

[350] Pfaendtner, J. *Biomolecular Simulations*; Springer, 2019; pp 179–200.

[351] Prakash, M. K.; Barducci, A.; Parrinello, M. *Journal of chemical theory and computation* **2011**, *7*, 2025–2027.

[352] Jorgensen, W. L.; Tirado-Rives, J. *Journal of the American Chemical Society* **1988**, *110*, 1657–1666.

[353] Luzar, A.; Chandler, D. *Nature* **1996**, *379*, 55.

[354] Friend, L. R.; Landsberg, M. J.; Nouwens, A. S.; Wei, Y.; Rothnagel, J. A.; Smith, R. *PloS one* **2013**, *8*, e75669.

[355] Edgar, R. C. *Nucleic Acids Res.* **2004**, *32*, 1792–1797.

[356] Kučerka, N.; Tristram-Nagle, S.; Nagle, J. F. *J. Mem. Biol.* **2006**, *208*, 193–202.

[357] Kučerka, N.; Nieh, M.-P.; Katsaras, J. *Biochimica et Biophysica Acta (BBA)-Biomembranes* **2011**, *1808*, 2761–2771.

[358] Seelig, J.; Waespe-Sarcevic, N. *Biochemistry* **1978**, *17*, 3310–3315.

Appendix A

List of amino acid sequences

A.1 Sequence alignment of VASA protein and LAF-1 helicase

Since there is no solved structure for the helicase domain of LAF-1, we started by predicting its structure from its homologue VASA. Both LAF-1 and VASA belong to the DEAD-Box family and the structure of *Drosophila* VASA has been solved [274]. We first aligned the LAF-1 sequence to the structured part of VASA using the MUSCLE v3.8 web service [355]. The sequence similarity is 51% and the alignment of the structured part is shown below:

VASA; Residue 202-621; chapter 3

LAF-1; Residue 187-623; chapter 3

```
VASA    YIPPEPSNDAIEI-FSSGIASGIHFSKYNNIPVKVTGSDVPQPIQHFTSADLRDIIIDNV
LAF-1   WENRGARDERIEQELFSGQLSGINFDKYEEIPVEATGDDVPQPISLFSDLSLHEWIEENI
        :    . :: **  : **  ***:*.**::***:.**.******. *:. .*.: * :*:
VASA    NKSGYKIPTPIQKCSIPVISSGRDLMACAQTGSGKTAAFLLPILSKLLED-PHELEL---
LAF-1   KTAGYDRPTPVQKYSIPALQGGRDLMSCAQTGSGKTAAFLVPLVNAILQDGPDAVHRSVT
        :.:**. ***:** ***.:..*****:*************:*::. :*:* *  :
VASA    ---GR----PQVVIVSPTRELAIQIFNEARKFAFESYLKIGIVYGG-TSFRHQNECITRG
LAF-1   SSGGRKKQYPSALVLSPTRELSLQIFNESRKFAYRTPITSALLYGGRENYKDQIHKLRLG
           **    *..:::******::*****:****: : :. .::***  .:. *   :  *
VASA    CHVVIATPGRLLDFVDRTFITFEDTRFVVLDEADRMLDMGFSEDMRRIM--THVTMRPEH
LAF-1   CHILIATPGRLIDVMDQGLIGMEGCRYLVLDEADRMLDMGFEPQIRQIVECNRMPSKEER
        **::*******:*.:*. :* :*. *::*************. ::*.*:  ..:. . *.
VASA    QTLMFSATFPEEIQRMAGEFLK-NYVFVAIGIVGGACSDVKQTIYEVNKYAKRSKLIEIL
```

```
LAF-1   ITAMFSATFPKEIQLLAQDFLKENYVFLAVGRVGSTSENIMQKIVWVEEDEKRSYLMDLL
         * *******:*** :* :*** ****:*:* **.:..:: *.*  *::  *** *:::*
VASA    SEQADG--TIVFVETKRGADFLASFLSEKEFPTTSIHGDRLQSQREQALRDFKNGSMKVL
LAF-1   DATGDSSLTLVFVETKRGASDLAYYLNRQNYEVVTIHGDLKQFEREKHLDLFRTGTAPIL
        .  .*.  *:*********. ** :*. ::: ..:****  * :**: *  *..*:  :*
VASA    IATSVASRGLDIKNIKHVINYDMPSKIDDYVHRIGRTGRVGNNGRATSFFDPEKDRAIAA
LAF-1   VATAVAARGLDIPNVKHVINYDLPSDVDEYVHRIGRTGRVGNVGLATSFFN-DKNRNIAR
        :**:**:***** *:*******:**.:*:************* * *****: :*:* **
VASA    DLVKILEGSGQTVPDFLR
LAF-1   ELMDLIVEANQELPDWLE
        :*:.::  :.* :**:*
```

We provided the VASA structure (PDB:2DB3) and the sequence alignment information shown above as the inputs to the Modeller software package v9.17 [273]. We then repeated the modelling process 100 times and picked the structure with the smallest energy as best model for the structure of the LAF-1 helicase domain (Fig. 3.8).

A.2 IDP sequences

Protein/peptide sequences will be listed with names and chapters in which they are used. Sequences are represented using single-letter amino acid representation with breaks every 10 residues. Residues which are altered or mutated from the wild-type variant of the respective protein are highlighted in red except for in extreme cases such as when the full sequence is shuffled.

hIAPP WT; chapter 2

```
KCNTATCATQ RLANFLVHSS NNFGAILSST NVGSNTY
```

CspTm; chapter 3

```
GPGMRGKVKW FDSKKGYGFI TKDEGGDVFV HWSAIEMEGF KTLKEGQVVE FEIQEGKKGG
QAAHVKV
```

HIV Integrase; chapter 3

```
GSHCFLDGID KAQEEHEKYH SNWRAMASDF NLPPVVAKEI VASCDKCQLK GEAMHGQVDC
```

ProTαN; chapter 3

```
GPSDAAVDTS SEITTKDLKE KKEVVEEAEN GRDAPANGNA ENEENGEQEA DNEVDEECEE
GGEEEEEEEE GDGEEEDGDE DEEAESATGK RAAEDDEDDD VDTKKQKTDE DD
```

ProTαC; chapter 3

```
MAHHHHHHSA ALEVLFQGPM SDAAVDTSSE ITTKDLKEKK EVVEEAENGR DAPANGNANE
ENGEQEADNE VDEECEEGGE EEEEEEEGDG EEEDGDEDEE AESATGKRAA EDDEDDDVDT
KKQKTDEDD
```

R15; chapter 3

```
KLKEANKQQN FNTGIKDFDF WLSEVEALLA SEDYGKDLAS VNNLLKKHQL LEADISAHED
RLKDLNSQAD SLMTSSAFDT SQVKDKRETI NGRFQRIKSM AAARRAKLNE SHRL
```

R17; chapter 3

```
RLEESLEYQQ FVANVEEEEA WINEKMTLVA SEDYGDTLAA IQGLLKKHEA FETDFTVHKD
RVNDVAANGE DLIKKNNHHV ENITAKMKGL KGKVSDLEKA
```

hCyp; chapter 3

```
SSFHRIIPGF MSQGGDFTRH NGTGGKSIYG EKFEDENFIL KHTGPGILSM ANAGPNTNGS
QFFISTAKTE FLDGKHVVFG KVKEGMNIVE AMERFGSRNG KTSKKITIAD SGQLE
```

Protein L; chapter 3

```
MEEVTIKANL IFANGSTQTA EFKGTFEKAT SEAYAYADTL KKDNGEWTVD VADKGYTLNI
KFAG
```

ACTR; chapter 3

```
GTQNRPLLRN SLDDLVGPPS NLEGQSDERA LLDQLHTLLS NTDATGLEEI DRALGIPELV
NQGQALEPKQ D
```

hNHE1cdt; chapter 3

```
MVPAHKLDSP TMSRARIGSD PLAYEPKEDL PVITIDPASP QSPESVDLVN EELKGKVLGL
SRDPAKVAEE DEDDDGGIMM RSKETSSPGT DDVFTPAPSD SPSSQRIQRC LSDPGPHPEP
GEGEPFFPKG Q
```

sNase; chapter 3

```
ATSTKKLHKE PATLIKAIDG DTVKLMYKGQ PMTFRLLLVD TPETKHPKKG VEKYGPEASA
FTKKMVENAK KIEVEFDKGQ RTDKYGRGLA YIYADGKMVN EALVRQGLAK VAYVYKPNNT
HEQHLRKSEA QAKKEK
```

α-synuclein; chapter 3

```
MDVFMKGLSK AKEGVVAAAE KTKQGVAEAA GKTKEGVLYV GSKTKEGVVH GVATVAEKTK
EQVTNVGGAV VTGVTAVAQK TVEGAGSIAA ATGFVKKDQL GKNEEGAPQE GILEDMPVDP
DNEAYEMPSE EGYQDYEPEA
```

FUS LC WT; chapters 3,4

```
MASNDYTQQA TQSYGAYPTQ PGQGYSQQSS QPYGQQSYSG YSQSTDTSGY GQSSYSSYGQ
SQNTGYGTQS TPQGYGSTGG YGSSQSSQSS YGQQSSYPGY GQQPAPSSTS GSYGSSSQSS
SYGQPQSGSY SQQPSYGGQQ QSYGQQQSYN PPQGYGQQNQ YNS
```

FUS 6E; chapters 3,4

```
MASNDYTQQA TQSYGAYPTQ PGQGYEQQSE QPYGQQSYSG YSQSTDTSGY GQSSYSSYGQ
SQNTGYGEQS TPQGYGSTGG YGSEQSEQSS YGQQSSYPGY GQQPAPSSTS GSYGSSEQSS
SYGQPQSGSY SQQPSYGGQQ QSYGQQQSYN PPQGYGQQNQ YNS
```

FUS 6E'; chapters 3,4

```
MASNDYTQQA TQSYGAYPEQ PGQGYEQQSE QPYGQQSYSG YEQSTDTSGY GQSSYSSYGQ
EQNTGYGTQS TPQGYGSTGG YGSEQSSQSS YGQQSSYPGY GQQPAPSSTS GSYGSSSQSS
SYGQPQSGSY SQQPSYGGQQ QSYGQQQSYN PPQGYGQQNQ YNS
```

FUS 6E*; chapters 3,4

```
MASNDYEQQA TQSYGAYPTQ PGQGYEQQSS QPYGQQSYSG YSQSTDTSGY GQSSYSSYGQ
SQNTGYGTQS TPQGYGSTGG YGSEQSEQSS YGQQSSYPGY GQQPAPSSTS GSYGSSEQSS
SYGQPQSGSY EQQPSYGGQQ QSYGQQQSYN PPQGYGQQNQ YNS
```

FUS 12E; chapters 3,4

```
MASNDYEQQA EQSYGAYPEQ PGQGYEQQSE QPYGQQSYSG YEQSTDTSGY GQSSYSSYGQ
EQNTGYGEQS TPQGYGSTGG YGSEQSEQSS YGQQSSYPGY GQQPAPSSTS GSYGSSEQSS
SYGQPQSGSY EQQPSYGGQQ QSYGQQQSYN PPQGYGQQNQ YNS
```

FUS40; chapters 3,4

```
MASNDYTQQA TQSYGAYPTQ PGQGYSQQSS QPYGQQSYSG
```

[FUS40]$_2$; chapters 3,4

```
MASNDYTQQA TQSYGAYPTQ PGQGYSQQSS QPYGQQSYSG MASNDYTQQA TQSYGAYPTQ
PGQGYSQQSS QPYGQQSYSG
```

[FUS40]$_3$; chapters 3,4

MASNDYTQQA TQSYGAYPTQ PGQGYSQQSS QPYGQQSYSG MASNDYTQQA TQSYGAYPTQ

PGQGYSQQSS QPYGQQSYSG MASNDYTQQA TQSYGAYPTQ PGQGYSQQSS QPYGQQSYSG

[FUS40]$_4$; chapters 3,4

MASNDYTQQA TQSYGAYPTQ PGQGYSQQSS QPYGQQSYSG MASNDYTQQA TQSYGAYPTQ

PGQGYSQQSS QPYGQQSYSG MASNDYTQQA TQSYGAYPTQ PGQGYSQQSS QPYGQQSYSG

MASNDYTQQA TQSYGAYPTQ PGQGYSQQSS QPYGQQSYSG

[FUS40]$_5$; chapters 3,4

MASNDYTQQA TQSYGAYPTQ PGQGYSQQSS QPYGQQSYSG MASNDYTQQA TQSYGAYPTQ

PGQGYSQQSS QPYGQQSYSG MASNDYTQQA TQSYGAYPTQ PGQGYSQQSS QPYGQQSYSG

MASNDYTQQA TQSYGAYPTQ PGQGYSQQSS QPYGQQSYSG MASNDYTQQA TQSYGAYPTQ

PGQGYSQQSS QPYGQQSYSG

FUS YtoF; chapter 4

MASNDFTQQA TQSFGAFPTQ PGQGFSQQSS QPFGQQSFSG FSQSTDTSGF GQSSFSSFGQ

SQNTGFGTQS TPQGFGSTGG FGSSQSSQSS FGQQSSFPGF GQQPAPSSTS GSFGSSSQSS

SFGQPQSGSF SQQPSFGGQQ QSFGQQQSFN PPQGFGQQNQ FNS

hnRNPA2 CTD WT; chapter 4

GRGGNFGFGD SRGGGGNFGP GPGSNFRGGS DGYGSGRGFG DGYNGYGGGP GGGNFGGSPG

YGGGRGGYGG GGPGYGNQGG GYGGGYDNYG GGNYGSGNYN DFGNYNQQPS NYGPMKSGNF

GGSRNMGGPY GGGNYGPGGS GGSGGYGGRS RY

hnRNPA2 CTD D290V; chapter 4

GRGGNFGFGD SRGGGGNFGP GPGSNFRGGS DGYGSGRGFG DGYNGYGGGP GGGNFGGSPG

YGGGRGGYGG GGPGYGNQGG GYGGGYDNYG GGNYGSGNYN VFGNYNQQPS NYGPMKSGNF

GGSRNMGGPY GGGNYGPGGS GGSGGYGGRS RY

hnRNPA2 CTD P298L; chapter 4

GRGGNFGFGD SRGGGGNFGP GPGSNFRGGS DGYGSGRGFG DGYNGYGGGP GGGNFGGSPG

YGGGRGGYGG GGPGYGNQGG GYGGGYDNYG GGNYGSGNYN DFGNYNQQLS NYGPMKSGNF

GGSRNMGGPY GGGNYGPGGS GGSGGYGGRS RY

LAF-1 IDR WT; chapter 3,4

MESNQSNNGG SGNAALNRGG RYVPPHLRGG DGGAAAAASA GGDDRRGGAG GGGYRRGGGN

SGGGGGGGYD RGYNDNRDDR DNRGGSGGYG RDRNYEDRGY NGGGGGGGNR GYNNNRGGGG

GGYNRQDRGD GGSSNFSRGG YNNRDEGSDN RGSGRSYNND RRDNGGDG

LAF-1 IDR P24G/P25G; chapter 4

MESNQSNNGG SGNAALNRGG RYVGGHLRGG DGGAAAAASA GGDDRRGGAG GGGYRRGGGN

SGGGGGGGYD RGYNDNRDDR DNRGGSGGYG RDRNYEDRGY NGGGGGGGNR GYNNNRGGGG

GGYNRQDRGD GGSSNFSRGG YNNRDEGSDN RGSGRSYNND RRDNGGDG

LAF-1 IDR (scramble 21-28); chapter 4

RMESNQSNNG GSGNAALNRG GYGGDGGAAA AASAGGDDRR GGVAGGGGYR RGGGNSGGGG

GGGYDRPGYN DNRDDRDNRG GSGGYGRDRN YEDRPGYNGG GGGGGNRGYN NNRGGGGGGH

YNRQDRGDGG SSNFSRGGYN NRLDEGSDNR GSGRSYNNDR RDNGGRDG

LAF-1 Shuffle; chapter 4

AGLNYGSDGG YNGDNAHGGN GRNGGNGRDR YYRRNRYRGG GGGERNRGDN GGNGNPGRGG

RNGAGSSRGG NGSGGQEAGGA YGGDVRGDDY GFGDGNNNDY QGASRGRGDR SGNGGGRDGG

SARGGRRNGD PGDSGNYSAG GRRNREDSGL GASDYGDDRG MYSGNNGN

TDP-43 CTD WT; chapter 4

GRFGGNPGGF GNQGGFGNSR GGGAGLGNNQ GSNMGGGMNF GAFSINPAMM AAAQAALQSS

WGMMGMLASQ QNQSGPSGNN QNQGNMQREP NQAFGSGNNS YSGSNSGAAI GWGSASNAGS

GSGFNGGFGS SMDSKSSGWG M

SV series[37, 79]; chapter 4

sv1: EKEKEKEKEK EKEKEKEKEK EKEKEKEKEK EKEKEKEKEK EKEKEKEKEK

sv2: EEEKKKEEEK KKEEEKKKEE EKKKEEEKKK EEEKKKEEEK KKEEEKKKEK

sv3: KEKKKEKKEE KKEEKEKEKE KEEKKKEEKE KEKEKKKEEK EKEEKKEEEE

sv4: KEKEKKEEKE KKEEEKKEKE KEKKKEEKKK EEKEEKKEEK KKEEKEEEKE

sv5: KEKEEKEKKK EEEEKEKKKK EEKEKEKEKE EKKEEKKKKE EKEEKEKEKE

sv6: EEEKKEKKEE KEEKKEKKEK EEEKKKEKEE KKEEEKKKEK EEEEKKKKEK

sv7: EEEEKKKKEE EEKKKKEEEE KKKKEEEEKK KKEEEEKKKK EEEEKKKKEK

sv8: KKKKEEEEKK KKEEEEKKKK EEEEKKKKEE EEKKKKEEEE KKKKEEEEKE

sv9: EEKKEEEKEK EKEEEEEKKE KKEKKEKKKE EKEKEKKKEK KKKEKEEEKE

sv10: EKKKKKKEEK KKEEEEEKKK EEEKKKEKKE EKEKEEKEKK EKKEEKEEEE

sv11: EKEKKKKKEE EKKEKEEEEK EEEEKKKKKE KEEEKEEKKE EKEKKKEEKK

sv12: EKKEEEEEEK EKKEEEEKEK EKKEKEEKEK KEKKKEKKEE EKEKKKKEKK sv13: KEKKKEKEKK EKKKEEEKKK EEEKEKKKEE KKEKKEKKEE EEEEEKEEKE

sv14: EKKEKEEKEE EEKKKKKEEK EKKEKKKKEK KKKKEEEEEE KEEKEKEKEE

sv15: KKEKKEKKKE KKEKKEEEKE KEKKEKKKKE KEKKEEEEEE EEKEEKKEEE

sv16: EKEKEEKKKE EKKKKEKKEK EEKKEKEKEK KEEEEEEEEE KEKKEKKKKE

sv17: EKEKKKKKKE KEKKKKEKEK KEKKEKEEEK EEKEKEKKEE KKEEEEEEEE

sv18: KEEKKEEEEE EEKEEKKKKK EKKKEKKEEE KKKEEKKKEE EEEEKKKKEK

sv19: EEEEEKKKKK EEEEEKKKKK EEEEEKKKKK EEEEEKKKKK EEEEEKKKKK

sv20: EEKEEEEEEK EEEKEEKKEE EKEKKEKKEK EEKKEKKKKK KKKKKKKEEE

sv21: EEEEEEEEEK EKKKKKEKEE KKKKKKEKKE KKKKEKKEEE EEEKEEEKKK

sv22: KEEEEKEEKE EKKKKEKEEK EKKKKKKKKK KKKEKKEEEE EEEEKEKEEE

sv23: EEEEEKEEEE EEEEEEEKEE KEKKKKKKEK KKKKKKEKEK KKKEKKEEKK

sv24: EEEEKEEEEE KEEEEEEEEE EEEKKKEEKK KKKEKKKKKK KEKKKKKKKK

sv25: EEEEEEEEEE EKEEEEKEEK EEKEKKKKKK KKKKKKKKKK KKEEKKEEKE

sv26: KEEEEEEEKE EKEEEEEEEE EKEEEEKEEK KKKKKKKKKK KKKKKKKKKE

sv27: KKEKKKEKKE EEEEEEEEEE EEEEEEEEEK EEKKKKKKKK KKKKKKKEKK

sv28: EKKKKKKKKK KKKKKKKKKK KKEEEEEEEE EEEEEEEEEE KKEEEEEEKEK

sv29: KEEEEKEEEE EEEEEEEEEE EEEEEEEKKK KKKKKKKKKK KKKKKKKKKK

sv30: EEEEEEEEEE EEEEEEEEEE EEEEEKKKKK KKKKKKKKKK KKKKKKKKKK

CspTm; chapter 5

MRGKVKWFDS KKGYGFITKD EGGDVFVHWS AIEMEGFKTL KEGQVVEFEI QEGKKGPQAA
HVKVVE

HIV Integrase; chapter 5

CFLDGIDKAQ EEHEKYHSNW RAMASDFNLP PVVAKGIVAS CDKCQLKGEA MHGQVDC

λ-repressor; chapter 5

GPCLTQEQLE DARRLKAIYE KKKNELGLSQ ESVADKMGMG QSGVGALFNG INALNAYNAA
LLAKILKVSV EEFSPSIARE CR

ProTαC; chapter 5

CEEGGEEEEE EEEGDGEEED GDEDEEAESA TGKRAAEDDE DDDVDTKKQK TDEDC

ProTαN; chapter 5

CDAAVDTSSE ITTKDLKEKK EVVEEAENGR DAPANGNAEN EENGEQEADN EVCEEC

hnRNPA2$_{190-233}$ unmodified; chapter 6

GRGGNFGFGD SRGGGGNFGP GPGSNFRGGS DGYGSGRGFG DGYN

hnRNPA2$_{190-233}$ ADMA; chapter 6 *Dimethylated residues highlighted in red.

GRGGNFGFGD SRGGGGNFGP GPGSNFRGGS DGYGSGRGFG DGYN

FUS$_{120-163}$; chapter 6

SSYGQPQSGS YSQQPSYGGQ QQSYGQQQSY NPPQGYGQQN QYNS

QC sequences; chapter 5

Table A.1: Sequences from Garcia-Quiroz et al.[26] and labels used for chapter 5, (Quiroz-Chilkoti/QC sequences). Groups 1, 2 and 3 correspond to sequences which undergo LCST, UCST and no phase separation respectively.

Name	Length	Sequence	Group	Name	Length	Sequence	Group
QC1	150	$[\text{AVPGVG}]_{25}$	1	QC2	390	$[\text{TVPGVG}]_{65}$	1
QC3	330	$[\text{TVPGAG}]_{55}$	3	QC4	180	$[\text{GVPGAV}]_{30}$	1
QC5	300	$[\text{GVPGVA}]_{50}$	1	QC6	120	$[\text{VAPGVG}]_{20}$	3
QC7	225	$[\text{APGVG}]_{45}$	3	QC8	175	$[\text{VPGVA}]_{35}$	1
QC9	150	$[\text{VPGVG}]_{30}$	1	QC10	125	$[\text{VHPGVG}]_{25}$	1
QC11	175	$[\text{VAPVG}]_{35}$	1	QC12	150	$[\text{VGPVG}]_{30}$	1
QC13	150	$[\text{VPAGVG}]_{25}$	1	QC14	240	$[\text{VPTGVG}]_{40}$	1
QC15	210	$[\text{APVGVG}]_{35}$	1	QC16	125	$[\text{VRPVG}]_{25}$	3
QC17	240	$[\text{APVGLG}]_{40}$	1	QC18	225	$[\text{VPAVG}]_{45}$	1
QC19	200	$[\text{VPHVG}]_{40}$	1	QC20	120	$[\text{VGPAVG}]_{20}$	1
QC21	150	$[\text{VTPAVG}]_{25}$	1	QC22	180	$[\text{TPVAVG}]_{30}$	1
QC23	189	$[\text{VPSALYGVG}]_{21}$	1	QC24	320	$[\text{GRGNSPYG}]_{40}$	2
QC25	448	$[\text{RGDSPHG}]_{64}$	2	QC26	160	$[\text{GRGDSPYG}]_{20}$	2
QC27	160	$[\text{GRDGSPYG}]_{20}$	2	QC28	168	$[\text{RGDSPYG}]_{24}$	2
QC29	160	$[\text{GRGDSPFG}]_{20}$	2	QC30	160	$[\text{GRGESPYG}]_{20}$	2
QC31	192	$[\text{RGDSPYQG}]_{24}$	2	QC32	224	$[\text{RGDAPYQG}]_{28}$	2
QC33	192	$[\text{QYPSDGRG}]_{24}$	2	QC34	140	$[\text{RGDSYPG}]_{20}$	2
QC35	320	$[\text{VPHSRNGG}]_{40}$	3	QC36	108	$[\text{VPSTDYGVG}]_{12}$	2
QC37	160	$[\text{GRPSDSYG}]_{20}$	1	QC38	261	$[\text{VPSDDYGVG}]_{29}$	3
QC39	180	$[\text{VPSDDYGQG}]_{20}$	3				

Appendix B

Validation of SLipids force field with TIP4P/2005 water

Derivation of SLipids parameters are performed in a way that is compatible with most FF Amber in general, and authors have presented that their test of compatibility with Amber03 in particular both for DOPC[209] and DOPS[210]. The compatibility tests presented by the authors of SLipids FF uses the combination of TIP3P water model with the lipids. As we used the TIP4P/2005 water model in this work, we presented below our compatibility tests with TIP4P/2005 in comparison with TIP3P work and experimental findings. These tests were performed for POPC bilayer composed of 16 lipids per leaflet and 1370 TIP4P/2005 water molecules. Analysis was performed on 40 ns serial production simulations which are ran in an isothermal-isobaric ensemble (T=300 K, P=1 bar) following the same methods described in the main text. We found that all structural parameters calculated for POPC bilayer in TIP4P/2005 environment agrees reasonably well with the experimental findings (Table B.1, Figures B.1 and B.2).

Table B.1: Comparison of structural properties of POPC bilayer from simulations and experiments. Experimental DB[356], 2DC[357] and area per lipid[357] measured at 303 K, TIP3P results[209] collected for 303K and TIP4P/2005 results (this study) collected at 300K. Error estimates on TIP4P/2005 simulation results are calculated standard error of the mean using two equal non-overlapping divisions of the simulation data

	Experiment	TIP3P	TIP4P/2005
D_B (Luzzati Thickness) [nm]	3.68	3.85	3.50±0.07
$2D_C$ (Hydrophobic Thickness) [nm]	2.88±0.06	2.84	2.57±0.11
Area per lipid [nm^2]	0.643±0.013	0.646±0.004	0.647±0.003

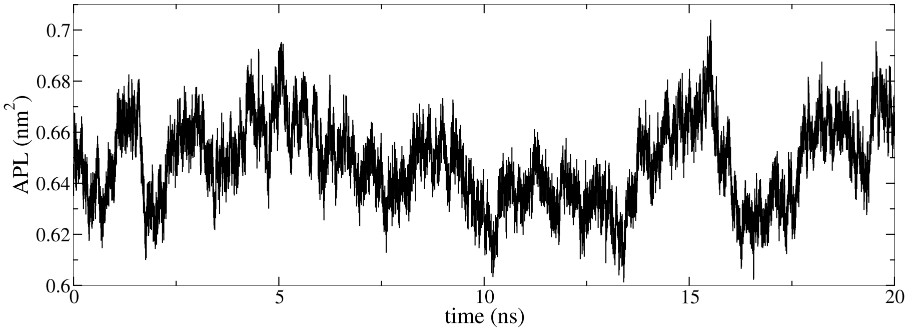

Figure B.1: Sampling of area per lipid with respect to time.

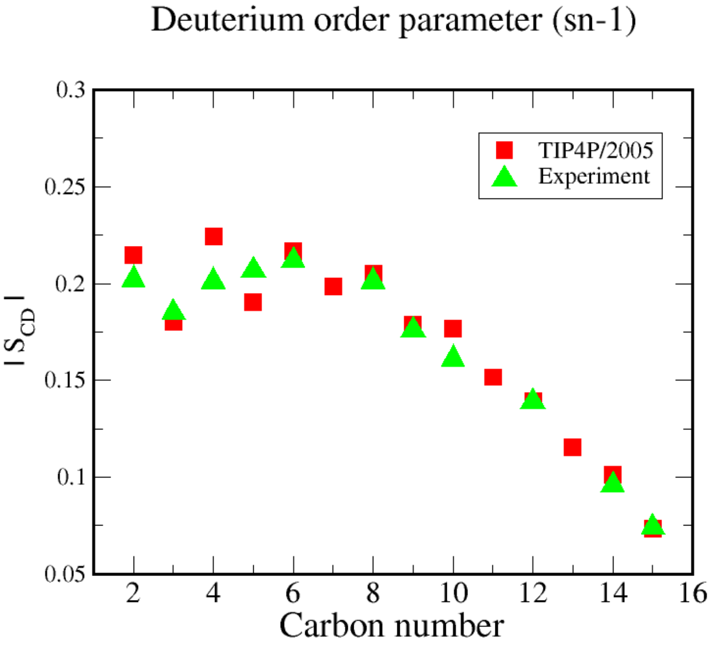

Figure B.2: NMR deuterium order parameter sn-1 quantified for POPC bilayer. Experimental data is reproduced from the work by Seelig and Waespe-Sarcevic[358].

Biography

Gregory Dignon was born on the 22nd of June, 1993 to Frederick and Nancy Dignon in Schenectady, New York. He completed a 2-year associates degree at Hudson Valley Community College (HVCC) while cross-registering for several courses at Rensselaer Polytechnic Institute (RPI). He then began attending RPI full-time and completed a Bachelor's Degree in Chemical Engineering in May of 2015. He then joined Lehigh University the following Fall, and began research in the lab of Professor Jeetain Mittal later in the semester. Throughout his undergraduate career, Gregory had the opportunity to work in the lab of Professor Peter Tessier during the semester, and over the summer in 2014, doing research on antibody self-interactions. He also also ran his own side-business as a private music teacher during his time at HVCC and RPI. Throughout his time at Lehigh, Gregory has published 8 papers in peer-reviewed journals, with 5 more unpublished manuscripts in various stages of completion, and attended 15 academic conferences and symposia to give a total of 9 oral presentations , and 12 poster presentations. In the Fall of 2017, Gregory received the Chevron Scholarship, which covered membership fees to AICHE for the next 5 years, and provided a sum of money to be used for travel to the AICHE annual meeting that year. The following year, Gregory was awarded the John C. Chen endowed fellowship, in recognition of productivity in research, and academic leadership. Gregory is also an accomplished musician, with professional-level proficiency with guitar and drums.