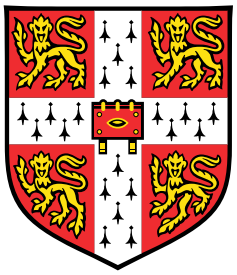
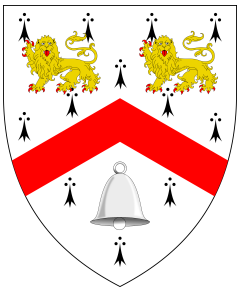


Multiscale coarse-grained models for biological phase separation

Development and applications



Ane Aguirre Gonzalez

Supervisor: Prof. Rosana Collepardo Guevara

Advisor: Prof. Jerelle A. Joseph

Department of Chemistry

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

Wolfson College

September 2023

To Edurne and my dear *amona*.

Nire bideko itsasargi eta ezkutuko zaindari,
elkarrekin mundua aldatzen ikusi dugulako
betirako ez garelako,
zuen bihotzen zati bat
eramango dudanaren itxaropenenean,
tesi hau zuengatik eta zuentzat da.

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text. It is not substantially the same as any work that has already been submitted, or, is being concurrently submitted, for any degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Ane Aguirre Gonzalez

September 2023

Abstract

Many proteins undergo liquid–liquid phase separation (LLPS) in order to generate membraneless organelles, by which the cell can organise its biomolecules and bioprocesses dynamically in space and time. Many of the experimental and theoretical methodologies used to study the formation of these condensates still struggles to capture the relation between microscopic, residue–level details of a protein with its phase behaviour in the bulk. In this thesis, we present three multiscale coarse-grained models with amino acid resolution aimed at studying phase separation of proteins.

The Mpipi model balances the dominant role of π–π and hybrid cation–π/π–π interactions with the rest of the interactions, while keeping R–based interactions stronger than K–based ones. The parameterisation is based on atomistic PMF calculations and bioinformatics data on π-based contacts. The Mpipi Recharged model is a finer and more optimised version of Mpipi, that appropriately balances the electrostatic interactions on a pair-by-pair basis, since which all-atom simulations prove the asymmetry between samely-charged and oppositely-charged residues. Both Mpipi and Mpipi recharged are capable of reproducing ensemble averaged experimental observables with high accuracy, from single-molecule properties to phase diagrams of an extensive set of proteins (*i.e*. hnRNPA1, FUS, *Laf*1, DDX4) and their corresponding mutations.

Lastly, we also investigated the role of Mg^{2+}ions in regulating LLPS of intranuclear proteins. Atomistic-resolution simulations proved that minimal quantities of Mg^{2+}ions present in the media can significantly alter the phase behaviour of proteins, especially ones with a high number of charged residues. The MagPi model arises from this observations and bioinformatics analysis of the proteome, and can reproduce the phase behaviour of a set of intranuclear proteins (*i.e*. MED1 IDR, BRD4 IDR, Nanog CTD, and DDX4 and DDX3 variants) qualitatively.

Overall, our multiscale modelling approach shows great potential at bridging the gap between atom-level observables, to single-molecule behaviour, to macroscopic phase transitions, as well as its ability to extend the range of the simulations to different solvent conditions or surroundings. Therefore, the work presented in this thesis poses a significant step towards the unification of experiments, computer simulations and real biological LLPS phenomena.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Prof. Rosana Collepardo-Guevara, for their unwavering support, guidance, and mentorship throughout the course of my PhD journey. Your expertise, patience, and encouragement have been invaluable to me, and I am genuinely grateful for the opportunity to learn and grow under your mentorship.

I want to extend my sincerest appreciation to my co-supervisor, Professor Jerelle A. Joseph, for their invaluable guidance, wisdom, and unwavering dedication. Their support has been instrumental in propelling my research forward, and their unwavering belief in my potential and commitment to my growth as a researcher has been a constant source of motivation and inspiration for me.

I would also like to extend my gratitude to Prof. Jorge Espinosa for his comprehensive feedback and his encouragement to think outside the box, as well as Dr. Andres Tejedor and Prof. Aleks Reinhardt for their time and effort in our collaborations.

My sincere thanks go out to my colleagues in the Collepardo group, especially my bestie Julia, Jan, Pin Yu, Kieran and Rob. The collaborative spirit, shared knowledge, and mutual support in moments of stress have made my time in the group both productive and enjoyable.

I gratefully acknowledge the Yusuf Hamied Department of Chemistry and the University of Cambridge for fostering an environment of academic excellence. I also wish to express my gratitude to European Research Committee for the financial support that has been crucial to my research.

Last but by no means least, I owe a deep debt of gratitude to my family. To my *ama* and *aita*, Mirari and Pedro, and my *amona*, for their unwavering belief in me, for instilling in me the values of hard work and perseverance, and for their sacrifices that paved the way for my career. To Idoia and Kai, for their constant encouragement and always being there to lift my spirits.

And to Galbi, my anchor and confidant, thank you for your meow-meows, sniffs, purrs, head bonks, and for standing by me every step of the way. You wanted treats and chicken, but your presence was all I needed to push forward.

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

ALS	Amyotrophic lateral sclerosis
CG	Coarse grained
COM	Center of mass
CS	Charge scrambled
CV	Collective variable
FUS	Fused in Sarcoma
LLPS	Liquid–liquid phase separation
MD	Molecular Dynamics
ODE	Ordinary differential equation
PDB	Protein Data Bank
PD	Parkinson's Disease
PMF	Potential of mean force
PTM	Post-translational modification
REMD	Replica Exchange Molecular Dynamics

TI Thermodynamic Integration

US Umbrella Sampling

WF Wang–Frenkel

WHAM Weighted Histogram Analysis Method

Chapter 1

Introduction

Inside every Eukaryotic cell, there is a complex blend of many biomolecules and are faced with the challenge of organizing their numerous components in space and time in order to manage various interconnected biochemical processes. Accordingly, different compartments are created, each with distinct chemical compositions and functions. While some of these compartments are separated from the surrounding intracellular environment by lipid membranes, the vast majority are membraneless. Research on membraneless compartments –more widely known as biomolecular condensates- has exploded in the past half-decade, attracting significant interest from experimentalists, computer simulators, and theoreticians alike. The formation of these compartments relies on a thermodynamic process termed liquid–liquid phase separation (LLPS), caused by the condensation of proteins and often, although not necessarily, nucleic acids into liquid-like droplets. These biomolecular condensates are formed by selecting some molecules to become concentrated while leaving others out. The importance of biomolecular condensates in intracellular compartmentalisation has led to a surge in research, calling for interdisciplinary expertise from the fields of biology, chemistry, and physics. After all, biomolecular condensates, which are comprised of proteins and/or nucleic acids, are not only critical for intracellular compartmentalisation but also sustained and regulated by the thermodynamic laws of phase transition. For this reason, the formation of these compartments can easily be studied beyond the laboratory, using theory and computational simulations.

1.1 Computational simulations of liquid–liquid phase separation

Computer simulations have been vital in understanding the molecular mechanisms that regulate biomolecular condensates. These simulations, conducted at various lengths and time scales, have helped to determine the molecular codes that enable biological phase separation under different conditions and have revealed how the environment can transform the percolating network that sustains the condensates. These simulations provide a detailed view of molecular-level interactions, allowing researchers to systematically vary parameters and design tailored model systems to eventually bridge the gap between residue-level interactions and their phase behaviour in bulk.

1.2 Thesis overview

This thesis delves into the intricate molecular mechanisms of LLPS by developing three distinct coarse-grained (CG) models, each tailored to capture specific facets of the phase separation process. We follow a multiscale approach that consists on carrying out simulations at high–resolution simulations at the per atom level of amino–acid residue pairs in solution in order to build an approximate medium–resolution models or representations of protein systems to study their macromolecular phase behaviour. Through this, these models bridge the gap between atomistic details and the macroscopic behaviour observed in LLPS, providing a comprehensive framework to study the dynamics and interactions that drive phase separation.

The first model, termed Mpipi, focuses on the fundamental interactions between proteins, capturing the relevant relative interaction strengths between amino acids, and putting an emphasis on the role of π-π can cation–π contacts. The second model, which we called Mpipi Recharged, extends this by incorporating a tunable term for electrostatic interactions, which we observe to be asymmetric in atomistic simulations, and play a pivotal role in driving LLPS. The third model, termed MagPi, integrates the effects of magnesium ions in regulating the strength of contacts, and how minimal additions of these ions can alter

the phase behaviour of intranuclear proteins. By seamlessly transitioning between different scales, from individual molecular interactions in atomistic simulations to collective phase behaviour at mesoscopic scales, this thesis offers a holistic understanding of LLPS, paving the way for future computational research and providing methodologies to develop CG models of biocondensates reducing the gap with experimental *in vivo* studies.

Chapter 2

Background theory - study of liquid-liquid phase separation

Liquid-liquid phase separation (LLPS) has emerged as a pivotal concept in the realm of cellular biology, offering profound insights into the spatial organization of cells and the dynamics of biomolecular condensates. This chapter provides a comprehensive review of LLPS in biological systems, exploring its fundamental principles, diverse applications in cellular functions, and its implications in both physiological and pathological contexts. In addition, we also discuss the main experimental techniques used in the field to study this phenomenon and characterise biocondensates, as well as the potential of computer simulations to underpin our understanding of LLPS by bridging observations from atomic to macroscopic scales.

Contents

2.1 Liquid–liquid phase separation in biology

Eukaryotic cells expertly manage thousands of physicochemical processes simultaneously, each within a brief time frame, without any interference or overlap between one another. This is achieved by creating micro-environments that isolate specific biomolecules such as proteins, nucleic acids, and ions, resulting in a highly controlled biochemical environment.

For a long time, the scientific community thought this organisation was attributable to membrane-bound organelles, which separate their internal environment from their surroundings through a lipid membrane. Examples of membrane-bound organelles are the mitochondria, the Golgi apparatus and the endoplasmic reticulum, *etc*. These structures keep their internal chemical compounds away from the cytosol while carrying out their functions with high precision. If the biomolecules leak outside of these organelles in an uncontrolled manner, it could have severe consequences and even be fatal.

The existence of other cellular subcompartments, not bound by lipid membranes that could maintain their own microenvironments, was, for a long time, difficult to reconcile.

However, several other cellular organelles have been found to mimic the functionality of the membrane–bound compartments in the absence of a partitioning layer. Structures like the nucleolus, P granules, Cajal bodies and stress granulles (SGs) are an example of these [116, 126], as shown by underlined organelles in Figure 2.1. Indeed, in many enzymatic catalytic reactions, the participating agents are concentrated and located at certain sites to ensure their availability during the process. The formation of these structures has been found to be due to liquid–liquid phase separation, LLPS [57, 128, 62].

2.2 Physicochemical determinants of biological LLPS

The idea that LLPS is responsible for the development of membrane-less organelles gained popularity after the discovery of the liquid-like characteristics of P granules in *C elegans* [27].

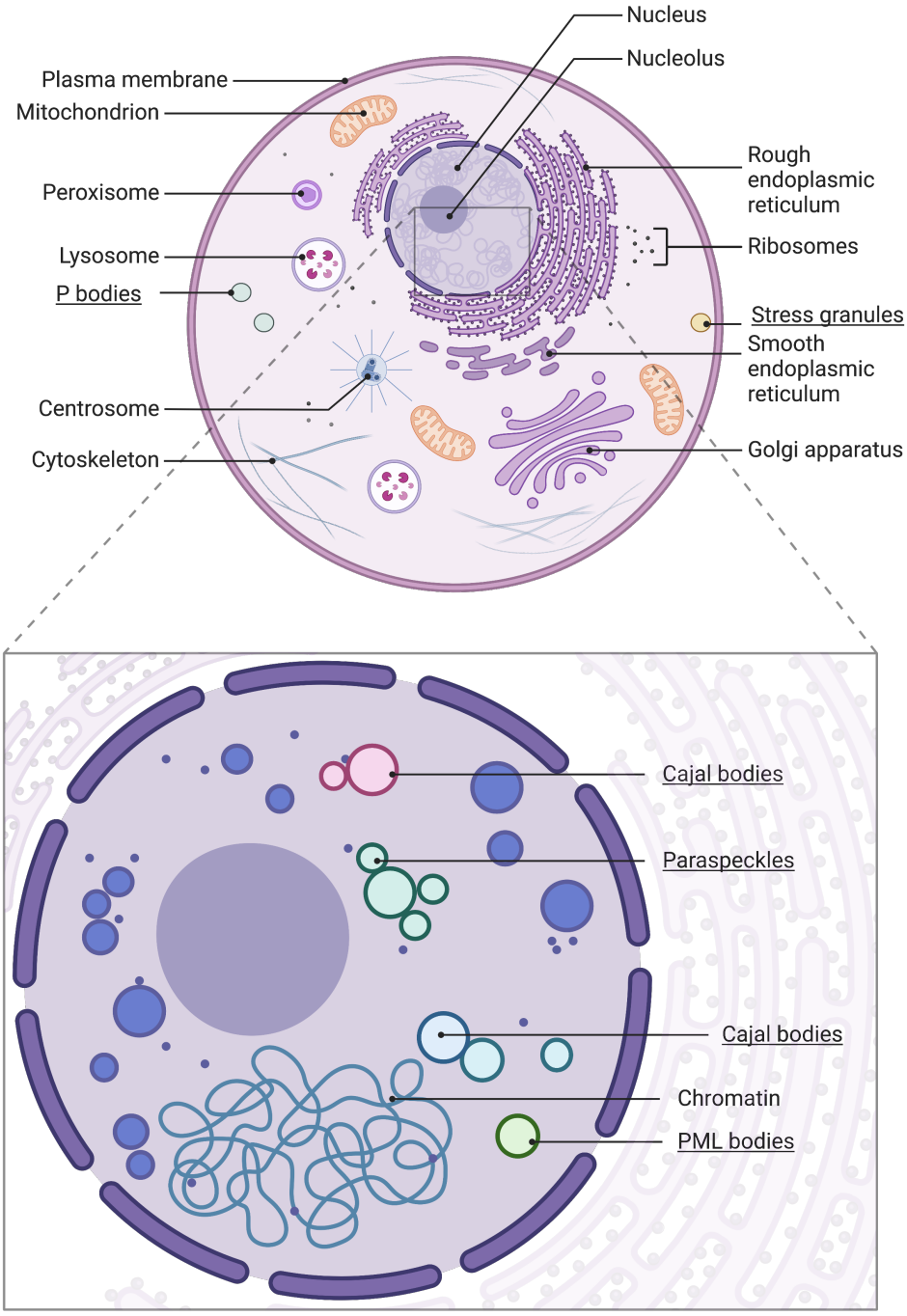


Fig. 2.1 **The Eukaryotic cell is comprised of a large number of organelles that organise the biomolecular processes and participant biomolecules in space and time**. Depicted here is a non-exhaustive list of organelles enclosed in a cell. Organelles with their names underlined are formed via LLPS and, thus, membrane-less.

P granules are composed of RNA and proteins and emerge when their components exceed a critical threshold, forming distinct droplets. This process occurs during zygote division, and the droplets that form eventually dissolve [27].

In both the cytoplasm and nucleus, biomolecular LLPS occurs where the fluid containing proteins and nucleic acids demixes into a dense and lighter phase, forming membrane-less organelles [27, 167]. Although the total concentration of proteins in the two phases combined may not differ, it is expected that the concentration of specific and related proteins will be higher in the dense phase. *In vivo* and *in vitro* experiments have shown that these droplets, most commonly known as biomolecular condensates, are liquid-like bodies with densities noticeably higher than the fluid in the cytoplasm and nucleoplasm [73].

Recent experimental studies [28, 27] have reported that these condensates share similar material properties with liquids: they form spherical droplets *in vivo* and *in vitro*, and as liquids do, they can flow, drip, and wet surfaces. Over the last couple of years, the term 'condensates' has broadened to also include gels, glasses and solids forming from LLPS. These characteristics were known a long time ago for nucleoli, which were observed to fuse together into a single body [28]. Their surface tension increases linearly with droplet size, as seen in liquids [27], and their structures are highly dynamic and constantly maintained at steady–state [116]. In fact, these condensates can exchange molecules with the surrounding nucleoplasm or cytosol, without a specific transfer mechanism, like the ones needed by membrane–bound organelles [116].

2.2.1 Thermodynamics of LLPS

From a thermodynamics perspective, LLPS of biomolecular solutions is a result of an interplay between enthalpic and entropic effects. For a well-mixed system to phase separate, the entropy loss stemming from the reduction of the number of microstates available in the demixed system needs to be overcome by the enthalpic gain coming from the associative interactions between biomolecules. The phase separation occurs when the overall Gibbs' free energy, $G = H - TS$, with H being the enthalpy and S being the entropy at temperature T, of the system is reduced [65].

Consider a hypothetical scenario where two solutions of two different types of molecules, A and B, are mixed together at constant volume. Upon mixing, it becomes evident that the intermolecular forces between similar molecules (A–A and B–B), also known as homotypic interactions, are stronger in comparison to those between different molecules (A–B), or heterotypic interactions. Consequently, the energy required to break down the A–A and B–B interactions, is higher than the energy gained by forming A-B interactions, and demixing is favoured [48]. Demixing leads to a free energy curve with two distinct basins (Figure 2.2, right panel), each one corresponding to states with different volume fractions, ϕ_L and ϕ_D, corresponding to a low-density phase and a dense phase, respectively. In this case, the free energy is minimised by demixing the systems into a dense phase and a dilute phase.

When A–A and B–B interactions do not overcome the entropy of mixing and A–B interactions, the free energy change curve with respect to the volume fraction is of concave shape, and the system is well-mixed (Figure 2.2, left panel). In a standard system, because the entropic contribution is weighed by the temperature, whether a solution would phase separate or not is linked to temperature. Once the system temperature surpasses the upper critical solution temperature (UCST), it becomes thermodynamically unfavourable for the mixture to undergo LLPS [93]. As a result, the system can only exist in a single mixed state.

The simplest biomolecular condensates can be considered as single-phase protein droplets surrounded by and in coexistence with pure water. The two phases are said to coexist when the chemical potential μ, pressure P and temperature T are equal in both phases. The conditions that satisfy such equality make up what is known in the field as a binodal curve.

The binodal curve describes the relationship between the composition of the system and the conditions (usually temperature, pH or salt concentration) under which phase separation occurs. It consists of two curves representing the boundaries between the two distinct phases, known as the coexisting liquid phases. As the binodal curve shows (see Figure 2.3), the system will be in a homogeneous mixed state at the conditions outside of the binodal dome.

Further aside from the nucleation regime, there is another regime where the system becomes significantly unstable. This occurs at the inflexion points of the Gibbs' free energy, when $\frac{d^2G}{d\phi^2} = 0$, and gives place to a regime called spinodal decomposition. The coexistence

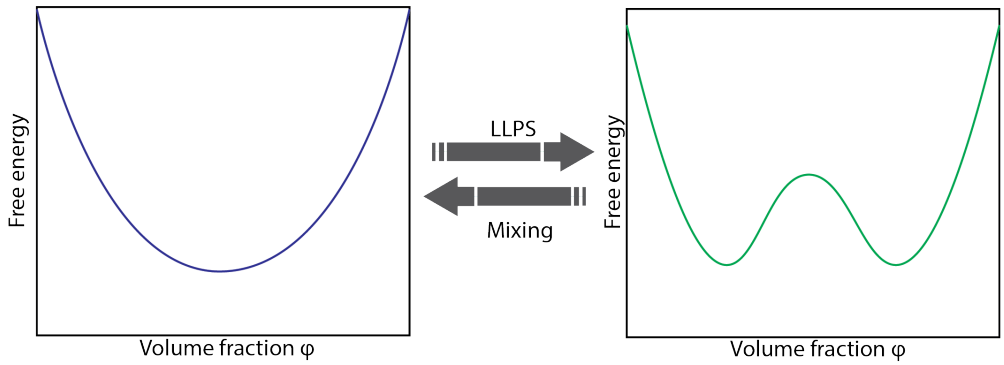


Fig. 2.2 **Comparison of free energy *versus* volume fraction of mixed (left) and demixed (right) systems.**

line between this regime and the nucleation regime is referred to as the spinodal curve, as represented in Figure 2.3.

Other types of systems, on the other hand, might present a lower critical solution temperature (LCST), the minimum temperature at which heating can induce phase separation [93]. Nevertheless, none of the research presented in this thesis pertains to such systems.

Another interesting kind of transitions are the so-called reentrant phase transitions, that emerge when a system that has already undergone one phase transition, goes through another transition to a similar macroscopic state by the alteration of a single condition. A clear example of such behaviour is when the condensates of one biomolecule are regulated by another. For instance, many cellular proteins, which require low concentrations of salt to exist as liquid droplets, reenter a dissolved state upon the addition of high quantities of salt [94].

As mentioned, the binodal curve is the result of a cumulative of contributions, thus, a temperature–dependent binodal curve can have different heights and shapes depending on a number of factors, such as salt concentration, pH, etc, that tune the interaction network and energies of the system's components [94, 105, 109].

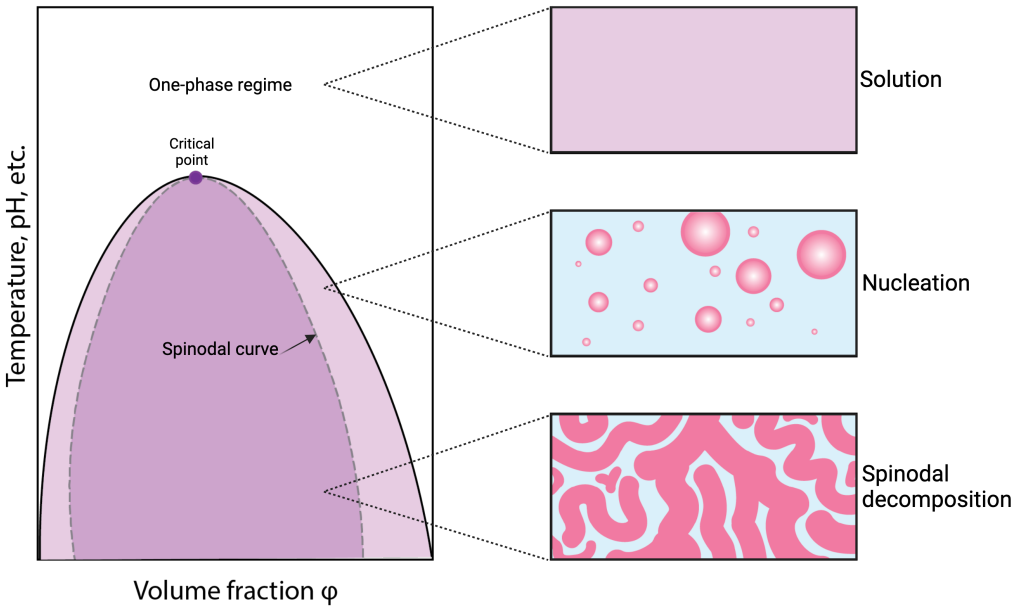


Fig. 2.3 **Phase diagram constructed by varying protein volume fraction versus solution conditions such as pH and temperature**. The solid line on the diagram shows the limit of solubility for molecules, beyond which they become immiscible with the surrounding solution (binodal curve). The dashed line represents the coexistence line, marking the point below which the system enters the spinodal decomposition regime (spinodal curve).

2.2.2 Biological functions of LLPS

The discovery of membraneless compartments has defied what all scientists knew about spatiotemporal organisation in cells for the past several decades. The properties and functions of these kinds of structures and the preferential formation over classic membrane-bound compartments have been a matter of discussion. Ever since the discovery of P granules on *C. elegans* cells [27], the interest of the wide biophysics community has picked up and, over the years, more and more biological functions of intracellular liquid–liquid phase separation have been discovered.

The main biological functions of membrane-less compartments postulated so far can be summarised as follows:

(a) Compartmentalisation of biomolecules: similar to membrane-bound organelles, biocondensates allow the segregation of specific biomolecules (proteins and RNA) from the surrounding solvent [103, 56]. These compartments show a significant concentration of LLPS-driving biomolecules and their high-affinity ligands. It can be argued that generating these compartments is more energy-efficient than forming organelles with physical barriers since they do not require the transport of membrane-building molecules or further chemical reactions to build such structures. Furthermore, these compartments can exchange molecules with the surrounding nucleoplasm or cytosol without the need for specific transfer mechanisms, as required by membrane-bound organelles [116].

(b) Concentration and buffering of specific molecules: the droplets formed by demixing in the cytoplasm or cell nucleus can maintain a highly dynamic exchange of molecules with the surroundings. For this reason, these biocondensates allow for rapid localisation of molecules, such as proteins, nucleic acids, ions or other compounds, to carry out specific chemical reactions [97]. By strategically sequestering in or out specific molecules, biomolecular condensates can efficiently regulate the yield of biochemical reactions.

(c) Stimuli sensing and signalling: recent studies have been uncovering the implication of biomolecular condensates in immune- and stress-related responses. For instance, Cai and colleagues [33] observed that SARS-CoV-2 N proteins bind RNA and protein G3BP1 and phase separate into droplets, blocking the immune response triggered by G3BP1 interacting with the cGAS DNA sensor. Neuronal signalling has also been connected to phase separation, such as the formation of presynaptic densities via the formation of RIM/RIM-BP and PSD-95/SynGAP condensates [188, 181].

(d) Organisation of chromatin: Phase separation plays a key role in the nucleus and, more specifically, in the transmission of genetic information, as many of the bimolecular condensates found inside the cell nucleus, such as transcriptional assemblies [21], actively participate in the structural organisation of DNA. Heterochromatin, which is

a condensed and transcriptionally repressed form of chromatin, can undergo LLPS to form distinct liquid-like droplets within the cellular nucleus. This process leads to the sequestration of heterochromatic regions and the establishment of nuclear domains [162]. The highly compacted structure of chromatin has been associated with condensate formation driven by histone proteins, and further changes such as post-translational modifications (PTMs) like acetylation, and DNA-linker length, regulate the formation of biocondensates or lack of thereof [66].

(e) Transcription regulation: Transcriptional condensates are membraneless compartments that form through LLPS and regulate gene expression. These condensates concentrate transcription factors (TFs), RNA polymerase (Pol), and other regulatory molecules, facilitating the efficient assembly of the transcriptional machinery. They are involved in the spatial and temporal regulation of gene expression. For instance, the formation of transcriptional condensates at enhancer regions promotes the activation of genes by bringing together enhancer elements and gene promoters [154]. Nuclear condensates have been correlated to transcription via RNA Pol II protein [24].

Despite all the cases where LLPS is involved in maintaining the correct functioning of cellular processes, plenty of studies have also depicted its role in cell dysfunction and disease. A widely investigated case of pathological LLPS is the one involving amyloid formation and aggregation. While most phase transitions are reversible, condensates formed through an irreversible process have been shown to disrupt cellular functions, as is the case of common diseases like Alzheimer's, Parkinson's disease (PD), amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD) [186, 172]. These conditions have been associated with aberrant phase separation of proteins Tau, α-Synuclein, Fused in Sarcoma (FUS) and chromosome C9ORF72, respectively [186, 143, 156, 147, 7, 159].

PTMs are thought to be the underlying cause of pathological aggregation of many proteins [190]. For instance, protein Tau can present phosphorylation of specific Serines and Tyrosines, leading to pathologic fibrillisation of the protein and causing Alzheimer's disease [190]. So is also the case of α-Synuclein, which forms the neurotoxic Lewy bodies from high amounts of the protein undergoing phosphorylation at residue S129 [34, 88]. A

distinctive hallmark of Lewy bodies, and therefore, PD, is the ubiquitination of some Lysines in α-Synuclein, which can also be present along small ubiquitin-like modifier (SUMO) protein [150].

Biocondensates have therapeutic potential beyond neuropathies and can be utilized in cancer treatment. In the context of breast cancer, the presence of MED1 condensates in the cell nucleus has been associated with unfavourable prognoses [92]. While pharmacological interventions such as tamoxifen and cisplatin have been found to target the Mediator complex via MED1, an overexpression of MED1 may result in treatment resistance, including tamoxifen resistance, and a less hopeful cancer prognosis.

2.2.3 Physicochemical forces driving LLPS

The exact chemical determinants of phase separation are yet to be understood, and due to the large variety of proteins and characteristics, it seems that there is no one-size-fits-all solution or unique framework encompassing the grammar regulating LLPS to this day. However, many advancements have been made to determine some fundamental interactions that participate in such a complex process.

One of the central principles that are believed to drive the behaviour of biomolecules, especially proteins, at both the single–molecule level and the macroscopic level is that sequence determines the function and behaviour of proteins. Following this logic, the sequence of a protein determines the chemical interactions between residues and the conformational structure the protein can adopt [9]. In this vain, one could postulate that detecting the most LLPS–driving residues would be enough to predict the phase behaviour of a protein. Some simplified models following this assumption include the 'stickers and spacers' model from Choi *et al*. [36].

However, amino acid composition alone has proven insufficient to predict protein function and, subsequently, phase behaviour. It is possible, for instance, for proteins to have a high identity score, yet show different 3D conformations and biological function [75, 4]. In fact, more so than residue composition, the sequence itself is the most relevant factor determining the conformation of a protein. The final conformation a protein can adopt, however, is

highly dependent on the conditions of its surroundings as well. On the other hand, it is also remarkable that vastly different sequences can fold to fundamentally the same conformation, such is the case of the lysozyme protein family [13].

The presence of folded or globular domains plays an important role in the formation of MLOs and biomolecular condensates at physiological conditions. These can bind specific binding partners and drive phase separation. For example, Li *et al*. [103] observed that folded SH3 and PRM domains threaded together by inert and flexible regions form liquid-like droplets due to the interactions between the two folded domains.

Proteins containing regions with no defined three-dimensional structure have puzzled scientists studying protein aggregation and phase separation alike. In spite of their disordered nature, these intrinsically disordered regions or proteins (IDRs or IDPs) can present repeated blocks of residue motifs responsible for the weak multivalent attractive interactions between molecules. This is the case for many condensates like stress granules and P bodies. Some of these residues include Glycine (G), Serine (S), Phenylalanine (F), Asparagine (N) and Tyrosine (Y) [87, 72]. IDRs in nuclear biocondensates do also present charged residues - lysine (K), arginine (R), aspartic acid (D) and glutamic acid (E) [14].

Phase separation is caused by a complex interplay between different types of interactions and is significantly context-dependent. So far, some of these kinds of interactions involve cation–π, π–sp^2 and π–π stacking contacts (see Figure 2.4) [170, 130]. For instance, this can be seen by mutating the aromatic residues in the intrinsically disordered nephrin intracellular domain (NICD), where such mutations lead to decreased ability to form droplets [139]. In DEAD-box Helicase 4 (Ddx4), electrostatic interactions between clusters of opposing charge and cation-π interactions between phenylalanine-rich and arginine-rich motifs have been invoked as key phase-separation driving interactions [137]. Dipolar interactions have also been accounted as key role players in driving phase separation in proteins rich in polar residues - S, Q and N - [82, 139].

Experimental work has revealed that even though biomolecular condensates can contain many different proteins, not every one of them is necessary to drive phase separation [47, 81, 146]. In some cases, proteins that drive phase separation bind to other proteins and recruit

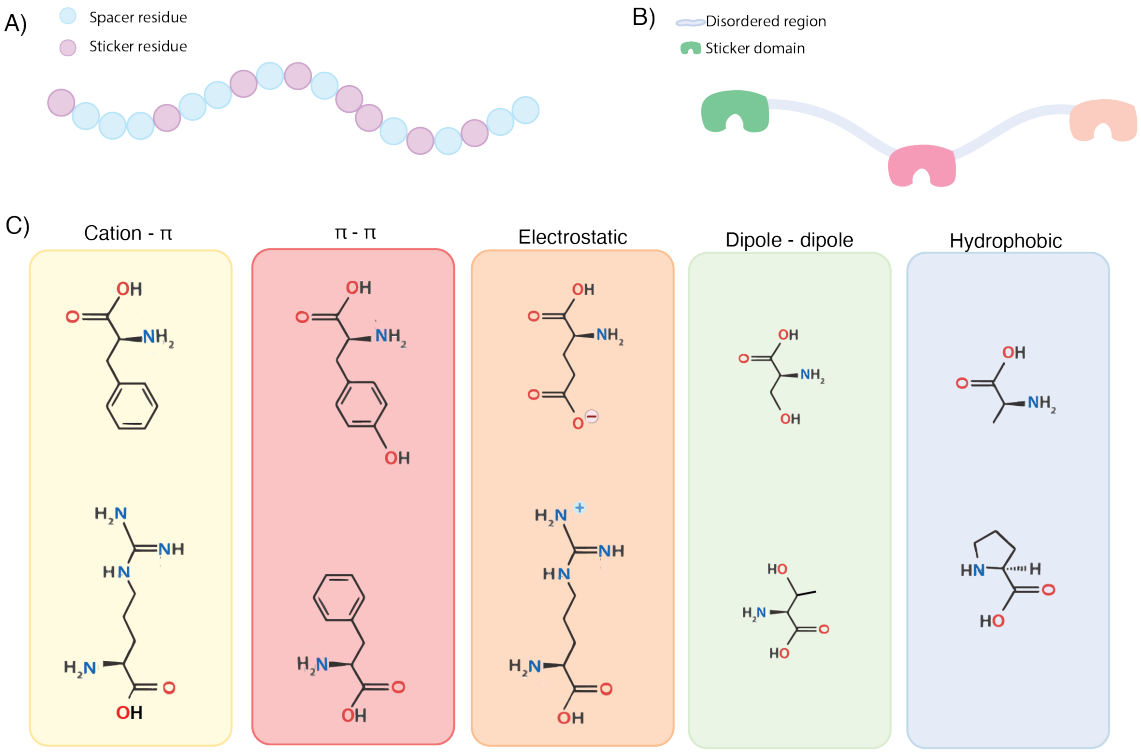


Fig. 2.4 **Molecular grammar driving LLPS**. (A) Peptide chain containing adhesive 'sticker' residues, with 'spacer' residues interswept in-between. (B) Peptide chain with intrinsically disordered (light blue) and globular (coloured) domains. (C) Chemical structure of amino acids that are often involved in the interaction between IDRs, and the modes of interaction each residue can have, categorised by type.

them into the newly formed dense phase. For instance, recent work by Banani *et al*. [15] demonstrated that droplets formed by the assembly of polySIM and polySUMO proteins also attracted SUMO and SIM monomers into them [15]. Components that are necessary to drive LLPS are termed 'scaffolds', and the ones that, although non–essential, are recruited into the droplets, are named 'clients', as proposed by Banani et al [15]. Clients are dispensable for LLPS, but recruited into the biomolecular condensates due to their interactions with scaffolds, such as the small RNA molecules recruited into DDX4 droplets [136].

Scientists have traditionally studied protein liquid-liquid phase separation and protein aggregation as two separate processes. However, both are driven by interactions between biomolecules, both within and between molecules. Liquid-liquid phase separation results

in dynamic liquid compartments, while aggregation results in the irreversible clumping of proteins into large, insoluble complexes. Interestingly, evidence shows that the droplet ageing–mechanism does, in some cases, overlap both phenomena [133].

The compositional regulation of these droplets is based on parameters like the binding affinities between proteins, their stoichiometric ratios, and the valencies [47, 3], and environmental factors such as pH, temperature and pressure [94, 109, 105].

2.3 Experimental characterisation of LLPS

To understand the intricate mechanisms underlying LLPS, scientists employ different experimental methodologies which can be categorised as either *in vivo* or *in vitro* methodologies. In the former, the experiments are carried out 'inside' living systems, more commonly in cells. On the latter, the experiments are usually carried out in solution or suspensions.

The usual first approach to study a system that undergoes LLPS is to reconstitute its essential components *in vitro*, which can be achieved with as easy as a sedimentation assay. The supernatant recovered after centrifugation of a mixture can contain the liquid and non-aggregated droplets of the phase-separating components [188], and turbidity measurements can assist in quantifying the degree of phase separation [121, 7].

The addition of fluorophore agents (i.e.: GFP, mCherry tags) to phase-separating proteins, joined with imaging techniques, aids in individually identifying, localising and measuring the liquid droplets. When rigorous controls are in place, these techniques can provide a wide array of information about the system studied. However, the addition of a fluorescent tag can significantly affect the phase behaviour of the system [187] and even disrupt LLPS [3].

Methods like fluorescence recovery after photobleaching (FRAP) and atomic force microscopy (AFM) are able to provide information about the kinetics and material properties of the condensates, such as diffusivity and viscosity [12, 143]. The recovery of fluorescence depends on diffusive motion across different phases and associative and dissociative contacts; therefore, the measures captured by FRAP contain a high error. Although they could provide

reasonably approximated calculations, one should be aware of the significant inaccuracy of the method [129].

In addition, many of these condensates, which are initially fluid, see an increase in their viscoelasticity with time, resulting in a reduction in molecular interchange with the solvent [187, 110]. This phenomenon is known as condensate ageing and is been a subject under study for which the underlying causes are still yet to be deciphered.

LLPS has also been investigated through spectroscopic methods such as nuclear magnetic resonance (NMR), X-ray diffraction and circular dicroism (CD) [7, 149, 25]. For instance, it has been observed that biomolecules show different NMR spectra and overall shift in backbone resonance when they form liquid droplets *vs.* are mixed [7].

Cryo-transmission electron microscopy (cryo-TEM) has been used to image coacervate microdroplets with a resolution below the micron range. Nevertheless, given the strict conditions required for the preparation of the sample, cryoTEM only provides information on the final structures stemming from LLPS. An interesting case study is the one of the Rubisco-CcmM protein complex that is involved in the formation of β–carboxysome in cyanobacteria, where Wang and colleagues [173] were able to solve the structure of the complex that gives rise to phase separation by joining cryo-electron microscopy (cryo-EM) and cryo-electron tomography (cryo-ET).

Observing the early stages of LLPS, specifically the formation of collapsed protein nanoclusters, has been difficult until now. *In situ* liquid TEM is an excellent method for this purpose as it enables real-time imaging, facilitating the study of dynamic processes, including droplet formation, dissolution, and rearrangement. For instance, Le Ferrand and colleagues [101] used this technique to study the behaviour of intrinsically disordered Histidine-rich Beak Protein 2 (HBP-2), monitoring the transitions from random coil to oligomers and further to nanoclusters stabilised by β–sheet formation in the early stages of LLPS.

Research has shown that thermodynamic processes play a key role in the creation of membrane-less organelles. Overall, while current experimental techniques can be used to study these processes both in the lab and in living organisms, they have limitations in terms

of their ability to examine molecular-level events and their resulting bulk behaviour of certain proteins. Specifically, they struggle to provide high resolution and insight into the dynamics of IDRs and other proteins that separate into distinct phases.

2.4 Computer-aided study of LLPS

Simulations of varying lengths and time scales have been beneficial in discovering the essential biophysical and molecular mechanisms that regulate biomolecular condensates [98]. These simulations have played a significant role in identifying the molecular codes that facilitate biological phase separation under diverse conditions [174, 29]. Additionally, they have revealed how the percolating[1] network that supports condensates undergoes transformation due to environmental changes [53, 94].

In the field of biomolecular condensates, computer models are confronted with the challenging task of accommodating both collective behaviour and physicochemical diversity. To ensure the accuracy of the models, a delicate balance must be struck between two competing criteria: the inclusion of sufficient details and chemical realism to capture the subtle effects of protein and nucleic acid sequence and structure and the trade-off of some of these details to achieve a high degree of computational efficiency capable of handling a large number of degrees of freedom and long simulation time scales that are relevant to condensates.

When it comes to computer simulations, there are generally two main techniques that are utilized: Molecular Dynamics (MD) and Monte Carlo (MC). MD involves the numerical integration of Newton's equations of motion, which allows monitoring of the configurations of the system through time. More about MD will be covered in Chapter 3. On the other hand, MC involves sampling the energy landscape of a system by randomly generating new microstates or configurations. These new states can be accepted or rejected based on certain

[1]When new connections are added to a network, a phase transition occurs, resulting in a sudden emergence of connectedness in a significant portion of the network. This phenomenon is specifically termed a phase transition in graphs.

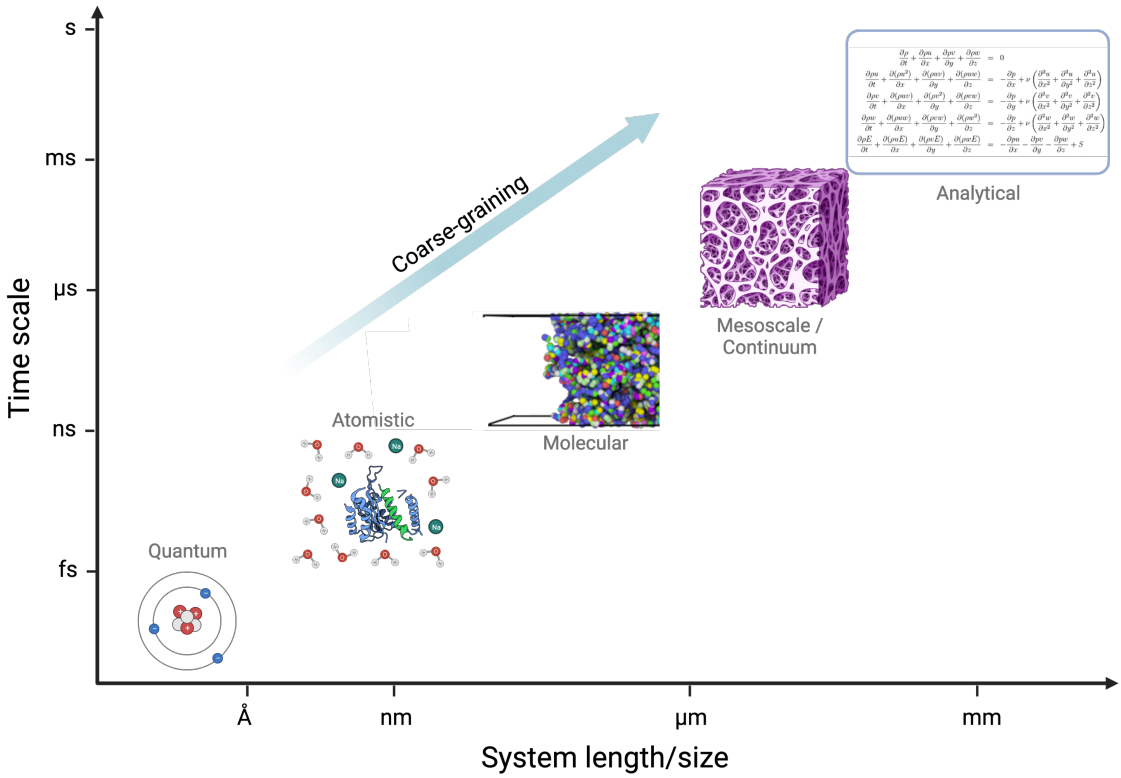


Fig. 2.5 **Time scale vs Length scale representation of various computational MD simulation resolutions usually applied in biology**.

criteria. Regardless of the technique used, the force field plays a critical role in controlling the total energy (in MC) and the forces (in MD) that ultimately determine the system's evolution.

While atomistic models hold great promise for achieving accuracy, they currently pose a significant challenge due to their high computational cost. However, the recent advancements in computational power and algorithms have enabled the application of atomistic molecular dynamics simulations to explore the properties of biomolecular condensates. Coarse-grained (CG) models may approximate whole amino acids, nucleic acids, proteins, DNA, RNA, or even groups of biomolecules by a single particle and generally use simplified physical potentials to represent the interactions between particles. Therefore, when compared to atomistic descriptions, CG models can enormously reduce the degrees of freedom used to describe a system, enabling the investigation of phase separation on biologically relevant spatiotemporal scales. However, because coarse-grained models sacrifice details for com-

putational efficiency, they require intensive care in their design and thorough experimental validation.

Across resolutions, a model's efficiency, transferability and accuracy are dictated by the approximations the model entails (e.g., degrees of freedom averaged out, solvent treatment, functional form of the energy functions) and by the quality of the parameters it uses. Molecular models can be designed based on fundamental physical and chemical knowledge of the constituent biomolecules ('mechanistic' or 'physics-based') and/or trained on experimental data sets ('data-driven' or 'machine-learned potentials').

The parameters can be fitted from the ' top–down', i.e. optimized to reproduce an experimental observable that reflects collective properties of molecules (e.g., partitioning coefficients, molecular sizes, diffusion coefficients etc.), or from the ' bottom–up', i.e. parameterized to reproduce quantities computed at higher resolution (e.g., all-atom simulations, quantum mechanical calculations).

Some CG models, such as the Martini model [119], have been parameterised from a combination of both 'top–down' and 'bottom–up' approaches. In all cases, care should be taken to use the model within the specific solution conditions it was developed and validated for (e.g., salt, pH, temperature, pressure). To ensure accurate results, it is crucial to use the models—from those including near-atomistic details to those grouping multiple biomolecules in a single particle— under the specific solution conditions, they were developed and validated for, such as salt, pH, temperature, and pressure.

CG models are proving to be helpful in answering questions related to biomolecular phase separation, as it occurs on multiple spatiotemporal scales. However, the type of questions that can be answered depends on the specific CG simulation approach used. The approach is closely linked to the design, resolution, approximations, and parameters of the model, as well as the sampling protocol used and whether or not enhanced sampling techniques are employed.

Chapter 3

General methods

In this chapter, we embark on a detailed exploration of the foundations of the computational methods employed throughout this thesis, with a particular emphasis on the realm of molecular dynamics simulations. As the bedrock of our work, molecular simulations provide a powerful lens through which we can probe and understand the intricate dance of molecules at the atomic and macroscopic levels. Moving beyond the basics of these simulations, we will delve into a specific kind of simulations, termed direct coexistence simulations, designed to calculate phase diagrams directly from molecular dynamics simulations. This technique, which is pivotal to our research, affords us a more nuanced understanding of phase transitions and the conditions under which they take place. Furthermore, we discuss further methods used throughout this thesis to conduct analysis from our simulation data.

Contents

3.1 Molecular Dynamics

All the simulations carried out in the following chapters are based on the Molecular Dynamics (MD) technique. MD is a computational method to solve numerically Newton's classical equations of motion. MD is useful for biomolecular systems because these are many-body systems and analytical solutions for the equations of motions of such systems cannot be obtained. Specifically, MD provides a way to estimate the coordinates of a system, r, of N atoms in 3 dimensions over time by numerically integrating Newton's equation of motion $F = m \cdot a$ of classical mechanics [60]. Let's consider the system is initialised with a momentum $p = m \cdot v$, where v is the velocity of the system with mass m. The system is then propagated over time t, following the equation below:

$$\vec{F} = m \cdot \frac{d^2\vec{r}}{dt^2} = -\nabla E(\vec{r}) \tag{3.1}$$

where $E(\vec{r}) = U(\vec{r}) + K(\vec{r})$, is the sum of the potential energy U and kinetic energy K:

$$E(\vec{r}) = \sum_{j>i}^{N} V(\vec{r_{ij}}) + \sum_{i=1}^{N} \frac{m\vec{r}^{\,2}}{2} \tag{3.2}$$

where $V(\vec{r_{ij}}$ describes the pairwise terms, also known as force field, and the pairwise $\vec{r_{ij}} \equiv \vec{r_j} - \vec{r_i}$.

Generally, the process of carrying out a simulation of a system can be broken down in the steps shown in Figure 3.1: (1) Generation of the topology of the system from initial coordinates, and the system's energy function. This topology usually requires the generation of a simulation box and the addition of solvent and ions - if the energy function describes them explicitly. (2) Short equilibration of the system, that first relaxes the system and brings it down to a low energy state,

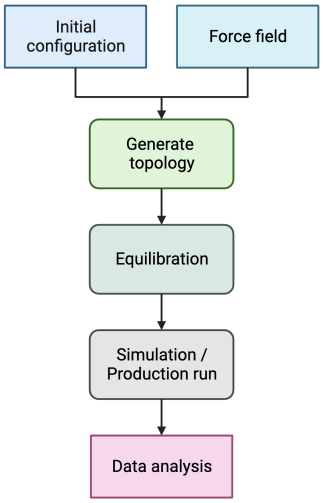


Fig. 3.1 **Flowchart of generic MD simulation process**.

and then another step performed to bring the system to a specific set of conditions (*i.e.* temperature, pressure, volume, *etc*.) that will be used in the production runs. (3) Production runs of the simulations and (4) analysis of simulation trajectories and computation of desired metrics/measurements.

It is important to note that a force field is a numerical description of the energy of the system, but limited to the system's predefined bonds, which means that bond breaking and forming events cannot be sampled through MD simulations.

3.1.1 Forces and integrators

Although Newton's equation is a second-order ordinary differential equation (ODE), it is impossible to solve analytically for systems of more than just a few atoms. Therefore, a numerical way to integrate these equations is required.

The forces atoms/particles of the system exercise on each other can be calculated from the potential energy, but this step is the most computationally expensive and time-consuming of the simulation. For instance, to compute the x-component of the force, we follow:

$$\begin{aligned} f_x &= -\frac{\partial U(\vec{r})}{x} \\ &= \left(\frac{x}{r}\right)\left(\frac{\partial U(r)}{\partial r}\right) \end{aligned} \tag{3.3}$$

The most straightforward way to integrate equation 3.1 is using a timestep integrator, which advances the trajectory of the system on each timestep or frame Δt. We, therefore, need to use finite differences:

$$\frac{\partial f(x)}{\partial x} = \frac{f(x+h) - f(x)}{h} \tag{3.4}$$

The mechanism of many integration algorithms can be understood by expanding $\vec{v}_{n+1} = \vec{v}(t_n + \Delta t)$ and $\vec{r}_{n+1} = \vec{r}(t_n + \Delta t)$ in a Taylor series. The Velocity Verlet algorithm [5] is one of the mostly used integrators. Starting from a system at t_0 with velocity $\vec{v_0}$, the next frame is computed using a Taylor's expansion, as below:

$$\vec{r}(t_0+\Delta t) = \vec{r}(t_0) + \vec{v}(t_0)\Delta t + \frac{1}{2}\left(\frac{\vec{f}(t_0)}{m}\right)\left(\Delta t^2\right), \tag{3.5}$$

then, after calculating $\vec{f}(t_0+\Delta_t)$:

$$\vec{v}(t_0+\Delta t) = \vec{v_0} + \frac{1}{2}\left[\frac{\vec{f}(t_0)}{m} + \frac{\vec{f}(t_0+\Delta_t)}{m}\right] \tag{3.6}$$

This process is then iterated for L timesteps or frames, so that $t = L\Delta t$, till the end of the MD simulation. One of the main benefits of using the Velocity Verlet algorithm versus other basic algorithms is that it is a sympletic integrator, meaning that the time evolution of the system's Hamiltonian conserves the form $dp \wedge dr$. With p as the momentum and r as the coordinates of the system, a sympletic intergrator would fulfill the requirement:

$$r = \frac{\partial H}{\partial p} \quad p = \frac{\partial H}{\partial r} \tag{3.7}$$

Moreover, this algorithm ensures time-reversibility, which is a critical feature to conserve energy, and it is highly accurate at reasonable time-steps.

Another important aspect to consider when using MD simulations is the ability to describe the system using a statistical ensemble, which consists of multiple copies of the system that represent possible states. This ensemble, also known as a thermodynamic ensemble, is in statistical equilibrium and allows for the derivation of properties of real thermodynamic systems from classical mechanics laws. Although there is a plethora of possible thermodynamical ensembles, we will consider the following three:

- Microcanonical ensemble (NVE): in this kind of simulations, a system of N particles is isolated from the surroundings and the volume does not change. Therefore, there is no energy transfer with the surroundings and the energy is the same throughout the simulation.

- Canonical ensemble (NVT): the system is kept at fixed volume but there is energy transfer with the surroundings, which allows for the definition of a temperature. The

temperature is kept fixed at a value T using a thermostat. A simple representation of this would be a system held in a thermal bath so that it transfer energy with one another. The energy is no longer constant.

- Isothermal–isobaric ensemble (NPT): the system is allowed to dilate and shrink while it tranfers energy with its encapsulating thermal bath. This ensemble undoubtedly plays a crucial role in chemistry, as the majority of significant chemical reactions occur under constant pressure and temperature conditions.

The timestep problem, however, is still a challenge in the computational fields. It's important to keep in mind that the selected timestep can have a significant impact on the results of a simulation. Therefore, it's crucial to choose a timestep that will provide accurate and reliable results. At higher timestep or Δt values, the energy might drift too much from its

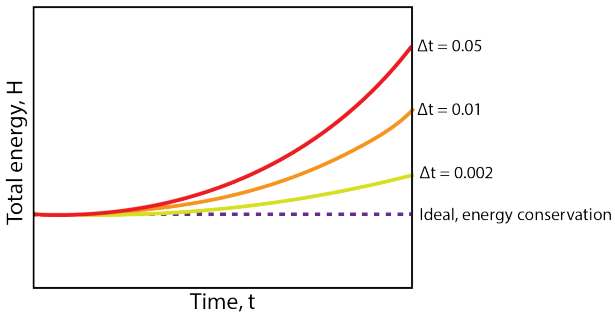


Fig. 3.2 **Energy drift of MD simulations at different timestep Δt values**. Example of the microcanonical ensemble (NVE).

'real' value, but at low values of Δt, the simulations become too inefficient as only a very short timeframe can be simulated (see Figure 3.2). With careful consideration and analysis, the right timestep can be confidently chosen for optimal simulation results.

3.1.2 Force fields

The potential energy of the system, $U(\vec{r})$, is comprised of bonded, short-range and long-range pairwise interactions that altogether describe the geometry and the properties of the system.

This is also commonly referred to as a force field. In biomolecular systems, most force fields can be divided into the following terms:

- Bonded interactions: referring to 2, 3 and 4-body interactions. Regarding bond stretching motion or covalent bonds, the simplest can be described by a harmonic potential:

$$V_{\text{bonded}}(\vec{r_{ij}}) = \frac{1}{2}k_{ij}\left(\vec{r_{ij}} - b_{ij}\right), \tag{3.8}$$

  where k_{ij} is the spring constant, and b_{ij} is the reference bond length. The 3-body bonded term is referred to as bond angle motion. This kind of vibration between 3 consecutive atoms can also be described through a harmonic expression;

$$V_{\text{angle}}(\theta_{ijk}) = \frac{1}{2}k^{\theta}_{ijk}(\theta_{ijk} - \theta^{0}_{ijk}) \tag{3.9}$$

  The energy required to twist a bond due to bond order (*i.e.*, double bonds) and neighbouring bonds by the 4-body bonded interaction term, and a bond can have multiple such terms. The total torsional energy is expressed as a Fourier series, as below:

$$V_{\text{torsion}}(\omega_i, \gamma_i) = \sum_i \sum_n \frac{1}{2}V_i^n\left[1 - \cos(n\omega_i - \gamma_i)\right] \tag{3.10}$$

- Non-bonded short-ranged pairwise interactions: such as van der Waals interactions, can be expressed with as simple as a Lennard-Jones potential:

$$V_{\text{vdW}}(r_{ij}) = \sum_{j=1}^{N-1}\sum_{i=j+1}^{N} \varepsilon\left[\left(\frac{r^0_{ij}}{r_{ij}}\right)^{12} - 2\left(\frac{r^0_{ij}}{r_{ij}}\right)^{6}\right] \tag{3.11}$$

- Long-range electrostatic interactions: interactions between charged atoms are usually described via Coulomb screening expressions:

$$V_{\text{elec}}(r_{ij}) = \sum_{j=1}^{N-1}\sum_{i=j+1}^{N} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}, \tag{3.12}$$

where ε_0 is the permitivity of vacuum ($\varepsilon_0 \approx 8.8542 \cdot 10^{-12} \cdot F \cdot m^{-1}$), and q refers to the point charge of each atom/particle, in e. The electrostatic force between two charges is repulsive if they have the same sign and attractive if they have different signs, and it acts along the straight line joining them.

In summary, a force field is comprised of the expressions that describe the interactions in the system that contribute to the total potential energy and their corresponding sets of parameters for the different types of particles or atoms.

3.2 Direct Coexistence Simulations

Computing phase diagrams is a crucial part of studying phase separation phenomena, both through experiments or computational simulations. One way of testing and validating our model is through Direct Co-existence (DC) simulations, which allow us to obtain the phase diagrams of mixtures of proteins in the temperature-concentration space. The Direct Coexistence simulation method consists of simulating the diluted and condensed phases both in the same simulation box separated by an interface.

The temperature-density (T–ρ) phase diagram of a phase-separating biomolecule can be systematically studied by performing simulations close to the phase boundary that defines the coexistence region for different simulation box sizes and number of molecules, followed by a finite size scaling extrapolation to the larger dimensions.

For a long time, computational researchers used droplet analysis in cubic simulation boxes to simulate the dilute and dense phases and the interface of a phase-separating system. However, this approach is very prone to suffering finite-size effects and the uncerteinties in the calculations are high, as proved by Nilson et al [134]. An alternative approach, which we employ in simulations of LLPS throughout the works presented in this manuscript, is described next.

3.2.1 The ‘slab’ method

In this type of simulations, the initial configuration is a slab of N molecules of the protein being simulated, prepared by compressing the mixture in XYZ directions at constant pressure.

The number of proteins placed in the simulation box and the box dimensions ought to be carefully chosen to avoid finite-size effects. Simulations with too few proteins might not represent the accurate behaviour of the system, and it might be difficult to simulate all the distinct phases. Although simulations with a higher number of proteins minimise the introduction of artifacts into the simulation and allow us to draw more meaningful information, an overly big simulation is computationally too costly and inefficient. In order to achieve a reasonable number of proteins to be simulated and the proper dimensions of the simulation box, it is necessary to perform finite-size analysis. The slab is placed in the

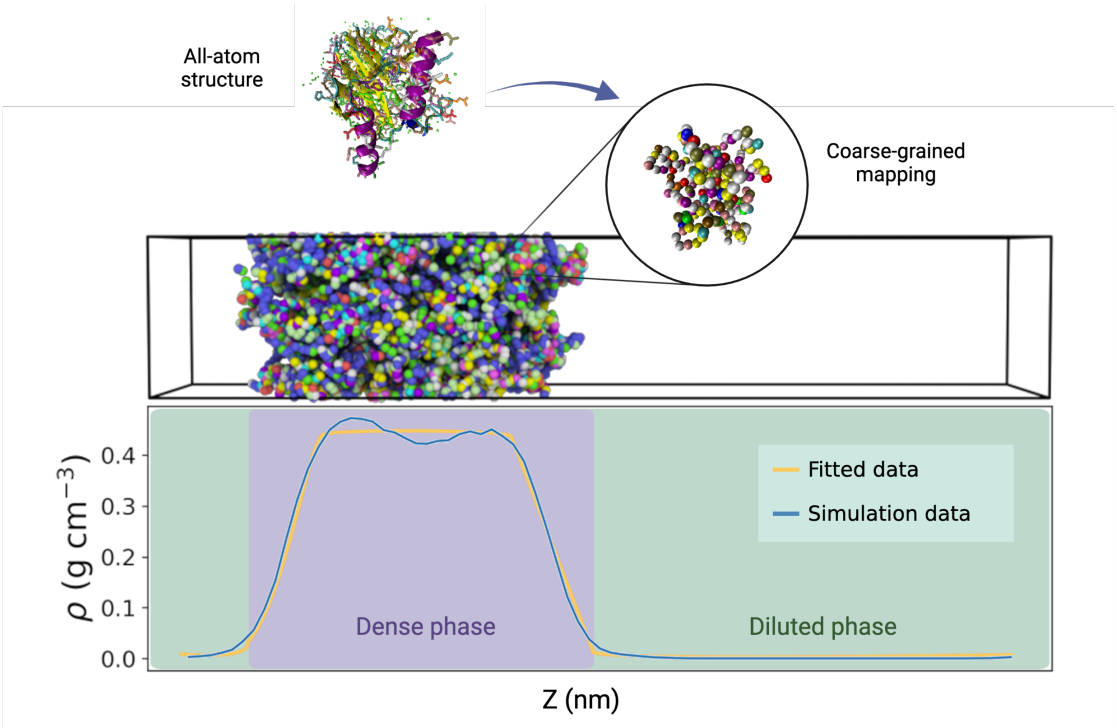


Fig. 3.3 **Summary of slab method to simulate liquid-like droplets using a coarse-grained model**.

middle of the simulation box, and the length of X and Y is the same as the length of the

slab, while the box is elongated in Z (see Figure 3.3), to about 4+ times the length of the cross-section. Additionally, as a rule of thumb, the length of the cross-section is set to at least about 2-3 times the radius of gyration of a single protein molecule.

From this starting configuration, the system is maintained at constant temperature and volume, and it will eventually converge to a stable state. In all of our simulations, the temperature is controlled by a Langevin thermostat [99], and so the pressure is by a Berendsen barostat [20]. All LLPS simulations in this work have been carried out using the LAMMPS software suite [165].

Through this, we can verify whether a protein will phase separate at a certain temperature T. In a temperature-driven phase diagram, there is a temperature above which no phase separation is observed in the system. This temperature is referred to as critical temperature or T_c. By running Canonical ensemble or NVT simulations of a system at several temperatures, we can systematically determine the critical temperature and critical density.

After sufficient equilibration of the simulation, the density of both phases, ρ_{high} and ρ_{low}, can be calculated from the density profile, which determines the phase boundary at the simulation temperature, as in Figure 3.3.

The critical temperature is hence obtained by fitting the densities of the dense and dilute phases at a range of temperatures by using the law of coexistence densities [151]:

$$(\rho_{high} - \rho_{low})^{3.06} = d(1 - \frac{T}{T_c}), \tag{3.13}$$

where ρ_{high} and ρ_{low} are the densities of the high and low-density phases in the mixture, and d is a fitting parameter. The critical temperature is that at which the densities of both phases are equal (see Figure 2.3). The density at the critical temperature, which is also referred to as the critical density ρ_c can be obtained from the law of rectilinear diameter [151]:

$$\rho_{high}(T) + \rho_{low}(T) = 2\rho_c + 2A(T - T_c), \tag{3.14}$$

with A being a fitting parameter. In summary, the slab method is highly efficient to compute binodals when care is taken to minimize finite size artifacts.

3.3 Contact Analysis

For each protein system we investigate, we can estimate the relative contribution of different amino acid pair contacts by computing a matrix of frequency of coarse-grained amino acid pair contacts. The contact matrix was computed at a temperature around ten per cent below the T_c. The cutoff to define a pairwise distance as a contact was dynamically calculated for each pair, as 1.2 times the average Van der Waals radius of both pair residues. Given residues i and j belong to different molecules, the contact maps were computed as shown below:

$$c_{i,j} = \sum_{i,j} \chi_{ij} \tag{3.15}$$

$$\chi_{ij} = \begin{cases} 1 & \text{if } \left|r_i - r_j\right| \leq r_{ij}^{\text{cutoff}} \\ 0 & \text{otherwise} \end{cases} \tag{3.16}$$

To get a more accurate picture of the main interactions driving phase separation for each variant, we normalised the contact maps to account for the occurrence of each residue in the sequence. This allowed us to account for the disparities in residue distributions in the sequences and extract the relevant residue-residue interactions. For each residue-residue pair, the normalisation was performed as follows:

$$C = \begin{vmatrix} \alpha_{M,M} & \alpha_{M,G} & \ldots & \alpha_{M,I} \\ \alpha_{G,M} & \alpha_{G,G} & \ldots & \alpha_{G,I} \\ \ldots & \ldots & \ldots & \ldots \\ \alpha_{I,M} & \alpha_{I,G} & \ldots & \alpha_{I,I} \end{vmatrix} \tag{3.17}$$

$$\alpha_{i,j} = \frac{c_{i,j}}{n_i \cdot n_j} \tag{3.18}$$

where $c_{i,j}$ is the number of contacts between residues types i and j (spanning the 20 naturally occurring amino acids- *i.e.* Met, Gly, *etc.*) and n_i is the number of i residues in the IDP sequence, making $\alpha_{i,j}$ the sequence-dependant ratio of contacts between residues i and

j . The resulting contact matrix is min-max normalised:

$$\beta_{i,j} = \frac{\alpha_{i,j} - \min C}{\max C - \min C} \tag{3.19}$$

$$C_{norm} = \begin{vmatrix} \beta_{M,M} & \beta_{M,G} & \dots & \beta_{M,I} \\ \beta_{G,M} & \beta_{G,G} & \dots & \beta_{G,I} \\ \dots & \dots & \dots & \dots \\ \beta_{I,M} & \beta_{I,G} & \dots & \beta_{I,I} \end{vmatrix} \tag{3.20}$$

where $\beta_{i,j}$ is the min-max normalised sequence-dependant ratio of contacts between residues i and j. The comparative contact maps, on the other hand, are computed as the difference in contacts of a protein variant respective to its wild type, as:

$$C_{comparative} = C_{variant} - C_{WT} \tag{3.21}$$

which is finally mean-normalised.

$$\gamma_{i,j} = \frac{\alpha_{i,j} - \overline{C_{comparative}}}{\max C_{comparative} - \min C_{comparative}} \tag{3.22}$$

$$C_{comparative,normed} = \begin{vmatrix} \gamma_{M,M} & \gamma_{M,G} & \cdots & \gamma_{M,I} \\ \gamma_{G,M} & \gamma_{G,G} & \cdots & \gamma_{G,I} \\ \dots & \dots & \dots & \dots \\ \gamma_{I,M} & \gamma_{I,G} & \cdots & \gamma_{I,I} \end{vmatrix} \tag{3.23}$$

Chapter 4

Everything you wanted to know about Potential of Mean Force calculations but were afraid to ask

Relative interaction strengths between amino acids are key guidelines on the parameterisation of the coarse–grained models presented in this thesis. Therefore, calculating free energies of binding poses a challenging task throughout our work and one of the techniques to do so is by computing potentials of mean force between two molecules. Calculating these profiles, however, is not a straightforward process and might be quite troublesome for those new to running umbrella sampling simulations. In this chapter, we review the statistical mechanics and theory foundations of this technique, as well as tips on how to carry out these simulations efficiently and minimise the errors, accompanied by a step-by-step tutorial on how to calculate the PMF curve of an interacting pair of amino acids.

Contents

4.1 Introduction

The measurement of free energy of binding between two molecules is crucial in understanding chemical and biological processes, with potential applications in medicine, drug development and materials science [179]. The change in free energy provides information on the direction and magnitude of binding and unbinding kinetics, non-bonded interactions, *etc*. As a result, it is a significant task for computational chemists to accurately calculate the free energy, particularly its profile, during a chemical or biological process.

In recent decades, several methods have been created for determining free energy through molecular dynamics (MD) simulations. These techniques include perturbation theory, thermodynamic integration (TI), umbrella sampling (US), the linear interaction energy approach, partition function based on the density of states, and MM/PBSA [67, 161].

The concept of potential of mean force (PMF) was developed by Kirkwood [90] and is a key concept in the statistical mechanics of fluids that corresponds to the ensemble average of the population distribution as a function of a collective variable (CV). A PMF can provide information about the conformational properties, transition rates, ligand-binding affinities of a system and many more, making it a fundamental piece in computational simulations [71].

Umbrella sampling (US) is an advanced sampling technique for obtaining the free energy profiles and thermodynamic data of a system by increasing the conformational sampling along a reaction coordinate (ξ) [166]. In other words, a simulation that, with standard techniques, would take extremely long timescales to sample configurations can achieve the same sampling by the introduction of a bias. Umbrella sampling is widely used in the field of computational chemistry for its quick convergence and ability to divide the MD simulations into several windows that can be run separately at the same time.

In this chapter, I introduce the framework for calculating potentials of mean force through all-atom MD simulations, providing an overview of the theory behind the method, a practical application with step-by-step instructions and general guidelines to troubleshoot the simulations and analysis. PMF calculations have played a key role in the development of the coarse-grained models presented in this thesis.

4.2 Method

4.2.1 Umbrella Sampling

In order to calculate the free energy of binding of a system, we can consider two states A and B, as the unbound and bound states, respectively.

$$\mathrm{A} \underset{k_{unbind}}{\overset{k_{bind}}{\rightleftharpoons}} \mathrm{B} \tag{4.1}$$

The free energy difference of a system, ΔG, which can exist in states A and B, at constant pressure, can be described as:

$$\Delta G = -\mathrm{RT}\ \ln K_D \tag{4.2}$$

$$\Delta G = -\mathrm{RT}\ \ln \left\langle e^{\left(\frac{-\mathrm{H_{AB}}}{\mathrm{RT}}\right)} \right\rangle_A, \tag{4.3}$$

, where R is the gas constant, T is the temperature, and H_{AB} is the Hamiltonian difference between states A and B, K_D is the dissociation rate between states A $\rightleftharpoons$ B , and $\langle\rangle$ rep-

resents an ensemble average [86, 184]. This Hamiltonian, however, cannot be calculated computationally without relying on statistical mechanics.

One can define a collective variable ξ, a reaction coordinate function of the atomic coordinates, r, that describes the transition between states A and B. The most common choice of $\xi\ (r)$ is based on geometry, like the center of mass (COM) distance. The probability distribution of the system as a function of $\xi\ (r)$ can be written as:

$$p(\xi) = \int \exp\left(-\beta\ E(r)\right) \delta(\xi - \xi(r))\ d^N r, \tag{4.4}$$

where β is $1/k_B T$, k_B being the Boltzmann constant, and N is the number of atomic coordinates.

Umbrella sampling was developed by Torrie and Valeau [166]. A biasing potential is introduced in the system to allow sufficient sampling of the space across the reaction coordinate ξ. This bias allows us to connect the energies of the different windows, each at different points along ξ, hence allowing us to overcome free energy barriers when sampling [86]. On each window i, the biasing potential affects the total energy of the system, $E_i^b(r)$ in the following fashion:

$$E_i^b\ (r) = E_i^u\ (r) - \omega_i(\xi) \tag{4.5}$$

were ω_i is the biasing potential on window i, and E_i^b and E_i^u are the biased and unbiased energies, respectively. The unbiased distribution $p_i^u\ (\xi)$ in an individual window is hence defined as:

$$p_i^u\ (\xi) = \frac{\int \exp\left(-\beta\ E(r)\right) \delta(\xi^b - \xi(r))\ d^N r}{\int \exp\left(-\beta\ E(r)\right)\ d^N r}, \tag{4.6}$$

which, when accounting for the biasing potential and assuming the system is ergodic - the entire phase space of the system is accessible within the simulation time-, the biased distribution $p^b(\xi)$ can be written as below:

$$p_i^b\ (\xi) = \frac{\int \exp\left(-\beta\ E(r) + \omega_i(\xi^b(r))\right) \delta(\xi^b - \xi(r))\ d^N r}{\int \exp\left(-\beta\ E(r) + \omega_i(\xi^b(r))\right)\ d^N r} \tag{4.7}$$

The population $p_i^b(\xi)$ can be further simplified since the bias depends on ξ only, and the

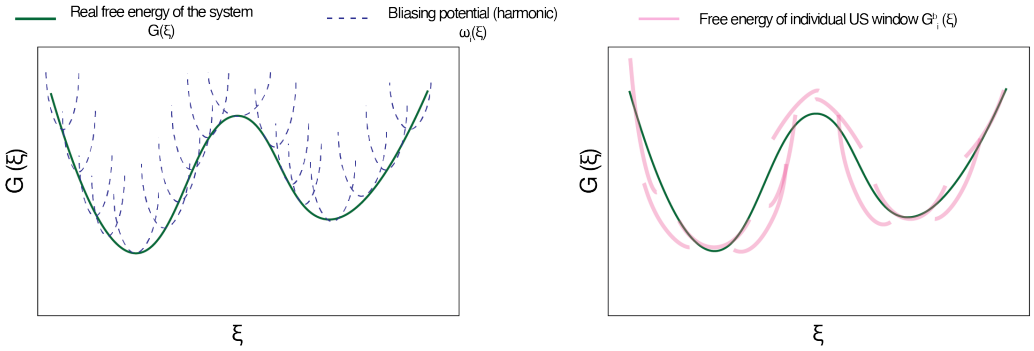


Fig. 4.1 **Calculation of free energy profile of a system through umbrella sampling**. Multiple biasing potentials (dashed line) are placed across the collective variable (ξ). The real free energy of the system is the dark green curve, which is unknown. Simulations are run for each window. The unbiased free energies G_i for each window are depicted by the faint pink curves, and are each offset by a different C_i each. These are used by WHAM to recover the free energy profile G (ξ).

integration is done over all N degrees of freedom of r.

$$p_i^b\ (\xi) = \exp\left(-\beta\,\omega_i(\xi)\right)\ \frac{\int \exp\left(\beta\ E(r)\right)\,\delta(\xi^b - \xi(r))\ d^N r}{\int \exp\left(\beta\ E(r) + \omega_i(\xi^b(r))\right)\ d^N r}, \tag{4.8}$$

thus,

$$\begin{aligned} p_i^u\ (\xi) &= p_i^b(\xi)\ \exp\left(\beta\,\omega_i(\xi)\right)\ \frac{\int \exp\left(\beta\ E(r) + \omega_i(\xi(r))\right)\ d^N r}{\int \exp\left(\beta\ E(r)\right)\ d^N r} \\ &= p_i^b(\xi)\ \exp\left(\beta\,\omega_i(\xi)\right) \left\langle \exp\left(\beta\,\omega_i(\xi)\right)\right\rangle \end{aligned} \tag{4.9}$$

MD simulations along N different windows allow us to calculate p_i^b for each window, and the biasing potential $\omega_i(\xi)$ is a factor not only we set up but also can obtain analytically. Generally, and in this case, as well, the biasing potential used is a simple harmonic spring potential with a spring constant k:

$$\omega_i(\xi) = \frac{1}{2}\,k\left(\xi - \xi_{i,ref}\right)^2 \tag{4.10}$$

Only if the sampling of each window is sufficient; the real, unbiased free energy G can be calculated through the following equation:

$$\begin{aligned} G_i(\xi) &= -(1/\beta)\ln\ p_i^b(\xi) - \omega_i(\xi) - (1/\beta)\ln\langle\exp\left(-\beta\,\omega_i(\xi)\right)\rangle \\ &= -(1/\beta)\ln\ p_i^b(\xi) - \omega_i(\xi) + C_i \end{aligned} \tag{4.11}$$

Obtaining the value of the constant C_i is not trivial since its value is unique for each window. To obtain the global value of $G(\xi)$ across the whole range of the reaction coordinate, one needs to combine the distributions of all windows. Several techniques have been developed to calculate the value of C_i and the global constant C from all the independent biased distributions. One of these techniques, termed the Weighted Histogram Analysis Method (WHAM), is commonly used along with umbrella sampling [96], and is discussed in the following section.

4.2.2 Weighted Histogram Analysis Method

The weighted histogram analysis method (WHAM) was developed by Kumar *et al*. [96] to remove the bias of the biased umbrella sampling simulation and quantify the constant C_i to eventually measure the potential of mean force of interest sampled with the smallest uncertainty. The WHAM first recovers the global unbiased probability distribution with the equation as follows:

$$P^u(\xi) = \sum_{i=1}^{N} \lambda_i p_i^u(\xi) \tag{4.12}$$

where λ_i is the weight of an individual window distribution. The value of λ_i is calculated so that the statistical error, σ^2 of the global probability distribution is minimised, hence $\frac{\partial\sigma^2(P^u(\xi))}{\partial\lambda_i} = 0$. The global free energy curve $G(\xi)$ is obtained by combining the windows and associating them through the bias potential employed. This is done by calculating the

ensemble average of the bias potential throughout the simulation:

$$\begin{aligned} \exp(-\beta C_i) &= \langle \exp(-\beta \omega_i(\xi)) \rangle \\ &= \int P^u(\xi) \exp(-\beta \omega_i(\xi)\, d\xi \end{aligned} \tag{4.13}$$

then,

$$= \int \lambda_i p_i^u(\xi) \exp\left(-\beta \omega_i(\xi)\right)\, d\xi$$

This process of calculating the optimal values of λ_i is iterated till convergence and minimisation of the statistical error in $P^u(\xi)$. Since WHAM employs normal distributions, the resulting PMF curves are rather smooth [96].

4.3 Test case: PMFs of sidechain-sidechain protein interactions using GROMACS

Although there are several umbrella sampling tutorials on the internet, often provided by organisations developing simulation software, very few go over the steps while mentioning their importance and implications in the final free energy curve. Here, we provide an overview of the steps to follow to compute the free energy profile of a sidechain–sidechain interaction between two amino acids using the GROMACS 2019.3 simulation software package [20] using the force field Amber *ff03ws* [22]. The calculation is carried out in the NPT ensemble at physiological conditions, such as 298.15 K, 1.0 bar pressure and 150 mM NaCl in water. The choice of ensemble is made carefully to reduce unwanted noise in the PMF curves and to mimic physiological conditions. In the NPT ensemble, the volume of the system fluctuates, which is crucial for simulating binding events. Maintaining a constant volume can result in inaccurate free energy calculations due to artificial pressures or constraints. In simulations involving binding, the presence of solvent molecules is critical. The binding process can alter the density and arrangement of solvent molecules. The NPT ensemble is preferred over the NVT ensemble as it allows for more natural adjustment of solvent density around the binding site, avoiding any undesired constraints.

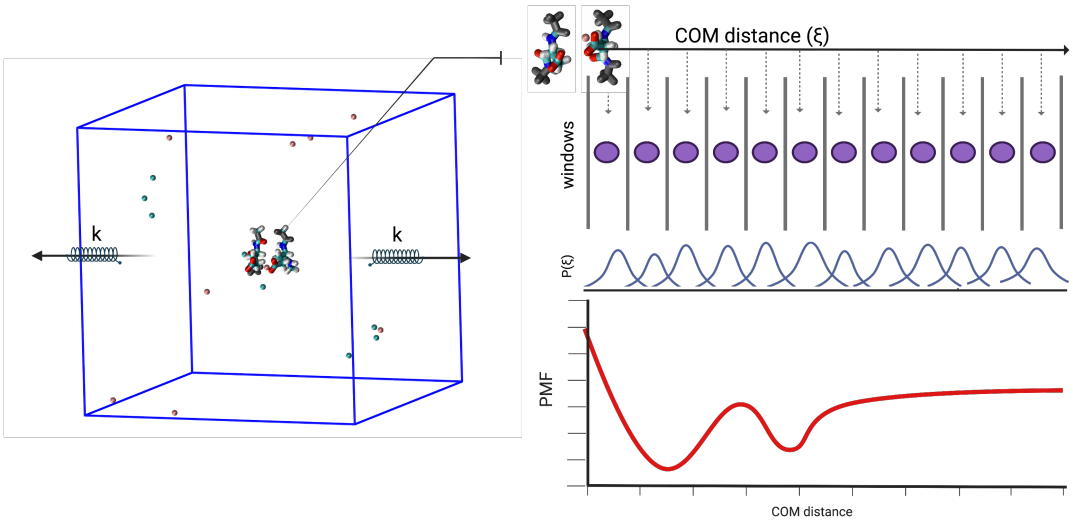


Fig. 4.2 **Diagram of Umbrella Sampling to PMF procedure**. The potential of mean force (PMF) is a free energy landscape that describes the thermodynamics of a system, in this case the interaction between two capped amino acids. Umbrella sampling is a computational technique used to calculate the PMF by sampling the potential energy of the system as a function of a reaction coordinate, a variable that describes the progress of the interaction. We used the distance between the centers of mass of the two amino acids. Define a set of n=34-40 windows along the reaction coordinate, divided into small intervals. In each window, apply a harmonic biasing potential k to keep the system close to a desired value of the reaction coordinate. The strength of the biasing potential should be chosen carefully to ensure that the system explores the entire range of the reaction coordinate. Run independent simulations in each window: For each window, run an independent simulation with the biasing potential applied. In each simulation, the system will explore the potential energy landscape in the vicinity of the chosen reaction coordinate value. Combine the results of the simulations: Collect data from each simulation to estimate the probability distribution of the system as a function of the reaction coordinate. Use the weighted histogram analysis method (WHAM) to combine the data from all windows and obtain the PMF.

Bash shell scripts are available to facilitate data preparation and file processing. These scripts can be executed on a regular UNIX system like Ubuntu or MacOS, or on a virtual Bash shell on Windows.

In order to calculate the PMF curve of two binding amino acids, we will follow these steps, which we will be discussing next:

1. Choose and set up the initial structure.

2. Prepare the Umbrella Sampling windows.

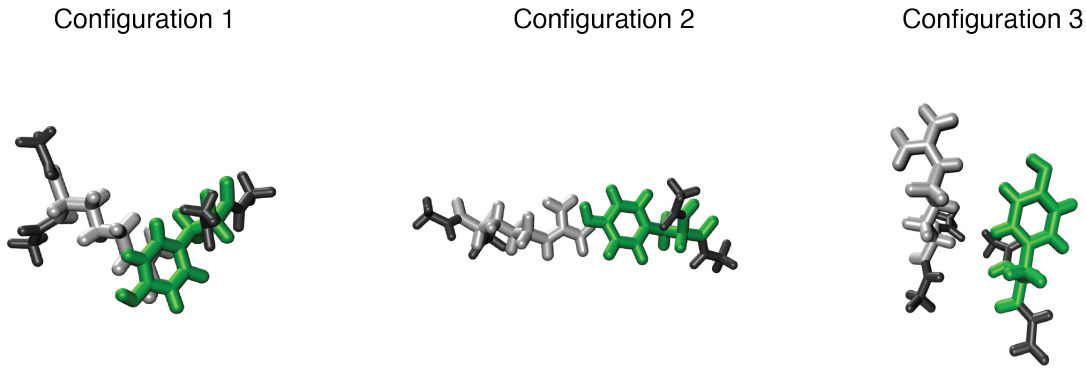


Fig. 4.3 **Initial structure of Arg–Tyr for Umbrella Sampling simulations in different sidechain–sidechain orientations.** While configurations 1 and 2 are potentially the most adequate for PMF calculations, configuration 3 is likely to introduce noise in the PMF curves due to cross-interactions between the capping groups.

3. Run Umbrella Sampling simulations.

4. Combine simulations and calculate PMF curve with WHAM.

4.3.1 Choosing and preparing the initial configuration

Before choosing collective variable (ξ) and generating windows for umbrella sampling, an initial key step is to choose the initial structure. This structure is the starting point and reference of the simulation and is critical for obtaining PMF curves that thoroughly sample the process of interest. As an example, here we focus on estimating the PMF between two amino acids. The election of the initial configuration for the interacting amino acids, in this case, is comprised of two chains, each containing a single amino acid capped by an acetyl group (-CO-CH_3) and an N-methyl (-N-CH_3) group, at the N- and C-terminals respectively, and oriented so that the side chain of each residue is facing one another. The function of the caps is to restrain the amino acids' amino and carboxyl groups from participating in the interaction and introducing undesirable noise. In order to determine the interaction energy of interaction between the two amino acids as if they were part of an inter-molecular interaction, we position the alpha carbon (C_α) of the amino acids strictly not adjacent to one another, except for configuration 3, where we purposefully orientate the pair with the cap-C_α-cap

axis parallel to one another, to visualise how the repulsive interactions between the capping groups distort the PMF curve.

In this example model, we use the centre-of-mass (COM) distance between the two whole chains as a collective variable, and the amino acids are Arg and Tyr. The Arg–Tyr interactions are known to be energetically one of the most favourable among proteins [63]. The most straightforward way to set up the windows with increasing inter-chain r_{COM} is to define one of the coordinate axes for the chain separation and the harmonic potential biasing the simulations. For this set, we use the Y axes as the biasing axes; therefore, the chains need to be oriented in perpendicular to the pulling axis.

4.3.2 Preparing the windows for umbrella sampling

4.3.2.1 Window generation, solvation and minimisation

A non-trivial requirement to keep in mind is that for umbrella sampling to be able to determine the free energy of the process sampled is for the resulting histograms to be overlapping with one another, so the range of distances and steps need to be chosen carefully. Although minimal increments in r_{COM} lead to overlapping histograms, this is not an efficient approach and can be computationally prohibitive. Usually, the minimum and most reasonable amount of windows and their consecutive separation are achieved through trial and error.

In these particular simulations, we set the windows at increasing distances between the amino acids along the Y-axis. In this specific case, we prepared 35 windows with distances ranging from -0.1 nm to 1.65 nm, specified in file `windows.txt`:

```
### windows.txt
### Inter-chain distances (nm)
    -0.10
    -0.05
    0.0
    ...
    1.60
```

```
    1.65
```

Once the distances are chosen, the initial structure, which has the smallest separation, is used as a reference to prepare the initial configurations of the restrained systems at the umbrella windows. We will translate one of the amino acids, Tyr, its corresponding distance in Y for each window, then set the spacing between each chain and the edge of the simulation box by 1 nm. Note that this distance is not arbitrary, as systems with longer chains would need a bigger separation with the edges of the simulation box in order to prevent periodic images of the system to interact with one another. To be able to select and translate an individual amino acid, it is necessary to create an index `.ndx` file using the `gmx make_ndx` function. The resulting system is then enclosed in a cubic box. In GROMACS, these are carried out with functions `gmx editconf` and `gmx genconf`.

```
aa1="arg"
aa2="tyr"
pair="${aa1}_${aa2}"
dir="<your-working-directory>"
ndx="${dir}/${pair}.ndx"
for i in $(cat windows.txt);
do
        fname="${dir}/${pair}.gro"
        oname="${dir}/newconf_${i}.gro"
        gmx editconf -f $fname -o $oname -translate 0  $i 0
             -n $ndx<<EOF
11 # Select index the AA to translate as in the .ndx file
1 # Return the whole system
EOF
        sleep 1
        fname=$oname
        oname="${dir}/renumber.gro"
```

```
        gmx genconf -f $fname -o $oname -renumber
        sleep 1
        fname=$oname
        oname="${dir}/newbox_${i}.gro"
        gmx editconf -f $fname -o $oname -bt cubic -c -d 1
        sleep 1

done
```

The starting windows are now set and ready to be solvated. With each windows having an increasing separation between chains, their corresponding simulation boxes also increase in size and, therefore, also in the number of solvent and ion particles. We opt for the TIP4P/2005 water model [169]. As we aim to sample the free energy of the Arg–Tyr interaction in physiological concentrations of NaCl, some solvent molecules will be replaced by Na^+ and Cl^- ions to reach a concentration of 150 mM. The ion model of choice is JC-SPC/E [19]. All the necessary parameters for the simulation are written and can be customised in the topology `.top` files. Note that since every window has a different number of solvent and ions, each window must have its corresponding topology file. The steps of solvation and ion addition are performed with `gmx solvate` and `gmx genion`, like in the code sample below:

```
aa1="arg"
aa2="tyr"
dir=calculations/${aa1}_${aa2}
for index in $(cat windows.txt);do
        fname=$dir/newbox_${index}.gro
        old_top=$dir/topol_${aa1}_${aa2}.top
        top=$dir/topol_${index}.top
        cp $old_top $top <-- Copy the base topology file
           for each window
```

```
        out=$dir/solv_${index}.gro
        watergro=<directory-with-ffs>/amber03ws.ff/
           tip4p2005.gro
        gmx solvate -cp $fname -cs $watergro -o $out -p
           $top
        sleep 2
        out=$dir/solv_ions_${index}.gro
        fname=$dir/solv_${index}.gro
        mdp=setup/ions.mdp
        tpr=$dir/ions.tpr
        gmx grompp -f $mdp -c $fname -p $top -o $tpr -
           maxwarn 10

        gmx genion -s $tpr -o $out -p $top -pname Na -nname
            Cl -conc 0.15 -neutral<<EOF
13 <-- Enter index of solvent group (SOL)
EOF
        old_top=$top
        top=$dir/win_${index}/topol_${index}_sub.top
        mdp=setup/em.mdp
        mkdir $dir/win_${index}/
        fname=$out
        cp $old_top $top
done
```

The next step before the umbrella sampling runs is crucial and consists of minimising the energy of the system in each window, since a non-equilibrated system is highly unstable and likely to blow up the simulation. We carry out the minimisation using the steepest descent [1] algorithm with a force tolerance of 500 J mol^{-1} pm^{-1}, and we restrain the heavy atoms of

the amino acids in all directions, by setting a restraining force of 200 J mol^{-1} pm^{-2} in the `posre.itp` file. These simulations are run in the NPT ensemble.

```
aa1="arg"
aa2="tyr"
for index in $( cat windows.txt  ) ; do
        dir=calculations/${aa1}_${aa2}
        fname=$dir/solv_ions_${index}.gro
        tpr=$dir/em_${index}.tpr
        top=$dir/topol_${index}.top
        ndx=$dir/solv_ions_${index}.ndx
        mdp=setup/em.mdp
        gmx grompp -f $mdp -p $top  -r $fname -maxwarn 1000
            -o $tpr -c $fname
        gmx mdrun -s $tpr -deffnm $dir/em_${index} -c $dir/
           em_${index}.gro -v

done
```

When estimating the force used in the positional restraints, significant care must be taken. If the restraining force is too low, the conformation after energy minimisation might differ notably respect to the initial configuration, and the ultimate free energy curve will not correspond to the desired configuration. However, if the restraints are too high, the likelihood of simulation failure increases and/or the post-minimisation conformation might turn unphysical. One needs to take a closer look at the configurations generated as well as the energy throughout the trajectory. An example of a well-minimised system can be found in Figure 4.4.

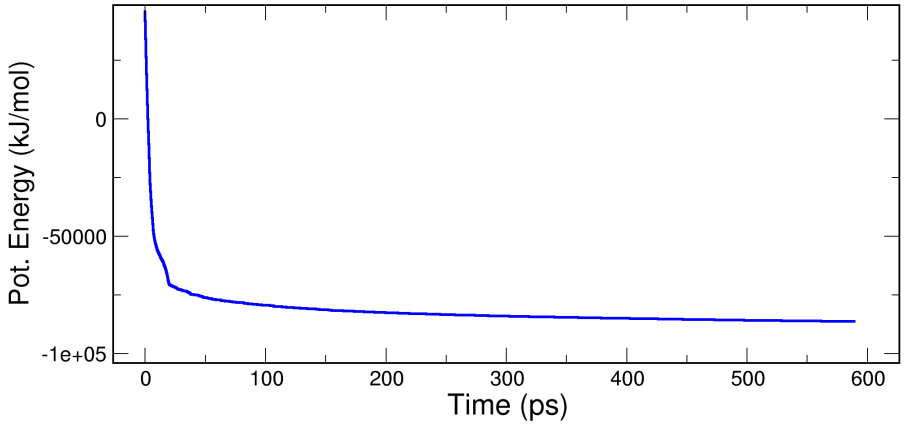


Fig. 4.4 **Potential energy throughout the energy minimisation step.** Exponential decay and plateau of the curve depict the system has rearranged and reached an energy minimum.

4.3.2.2 Umbrella sampling simulations and WHAM

Production runs of each window are ready to be carried out once all windows reach a stable conformation. For increased sampling, we run three production runs for each window, each using a different random seed. The first requirement is to create an index file, again, using `gmx make_ndx`, for each window defining each amino acid chain in separate groups, renaming each as chain A and B, respectively. A new index file should be created for each window, as the number of molecules varies following the size of the box.

In production runs, the force to restrain the heavy atoms of 1 J mol^{-1} pm^{-2}, is applied in the X and Z directions, while the pulling force is applied in Y. Therefore, the restraints file needs to be edited beforehand. The restraining force in Y needs to be set to zero.

We use a pulling force of 6000 J mol^{-1} pm^{-2} between chains A and B and set the distance as the geometric collective variable. The production runs output files measuring the

pulling force and inter-chain distances across the 30 nanoseconds–long simulations for each window.

From the resulting simulations, WHAM will require the topology files (`*.tpr`) and any of either the force files (`pullf*.xvg`) or distance files (`pullx*.xvg`) to calculate the histograms and calculate the free energy profile from the biased simulations. WHAM is implemented as a GROMACS function and it can be used as below:

```
gmx wham -it tpr-files.dat -if pullf-files.dat -o -hist
    -b 1000 -unit kCal -temp 298.15 -tol 1e-5 -bs-
   method b-hist -nBootstrap 100
```

To increase the sampling, we carry out the analysis using Bayesian bootstrapping [51]. This technique is based on resampling the observation n times, 100 in the code above, therefore calculating n independent free energy curves, hence computing the uncertainty of the PMF. The tolerance for bootstrapping iteration was set to 10^{-5} kcal mol^{-1}. The first one ns of simulation is discarded for each window in order to sample uniquely from converged data points. For larger systems, however, longer time might be deemed as 'equilibration time'.

4.4 Analysing and understanding your PMFs

Determining the necessary amount of sampling to achieve a converged PMF curve is, *a priori*, a challenging task. The convergence of a profile can be assessed in many ways. The most straightforward way is generally to split the trajectory in several chunks and calculate the individual free energy profiles of each chunk, then comparing the variance among all of them. Bootstrapping techniques, many of which have been introduced in simulation software suites, make this task much simpler and more efficient.

Although non-trivial, the interpretation of PMF curves can be summarised as measuring the ΔG of the sampled path by taking the maximum and minimum peaks in the curve, and associating the extrema to different configurations across the phase space.

In the interaction sampled in this case, the minima in the free energy profile represent the configurations where the Arg–Tyr interaction is the most attractive, while the flat plateau

represents the ones where the residues are not interacting with one another (see Figures 4.5 and 4.6).

4.4.1 Dependence of initial structure on PMF curve

The effect of the initial starting configuration on the final PMF curve must be carefully considered. As shown in Figure 4.6, the computed free energy profile can be significantly different depending on the choice of initial configuration. In this case, we observe the capping atoms add unwanted repulsive forces to the interaction, hence weakening it and producing a free energy profile with a very shallow well when umbrella sampling is initialised from configuration 3, where the amino acids are placed side to side.

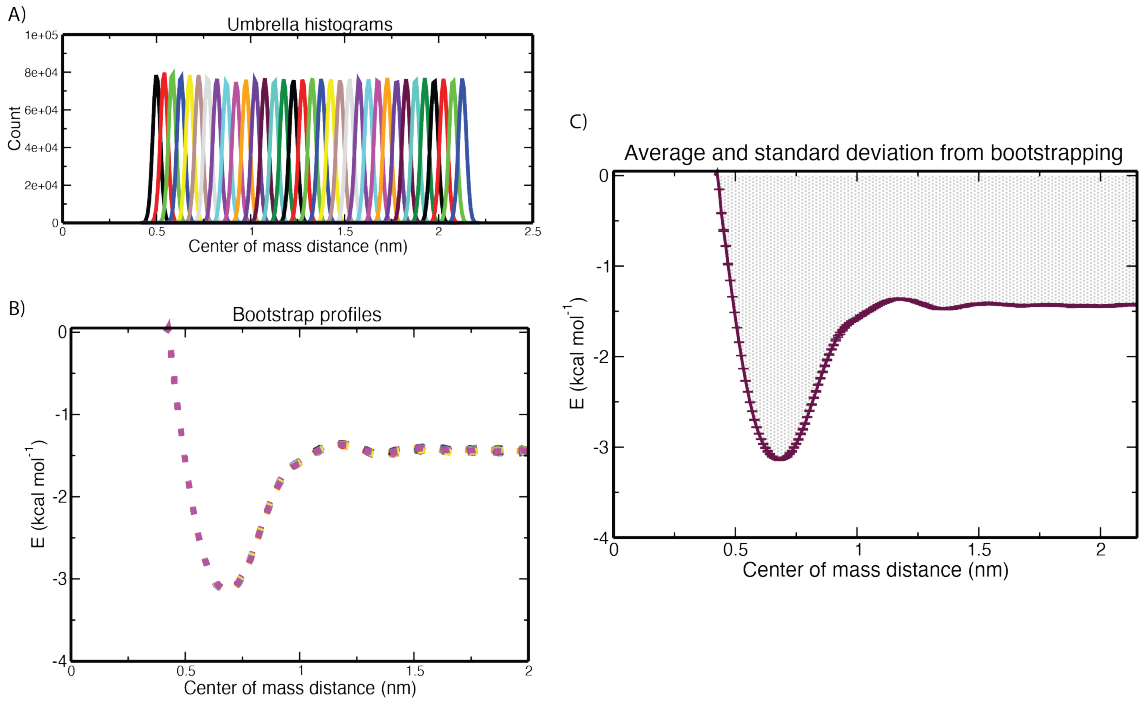


Fig. 4.5 **US–WHAM scheme to compute free energy of sidechain–sidechain interaction of RY.** A) Distribution for the distance between the centers of mass of Arg and Tyr capped residues. Each colored histogram represents an individual umbrella sampling window. B) Free energy along the reaction coordinate. Different colored lines represent individual interactions initiated by Bayesian bootstrapping. C) Raw average free energy profile along inter-chain Arg-Tyr center-of-mass separation. Error bars are associated to the standard deviation of the free energy, from bootstrapping.

Duarte and colleagues [95] elucidated the nature of the cation–π interaction between Arg and Tyr, which is further strengthened by the sp^2 hybridised non-aromatic orbital in the Arg sidechain. This can also be observed in the PMF curve of configuration 2 vs configuration 1 in Figure 4.6, where the initial structure of the former implies reduced interaction between the aromatic π orbitals from tyrosine and sp^2 hybridised non-aromatic orbital of the arginine residue side chain.

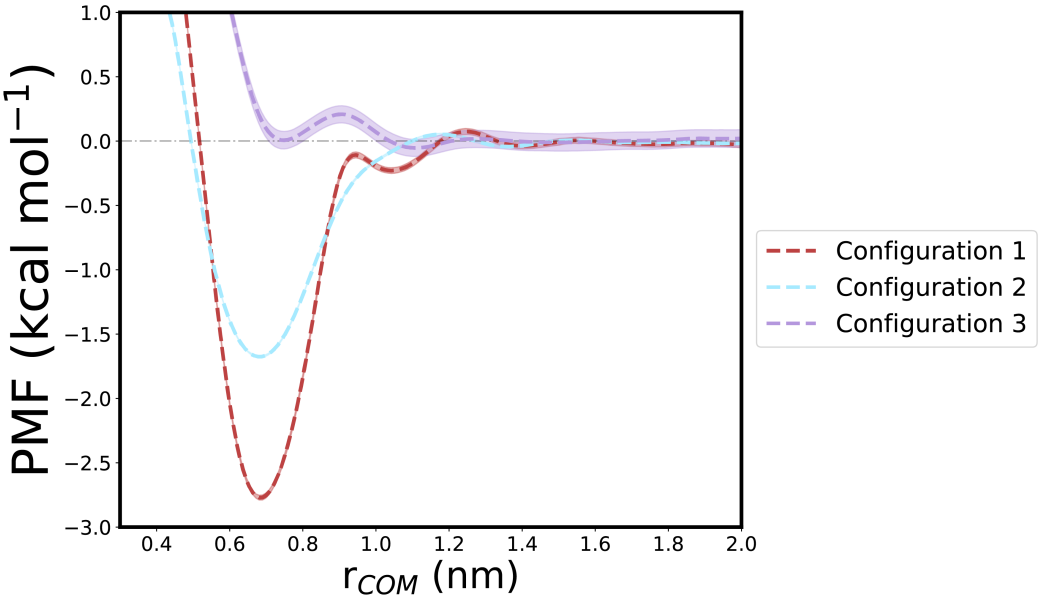


Fig. 4.6 **Free energy profiles of Arg–Tyr interaction computed from different starting configurations**. Dashed lines represent the calculated potential of mean force curves, while highlighted surfaces correspond to the associated error of the curves from Bayesian bootstrapping.

4.5 Troubleshooting

Despite the fact that the US-WHAM method is widely used in computational chemistry due to its simplicity, there are several obstacles and bottleneck issues that can make constructing a free energy profile from umbrella simulations a complex task for someone attempting a PMF for the first time.

Henceforth, one should carefully consider the following when simulations do not perform as expected:

- Overlapping of umbrella histograms: ensuring the selected pathway for umbrella sampling for the windows is able to capture the whole phase space along the reaction coordinate without significant gaps between adjacent windows is essential for the conditions that allow us to compute the free energies accurately. Failing to do so will, in the best of cases, lead to free energy curves with excessive noise and large errors, therefore failing to provide any meaningful information.

  When histograms do not overlap or are sparsely populated, it can make it difficult to identify energy barriers or transition states in the system. It is also not possible to compute the individual values of C_i for window by 'merging' or 'stitching' configurations that do not overlap. This can lead to inaccurate estimations of the heights and positions of these barriers, which can bias the PMF curves and hinder our understanding of the underlying dynamics. It is important to ensure sufficient data and a clear separation of histograms to avoid these issues.

- A sufficient sampling of phase space along the reaction coordinate: Similarly to the condition above, windows with a reduced sampling of ξ will be notably penalised when making up the free energy profile from the unbiased simulations. In some cases, it can also make the WHAM algorithm fail, and the resulting PMF curve will show as a linearly increasing curve. In this case, the histograms should be carefully visualised to identify the window with insufficient sampling. To overcome this, one should check the MD simulations for the specific window has run properly or add additional windows at nearby ξ.

- Pre-production equilibration of umbrella window conformations: as mentioned in the previous section, prior to running umbrella MD simulations, the starting window configurations must be well equilibrated. Otherwise, the production runs might turn energetically unstable and crash, and lead to poor histograms. The easiest approach to ensure this is avoided is to check the potential energy of each window decreases exponentially and reaches a stable plateau during the NPT equilibration runs. Specific windows which might contain unrealistic conformations (such as overlapping

molecules) must be carefully left out. Adequate choice of sampling path along the reaction coordinate is essential for this.

- Adequate strength of positional restraints in production MD runs: positional restraints (commonly specified in the topology files) play a crucial role in ensuring the correct configurations are sampled. Insufficient restraints in the process of window preparation might lead to alteration of the conformational sampling with respect to the one planned for. Excessively high positional restraints, both in the NPT equilibration or the US MD simulations, might alter the binding dynamics. Consequently, the resulting PMF curves may not accurately reflect the true free energy landscape and may exhibit artifacts or inaccuracies. A clear case in which the positional restraints are not appropriately set up is when the forces applied are so high that the simulation box blows up.

All of the aforementioned elements in umbrella sampling simulations are widely interconnected, and avoiding poor statistics or biased energy landscapes may require iterative adjustments and close-up analysis throughout the steps of the US-WHAM execution.

4.6 Discussion: challenges and alternatives

Here, we demonstrate that joint umbrella sampling and WHAM could be reliable methods for computing PMFs if a reasonable trajectory is provided and the biased simulations used in each umbrella window are well-converged.

In molecular simulations, long auto-correlations are a common occurrence. As a result, histograms generated from short umbrella simulations may not accurately represent all areas of the phase space. However, umbrella simulations, which generally require high-performance computing clusters, are computationally too costly to run above the microsecond range. Consequently, standard umbrella sampling simulations may not effectively sample processes that occur over long timescales or rare events. It may be more appropriate to consider advanced enhanced sampling methods or alternative approaches that are specifically designed for rare event sampling.

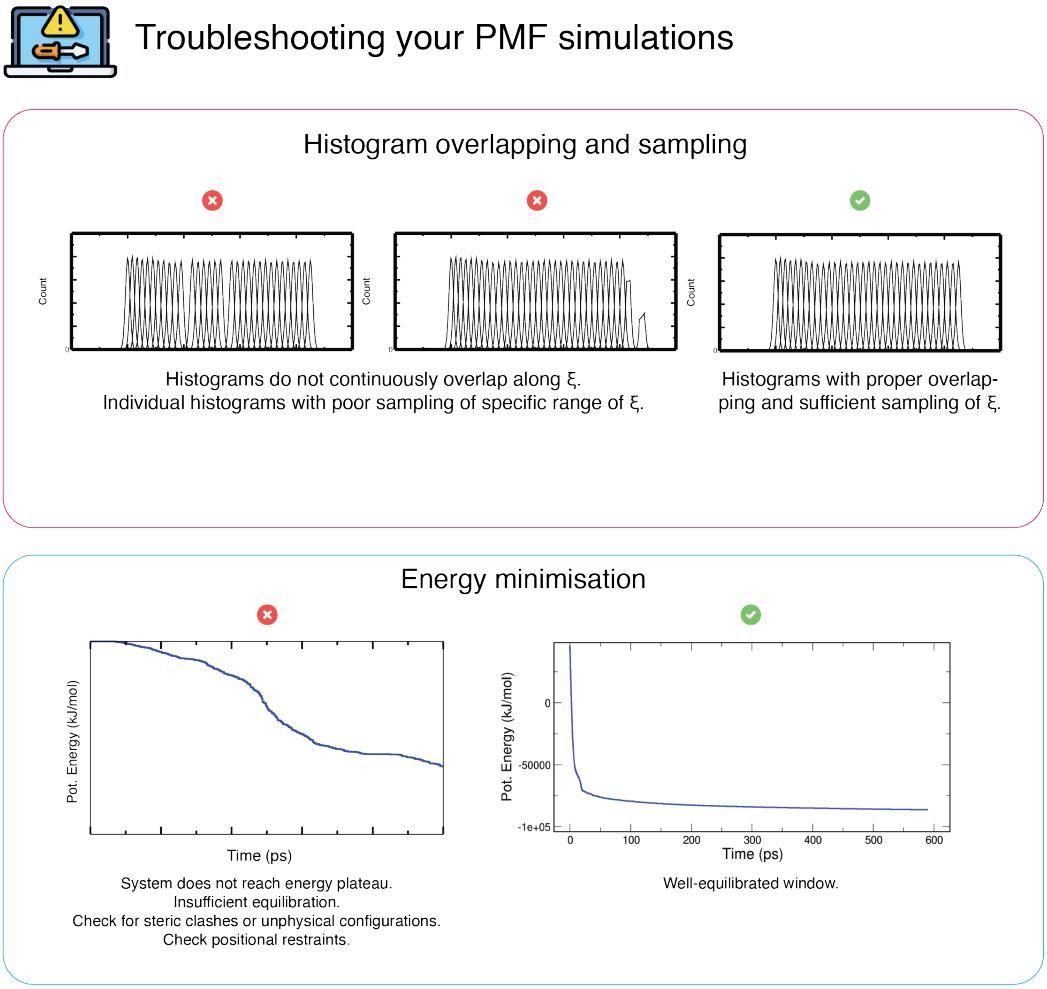


Fig. 4.7 **Brief guide to troubleshooting your US simulations.** Top) Examples of histograms that might result from umbrella simulations. Bottom) Examples of potential energy evolution throughout NPT equilibration of individual umbrella sampling windows. Both provide examples of correct and problematic/incorrect histograms and potential energies.

Additionally, umbrella sampling is ideal for sampling processes happening throughout a reduced number of collective variables but is not applicable to highly dynamic or high-dimensional systems since the number of collective variables needed to reach appropriate sampling would be computationally unfeasible.

Multiple alternative techniques exist to calculate free energies, some of which address the mentioned limitations of umbrella sampling.

For instance, metadynamics is an enhanced sampling method that introduces a history-dependent biasing Gaussian potential to biased regions, allowing the system to escape local minima. By analyzing the accumulated biasing potentials, free energy calculations can be performed. Metadynamics is especially useful for studying rare events and transitions [32].

Free Energy Perturbation (FEP) is another widely used method for calculating free energy differences. The process of transforming one system into another involves a series of intermediate states, while applying a coupling parameter. The free energy difference is computed by integrating the perturbation energy changes along the transformation path. While FEP requires running separate simulations for each intermediate state and can be computationally demanding, it is highly accurate [191].

Replica Exchange MD (REMD) is an enhanced sampling technique that runs multiple parallel simulations at different temperatures. Periodic exchanges of configurations between the replicas facilitate enhanced exploration of the conformational space. Free energy differences and thermodynamic properties can be estimated by analysing the temperature-dependent sampling. In REMD combined with umbrella sampling, commonly known as Hamiltonian REMD, the combination of temperature and Hamiltonian scaling promotes a wider range of conformational sampling [120]. This can be beneficial in capturing rare events, exploring multiple energy basins, or studying systems with flexible or disordered region. This methodology has been employed to study and develop coarse-grained DNA models [55].

Ultimately, the choice of methodology to calculate free energies comes down to the system's dimensionality, the timescale of the process we aim to study, computing resources, etc. In this chapter, however, we have presented an effective methodology to calculate residue–residue protein interactions in physiological conditions. This proxy is one of the building blocks of the modelling work carried out throughout my PhD and described in the upcoming chapters.

Chapter 5

Development of a physics-based coarse-grained model for LLPS

Sequence-dependent protein coarse-grained modelling has become the pillar of the computational side of the field, as a tool for the study of biomolecular phase separation and identification of its dominant physicochemical driving forces. In this chapter we present Mpipi, a multiscale coarse-grained model that describes with promising accuracy the change in protein critical temperatures as a function of amino acid sequence. The model is parameterized using both atomistic simulations and bioinformatics data and takes into account the dominant role of π–π and hybrid cation–π/π–π interactions, as well as the much stronger attractive contacts established by arginines compared to lysines. Our comprehensive set of benchmarks for Mpipi and seven other residue-level coarse-grained models against experimental radii of gyration and quantitative in vitro phase diagrams provide convincing evidence that Mpipi predictions are consistent with experimental data on both fronts. Furthermore, Mpipi accurately reproduces the experimental liquid–liquid phase separation trends for sequence mutations on FUS, DDX4 and Laf-1 proteins.

Contents

5.1 Preamble

This chapter introduces the initial phase of a simplified computational model for LLPS that takes into account the interplay between pairwise interactions among protein residues, intrinsically disordered sequences, and globular domains. It is important to note that I was also one of the authors of a larger scientific collaboration, which was initially published in a peer-reviewed journal article titled **"Physics-driven coarse-grained model for biomolecular phase separation with near-quantitative accuracy"** [83]. The aforementioned article, published in Nature Computational Science in 2021, serves as the primary reference for the research presented in this chapter.

As a second author, my contribution to this paper involved the initial parameterisations of the model using bioinformatics data and the validation of the subsequent parameterisation versions of the model through simulations of different proteins, namely FUS, LAF-1, G3BP1 and DDX4, and several mutated sequences, via direct co-existence simulations to obtain their corresponding phase diagrams.

Professor J. A. Joseph carried out the PMF calculations of a set of residue pairs to fine-tune the sequence-specificity of the coarse-grained model, and carried out comparative

analysis of the Mpipi model with other relevant models in the field at the time of publication. They also computed the single–molecule radii of gyration of IDPs. Dr. Reinhardt carried out the simulations of protein hnRNPA1 and a set of experimentally probed mutations. P.Y. Chew computed the phase diagrams of the multiphase protein–RNA systems.

Inside cells, the creation of membraneless organelles and compartments is made possible through a complex interplay of molecular interactions known as liquid-liquid phase separation (LLPS). This process is primarily driven by the unique properties of the molecules involved, especially proteins and nucleic acids like RNA. These molecules often have regions that lack a defined three-dimensional structure, allowing them to interact weakly through various means such as electrostatic interactions, hydrogen bonding, and hydrophobic interactions [15]. These interactions are reversible and are sensitive to factors like temperature, pH, and ionic strength, making LLPS a highly dynamic process [113].

Recently, our group observed a salt-mediated re-entrant phase transition for protein LLPS [94]. In that work, π–π interactions emerged as a major driver of LLPS at both low and high salt concentration. Specifically, our atomistic simulations revealed that, for proteins, the strongest pairwise interactions arise when the two amino acids in question both possess π electrons in their side chains, including both aromatic or non-aromatic residues with sp^2-hybridised groups. In fact, prior to that study, the influential stickers-and-spacers framework of Pappu and colleagues [74, 35] and the groundbreaking quantitative experimental phase diagrams of Mittag and colleagues together positioned aromatic residues as being chief drivers of biomolecular phase separation [119, 29]. The dominant role of π–π interactions in LLPS was also put forward by Vernon et al. [170], who identified an abundance of π–π contacts, involving not only aromatic but also non-aromatic residues, in protein structures via a comprehensive survey of the protein data bank. Moreover, it is evident that even within the subset of aromatic residues, tyrosine is a stronger contributor than phenylalanine to LLPS stability [174, 145, 29, 94].

Given the current computational techniques and needs for further exploration and representation of π-based interactions in LLPS, our main motivation for this work was to develop a coarse-grained model that takes into account the dominant role of $\pi - \pi$ and cation–π inter-

actions, and the higher relative interaction strength of Arginine than Lysine, which models known to that day had failed to capture. All in all, our goal was to develop a computational approach that provides a quantitative description of a biomolecular system by linking the amino-acid sequence to their macroscopic phase behaviour.

5.2 Methods

Atomistic PMF calculations

To determine the varying impacts of different interaction types at physiological salt levels, we conducted atomistic potential-of-mean-force (PMF) calculations for a select group of residue pairs. These pairs include aromatic–aromatic (WW, YY and FF), cation–aromatic (RY, RF, KY and KF), charged–charged (RE, RD, KE, and KD) and non-aromatic non-charged (AA, SS and PP) interaction pairs. All residue pairs from our previous work [94] were recomputed at a 150 mM concentration of NaCl, and we have sampled extra pairs.

Preparation of structures

The AMBER ff03ws force field [22] is used to carry out the calculations. The force field is ideal for investigating interactions between proteins. For modelling the solvent (water) and salt, we use the JC-SPC/E-ion/TIP4P/2005 models [19], as in our previous work [94]. Each amino acid's N- and C-terminal ends are capped with acetyl and N-methyl capping groups, respectively. Following the same proxy as in the test case in Chapter 4, the amino acids are positioned in such a way that their side chains face each other, following the common patterns found in protein structures. Multiple pairwise arrangements are tested to determine the strongest interaction mode in cases where the interaction preference is uncertain.

Each dimer is then introduced in a cubic box containing TIP4P/2005 water molecules with a minimum distance of 1 nm between the dimer and the edge of the box. Na^+ and Cl^- ions are added to achieve a salt concentration of $\sim$150 mM and to neutralise the charge of the

system. The resulting systems are then minimised (force tolerance = 500 $\mathrm{J\,mol^{-1}\,pm^{-1}}$), with positional restraints of 200 $\mathrm{J\,mol^{-1}\,pm^{-2}}$ applied to all the heavy atoms in each dimension.

Umbrella sampling

The interaction between each dimer is probed with umbrella sampling.For production runs, positional restraints of 1 $\mathrm{J\,mol^{-1}\,pm^{-2}}$ in directions perpendicular to the pulling direction are used to constrain heavy atoms. The centre-of-mass (COM) distance between interacting pairs is restrained with a harmonic umbrella potential (pulling spring constant k=6 $\mathrm{J\,mol^{-1}\,pm^{-2}}$). All bonds with hydrogens are constrained using the LINCS algorithm [76] and the integration time step was set to 2 fs. Periodic boundary conditions (PBC) were used during MD simulations. Electrostatics are computed using particle-mesh Ewald summations [54] with a Coulomb cutoff of 0.9 nm. All atomistic simulations and analyses are carried out using the simulation package.

For each umbrella sampling run, around 34 to 40 windows, spaced at 50 pm from 0 nm to 2 nm, are used for each dimer. Each window was simulated for 10 ns. Three independent simulations are conducted for each umbrella sampling window (i.e. an aggregate simulation time of 30 ns per window), using different random seeds. Umbrella sampling data is analysed using the weighted histogram analysis method (WHAM) [96]. The first 1 ns of simulations is used for equilibration and is discarded in the WHAM analysis. Error analysis is performed using the Bayesian bootstrap method.

In our calculations, we mainly focus on the distances between the centers of mass (COM) for fixed molecular orientations. However, we translate these distances to C_α–C_α distances in the coarse-grained potential. The choice of order parameter may affect the effective free energy, which is influenced by the Jacobian determinant of the transformation [160].

Nevertheless, the choice of order parameter does not affect the observable properties of the system. For a fixed molecular orientation, the two distances are linearly related, and using the PMFs consistently should result in the same ratios of interaction strengths, regardless of the order parameter selected.

Refitting of cation–π charges

Cation-π interactions involve significant polarization of the π electron clouds of aromatic side-chains in the vicinity of cationic side-chains, such as arginine and lysine, particularly under physiological salt conditions. Several attempts have been made to accurately capture cation-π interactions in atomistic force fields, utilizing both fixed-charge and polarisable force fields, as discussed by Liu and colleagues [111]. Recently, Paloni et al. demonstrated that the fixed-charge AMBER 99SB-disp force field correctly accounts for Arg/Lys–π interactions in the DDX4 NTD [140]. In another study, Liu and colleagues used quantum-mechanical calculations to reparameterize the Lennard-Jones parameters in the CHARMM36 force field to simulate cation-π pairs [111]. Their modified parameters improved the descriptions of the chosen folded proteins, resulting in a closer match to experimental crystal structures.

In order to model cation–π interactions at the all-atom level, we follow the approach by Krainer et al [94] and first refit the charges on tyrosine and phenylalanine side chains. Particularly, the pairs (Arg/Lys–Phe/Tyr) are first optimised using constrained geometry optimisations at MP2/6-31G(d) level of theory, where the backbone and capping group heavy atoms are frozen.

The electrostatic surface potential (ESP) is then computed for respective optimised pairs at HF/6-31G(d) level. These calculations are carried out using the Gaussian 09 code. Finally, the restrained electrostatic potential method in AMBER [18] was used to refit the side-chain charges of Tyr and Phe to the ESPs from the quantum-mechanical calculations; the charge symmetry of the rings is maintained during the refitting procedure. The refitted charges are then used when probing the pairwise interaction strengths via umbrella sampling, as described above.

5.2.1 Mpipi model

In the Mpipi model, each amino acid or nucleic acid is represented by a single bead with its corresponding mass, molecular diameter (σ), charge (q), and an energy scale reflecting the relative planar π–π contact frequency (ε), as shown in Table 5.1. We broadly follow the

approach of Dignon et al. [45] to compute the potential energy of a given protein or RNA molecule as

$$E_{\text{Mpipi}} = E_{\text{bond}} + E_{\text{elec}} + E_{\text{pair}}. \tag{5.1}$$

The bond energy is computed by using a harmonic bond potential,

$$E_{\text{bond}} = \sum_{\text{bonds } i} \frac{1}{2} k (r_i - r_{i,\text{ref}})^2, \tag{5.2}$$

where the spring constant k is set to $8.03\,\text{J}\,\text{mol}^{-1}\,\text{pm}^{-2}$ and r_i is the bond length: reference a bond length, $r_{i,\text{ref}}$, of 381 pm is used when bond i connects two protein beads. The electrostatic contribution to the potential energy is computed using a Coulomb term with Debye–Hückel electrostatic screening, [44],

$$E_{\text{elec}} = \sum_{i,j} \frac{q_i q_j}{4\pi\varepsilon_{\text{r}}\varepsilon_0 r_{ij}} \exp(-\kappa r_{ij}), \tag{5.3}$$

where $\varepsilon_{\text{r}} = 80$ is the relative dielectric constant of water, ε_0 is the electric constant and κ^{-1} =795 pm is the Debye screening length, corresponding to a concentration of sodium chloride of 0.15 M to be consistent with the PMF calculations. We use a Coulomb cutoff of 3.5 nm. The dielectric constant ε_0 and the Debye length κ control the range of ionic interactions and determine the relative importance of charges relative to all other interactions.

Finally, the nonbonded interactions between protein beads are modelled via the Wang–Frenkel (WF) potential [175]. The WF potential between two beads of types i and j a distance r apart is given by

$$\phi_{ij}(r) = \varepsilon_{ij}\alpha_{ij}\left[\left(\frac{\sigma_{ij}}{r}\right)^{2\mu_{ij}} - 1\right]\left[\left(\frac{R_{ij}}{r}\right)^{2\mu_{ij}} - 1\right]^{2\nu_{ij}}, \tag{5.4}$$

where

$$\alpha_{ij} = 2\nu_{ij}\left(\frac{R_{ij}}{\sigma_{ij}}\right)^{2\mu_{ij}}\left[\frac{2\nu_{ij}+1}{2\nu_{ij}\left(\left\{\frac{R_{ij}}{\sigma_{ij}}\right\}^{2\mu_{ij}} - 1\right)}\right]^{2\nu_{ij}+1}, \tag{5.5}$$

and σ_{ij}, ε_{ij} and μ_{ij} are parameters specified for each pair of interacting beads. We use $\nu_{ij} = 1$ and $R_{ij} = 3\sigma_{ij}$. The total pairwise energy E_{pair} is then taken as the sum over all pairs of beads evaluated within their respective interaction ranges (i.e. R_{ij}, at which ϕ_{ij} vanishes). Most

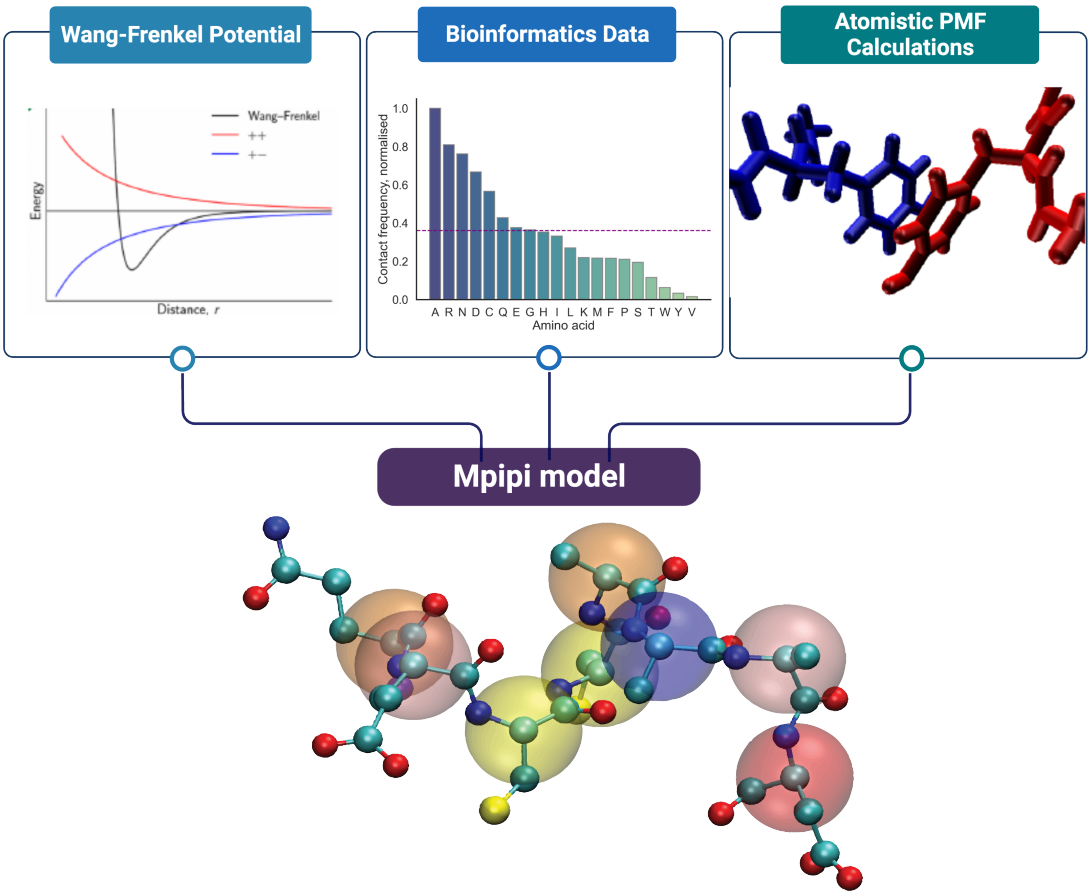


Fig. 5.1 **Anatomy of the Mpipi model**. The Mpipi is a coarse-grained model where each amino acid is represented by a single bead, and solvent effects are described implicitly. In a protein, these beads are bonded consecutively by a harmonic potential. Pairwise interactions are computed through the Wang-Frenkel potential [175], and charged interactions are computed with the Debye-Huckel Coulombic term. The parameterisation of the model is a result of bioinformatics data and all-atom PMF data.

importantly, the Wang–Frenkel potential is finite-ranged, vanishing quadratically to zero at the custom distance R_{ij}, and so obviates the need for truncating and shifting the potential. This key feature makes the Wang–Frenkel potential better suited for numerical calculations

and removes any ambiguities or inconsistencies that may arise from one implementation to the next.

For example, more typical Lennard-Jones-based potentials can exhibit significant undesirable finite-size effects as a function of the cutoff distance and subsequent tail corrections [77]. The computational performance of the Wang–Frenkel potential is comparable to the Lennard–Jones potential for the same cutoff. However, we have yet to do this here. If one wishes to simulate extensive systems, the Wang–Frenkel potential's more flexible functional form allows for optimising the distance at which the potential vanishes, which could enable a significant computational boost without degrading performance. Moreover, although from its scaling properties, the Lennard-Jones potential appears at first glance to account for London dispersion interactions, in reality this is not the case in solution, where the potential accounts for many interactions in a coarse-grained way; a further advantage of the Wang–Frenkel potential is that it removes this misleading appearance of physicality. To obtain the parameters that appear in the WF parameterisation, we first determine relative planar π–π contact frequencies of the amino acids from the work of Vernon et al. [170], determine Ashbaugh–Hatch-styleLennard-Jones interactions following Dignon et al. [45] obtain the initial WF parameters from these.

By increasing the μ parameter above unity, this framework can effectively adjust the steepness of the repulsive region and the width of the attractions in the potential (more about this exponent will be discussed in Chapter 6). To ensure accurate calculation of the overall interaction energy, we also modify the values of ε_{ij} through appropriate multiplication, such that the PMF curves of residue pairs i and j approximate their WF analogues, including any relevant screened charge-charge interaction. It is important to consider all factors to ensure accurate results. Although it has been suggested [102, 40] that simple arithmetic combination rules are often sufficient, unlike in previous models, the pairwise interactions for those residue pairs which dominate the phase behaviour are explicitly specified, giving the increased flexibility of Mpipi. It cannot be assumed beforehand that interactions between different species will be accurately described by averaging interactions within the same species. Specifically, we have observed that the interactions between arginine and lysine are

Full name	Code		Charge	π–π freq.
Alanine	Ala	A	0	0.091
Arginine	Arg	R	+	0.552
Asparagine	Asn	N	0	0.353
Aspartate	Asp	D	−	0.195
Cysteine	Cys	C	0	0.127
Glutamine	Gln	Q	0	0.365
Glutamate	Glu	E	−	0.211
Glycine	Gly	G	0	0.220
Histidine	His	H	+	0.668
Isoleucine	Ile	I	0	0.005
Leucine	Leu	L	0	0.021
Lysine	Lys	K	+	0.048
Methionine	Met	M	0	0.073
⋆ Phenylalanine	Phe	F	0	0.712
Proline	Pro	P	0	0.144
Serine	Ser	S	0	0.113
Threonine	Thr	T	0	0.057
⋆ Tryptophan	Trp	W	0	1.000
⋆ Tyrosine	Tyr	Y	0	0.762
Valine	Val	V	0	0.011

Table 5.1 **The 20 naturally occurring amino acids with their one- and three-letter codes, their charges and π–π contact frequencies**. In simulations with all models, the charge of His is set to $+0.375e$, whilst all other non-zero charges are set to $\pm 0.75e$, as appropriate. Amino acids marked with a '⋆' are aromatic. The last column represents the planar π–π contact interaction frequencies for each amino acid, extracted from Figure 1B of Ref. 170, and normalised to a range between 0 and 1.

noticeably distinct from what is predicted by mixing rules. We model disordered proteins and regions as completely flexible polymers, as shown in Figure 5.2.

Validation simulations use various previously reported models. Mostly, these are based on the functional form introduced in the work of Dignon et al. [45]. The functional form of both the bonded and electrostatic potentials is identical, with only slight variations in their constants. We provide a full parameter listing of Mpipi in Table 5.2.

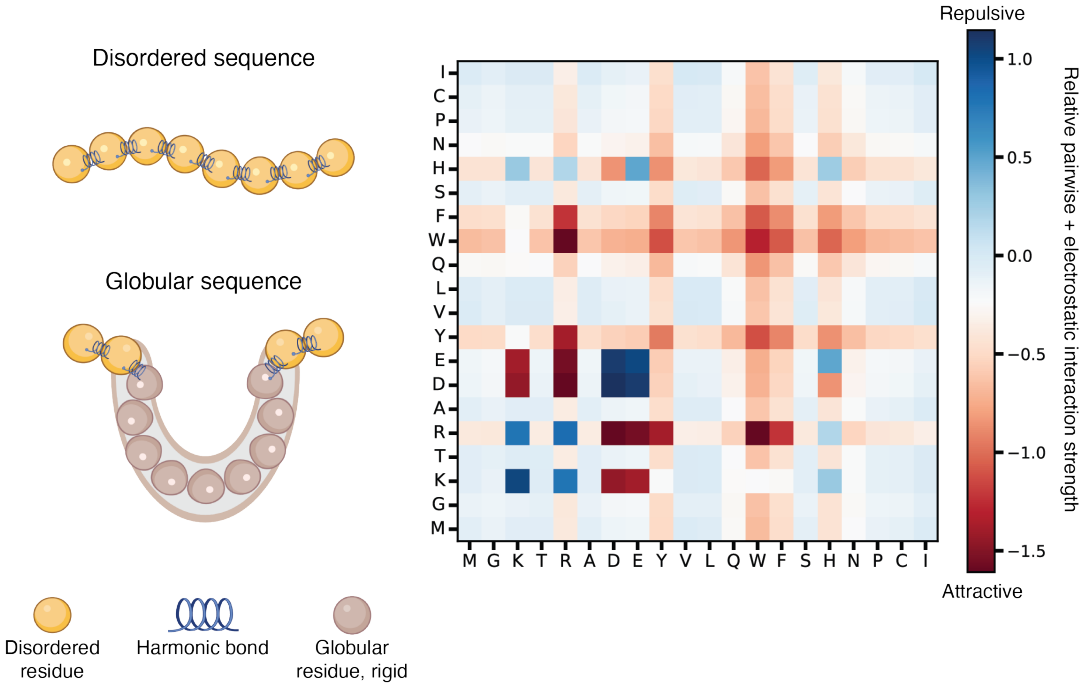


Fig. 5.2 **Description of bonds in Mpipi, and pairwise relative interaction energy**. On the left, disordered regions in proteins are modelled as flexible polymers, where each residue bead is linked consecutively by a harmonic potential. Globular sequences, on the other hand, are modelled as rigid bodies, and only the first and last globular residues are linked to their adjacent disorered residues through a bond. Additionally, all the interactions between IDR and globular residues have a WF ε parameter rescaled by $\sqrt{0.7}$ its disordered value, while globular-globular are rescaled by 0.7.

5.3 Results

5.3.1 Designing a multiscale coarse-grained model for probing biomolecular LLPS

We have designed a residue-level coarse-grained model for predicting biomolecular phase behavior (Fig. 5.1). In the Mpipi model, each amino acid (or nucleic acid) is mapped onto a unique bead based on simulation and experimental data. Following the work of Dignon et al. [45], the potential energy of molecules is computed as the sum of the bond energy,

Table 5.2 **Wang–Frenkel parameters for the Mpipi coarse-grained model**. For each residue, we provide two data sets. The row highlighted in red shows the value of $\varepsilon/\mathrm{kJ\,mol^{-1}}$ and the row highlighted in green is σ/nm. The value of μ is 2 for all entries except the ones highlighted in blue [$\mu(\mathrm{V–I}) = 4$, $\mu(\mathrm{I–I}) = 11$]. All charged amino acids have $q = \pm 0.75e$, as appropriate, except H, which has $q = 0.375e$, where e is the elementary charge.

	M	G	K	T	R	A	D	E	Y	V	L	Q	W	F	S	H	N	P	C	I
M	0.165 536	0.284 583	0.136 758	0.147 114	0.717 619	0.186 280	0.306 152	0.324 582	0.959 705	0.094 437	0.105 776	0.502 105	1.233 991	0.902 083	0.211 635	0.851 406	0.488 298	0.247 358	0.227 764	0.083 592
	0.646 795	0.557 618	0.656 778	0.617 823	0.664 396	0.586 850	0.614 146	0.631 818	0.659 004	0.632 057	0.648 344	0.636 519	0.675 573	0.653 821	0.593 894	0.639 251	0.618 802	0.613 446	0.609 429	0.649 639
G		0.403 630	0.255 806	0.266 161	0.836 666	0.305 327	0.425 199	0.443 630	1.078 752	0.213 484	0.224 823	0.621 152	1.353 039	1.021 130	0.330 682	0.970 454	0.607 345	0.366 405	0.346 812	0.202 639
		0.469 511	0.567 134	0.528 442	0.576 639	0.498 013	0.525 925	0.543 637	0.571 317	0.541 239	0.557 893	0.548 630	0.587 924	0.566 122	0.505 252	0.551 541	0.530 902	0.524 999	0.520 872	0.558 033
K			0.079 986	0.118 336	0.510 189	0.157 502	0.002 063	0.002 201	0.429 701	0.065 660	0.076 998	0.473 328	0.438 466	0.482 315	0.182 858	0.365 556	0.459 520	0.218 581	0.198 987	0.054 815
			0.667 134	0.627 940	0.673 819	0.596 669	0.623 699	0.641 358	0.668 408	0.643 354	0.659 256	0.645 969	0.684 969	0.663 228	0.603 618	0.648 660	0.628 255	0.623 106	0.619 115	0.661 800
T				0.128 691	0.699 197	0.167 858	0.287 729	0.306 160	0.941 283	0.076 015	0.087 354	0.483 683	1.215 565	0.883 661	0.193 213	0.832 984	0.469 876	0.228 936	0.209 342	0.065 170
				0.588 906	0.635 202	0.557 795	0.584 987	0.602 654	0.629 806	0.603 697	0.619 809	0.607 329	0.646 378	0.624 625	0.564 801	0.610 053	0.589 610	0.584 320	0.580 321	0.621 462
R					0.376 209	0.738 363	0.006 368	0.006 502	2.543 119	0.646 520	0.657 859	1.054 188	2.893 893	2.280 037	0.763 718	0.519 808	1.040 381	0.799 441	0.779 847	0.635 675
					0.683 905	0.604 909	0.632 986	0.650 709	0.678 624	0.647 780	0.664 461	0.655 828	0.695 255	0.673 424	0.612 191	0.658 837	0.638 103	0.631 985	0.627 828	0.664 491
A						0.207 024	0.326 896	0.345 326	0.980 449	0.115 181	0.126 520	0.522 849	1.254 736	0.922 827	0.232 379	0.872 151	0.509 046	0.268 102	0.248 509	0.104 336
						0.527 007	0.554 510	0.572 198	0.599 537	0.571 467	0.587 918	0.576 988	0.616 122	0.594 352	0.534 119	0.579 777	0.559 268	0.553 738	0.549 687	0.588 726
D							0.330 938	0.344 590	1.100 321	0.235 049	0.246 387	0.642 721	1.374 603	1.042 699	0.352 247	0.007 355	0.628 914	0.387 974	0.368 380	0.224 208
							0.582 352	0.600 058	0.627 647	0.597 874	0.614 509	0.605 004	0.644 244	0.622 456	0.561 730	0.607 876	0.587 280	0.581 455	0.577 340	0.614 699
E								0.358 242	1.118 751	0.253 479	0.264 818	0.661 151	1.393 034	1.061 129	0.370 677	0.007 494	0.647 344	0.406 404	0.386 811	0.242 639
								0.617 767	0.645 370	0.615 504	0.632 146	0.622 721	0.661 969	0.640 180	0.579 422	0.625 600	0.604 999	0.599 152	0.595 036	0.632 333
Y									1.753 874	0.888 606	0.899 945	1.296 274	2.028 156	1.696 252	1.005 804	1.645 576	1.282 467	1.041 527	1.021 934	0.877 761
									0.673 363	0.642 355	0.659 039	0.650 525	0.690 005	0.668 159	0.606 828	0.653 569	0.632 798	0.626 630	0.622 465	0.659 055
V										0.023 338	0.034 677	0.431 006	1.162 888	0.830 984	0.140 536	0.780 308	0.417 199	0.176 259	0.156 666	0.012 493
										0.626 600	0.639 093	0.619 973	0.658 912	0.637 178	0.578 166	0.622 610	0.602 261	0.597 433	0.593 567	0.671 801
L											0.046 016	0.442 345	1.174 227	0.842 323	0.151 875	0.791 646	0.428 538	0.187 598	0.168 004	0.023 832
											0.653 407	0.636 649	0.675 594	0.653 860	0.594 697	0.639 294	0.618 936	0.614 021	0.610 125	0.660 772
Q												0.838 674	1.570 556	1.238 652	0.548 204	1.187 976	0.824 867	0.583 927	0.564 334	0.420 161
												0.627 785	0.667 141	0.645 328	0.584 255	0.630 743	0.610 059	0.604 035	0.599 885	0.636 715
W													2.302 443	1.970 538	1.280 086	1.919 858	1.556 753	1.315 814	1.296 220	1.152 048
													0.706 655	0.684 798	0.623 414	0.670 209	0.649 416	0.643 217	0.639 053	0.675 603
F														1.638 630	0.948 182	1.587 954	1.224 845	0.983 905	0.964 312	0.820 139
														0.662 955	0.601 640	0.648 367	0.627 603	0.621 441	0.617 279	0.653 879
S															0.257 734	0.897 506	0.534 397	0.293 457	0.273 864	0.129 691
															0.541 267	0.587 063	0.566 533	0.560 932	0.556 848	0.595 253
H																0.113 872	1.174 168	0.933 229	0.913 635	0.769 463
																0.633 778	0.613 018	0.606 865	0.602 702	0.639 315
N																	0.811 064	0.570 124	0.550 531	0.406 358
																	0.592 335	0.586 308	0.582 164	0.618 999
P																		0.329 185	0.309 591	0.165 419
																		0.580 616	0.576 515	0.614 386
C																			0.289 997	0.145 825
																			0.572 436	0.610 587
I																				0.001 653
																				0.692 168

electrostatic and short-ranged energy terms which account for π–π, cation–π and other non-charged interactions. Bonds are modeled via harmonic springs and charge–charge interactions are approximated via a Coulomb term with Debye–Hückel screening. The main differences between the Mpipi model and other sequence-based coarse-grained models for LLPS are: (1) the functional form of short-ranged terms, (2) the parameterization of short-ranged interactions, and (3) the relative contribution of long-ranged electrostatics and short-ranged terms to the total energy.

For non-bonded short-range interactions, we use the recently developed Wang–Frenkel [175] pair potential (Fig. 5.1; see Methods). Like similar 'toy model' potentials, the Wang–Frenkel potential accounts for key physical interactions, namely a short-ranged excluded-volume repulsion and a longer-ranged attraction which gradually decays to zero.

However, the Wang–Frenkel potential has several advantages [175] over Lennard-Jones-like potentials that are commonly adopted in molecular simulations. Most importantly, the Wang–Frenkel potential is finite-ranged, vanishing quadratically to zero at the user-specified cutoff distance, and so obviates the need for truncating and shifting the potential. This key feature makes the Wang–Frenkel potential better suited for numerical calculations.

When determining the energy scale for short-ranged interactions, our goal is to find the right balance between π–π and non-π-based contacts. We achieve this by combining bioinformatics data and atomistic short-ranged free energy estimates. Initially, we use data from Vernon et al. [170] to determine the relative π–π contact frequencies for amino acids (see 5.1). The authors of Ref. 170 predicted planar π–π contact frequencies by surveying around 6000 high-resolution structures in the Protein Data Bank. We used these contact frequencies to set an initial energy scale for short-ranged interactions in our model.

Next, we refined this initial energy scale using atomistic PMF calculations. We focused on aromatic π–π (Fig. 5.3a), cation–π (Fig. 5.3b), and other non-π-based (Fig. 5.3c) interactions. Pappu and colleagues' seminal work highlights the importance of aromatic "stickers" as the main drivers of biomolecular LLPS [74, 35]. Our recent findings support the stickers-and-spacers model, indicating that even at extremely high salt concentrations, aromatic π–π interactions constitute dominant forces in LLPS [94].

In our research, we analyzed the PMF values between YY, FF, and WW (Fig. **??**a) at physiological salt concentration (see Methods). Our findings align with the bioinformatics data [170] and experiments [174, 145, 29], indicating that the aromatic π–π interactions are stronger in WW>YY>FF order (magenta bars in Fig. 5.3). Moreover, our analysis revealed that aromatic π–π interactions are at least twice as strong as non-π based interactions (dark yellow bars in Fig. 5.3d) that include non-polar, polar, and special residue interactions (e.g. Pro), commonly categorized as spacers by Pappu and others. Our model's spacer-type interactions have been adjusted by a factor between x % and y % to match the relative PMF well depths and breadths. Our focus was on the interplay between basic residues, particularly Arg and Lys, and aromatics, known as cation–π interactions. These interactions significantly contribute to the LLPS of biomolecules. According to research conducted by Wang et al. [174], LLPS systems stabilized by Arg–Tyr interactions have a saturation concentration about an order of magnitude lower than those predominantly stabilized by Tyr–Tyr contacts. Our PMF calculations also corroborate the stronger influence of Arg–Tyr interactions compared to Tyr–Tyr. It's crucial to find a balance between Arg–π and Lys–π interactions. Recently, we suggested that Arg–π interactions are best described as hybrid cation–π/π–π, while Lys–π contacts represent "purer" cation–π interactions [94]. The distinction is mainly due to the presence of π electrons in the Arg side-chain [50, 8, 170, 94, 58, 59]. These electrons enable Arg residues to interact much more strongly with π-binding partners than Lys can [49, 59, 58, 94].

Our model's cation–π interactions have been reparameterised to match the relative weights suggested by the atomistic simulations (Fig. 5.3). A summarized chart of the interaction energies between amino-acid pairs is provided in Fig. 5.2, right, including electrostatic and Wang–Frenkel contributions.

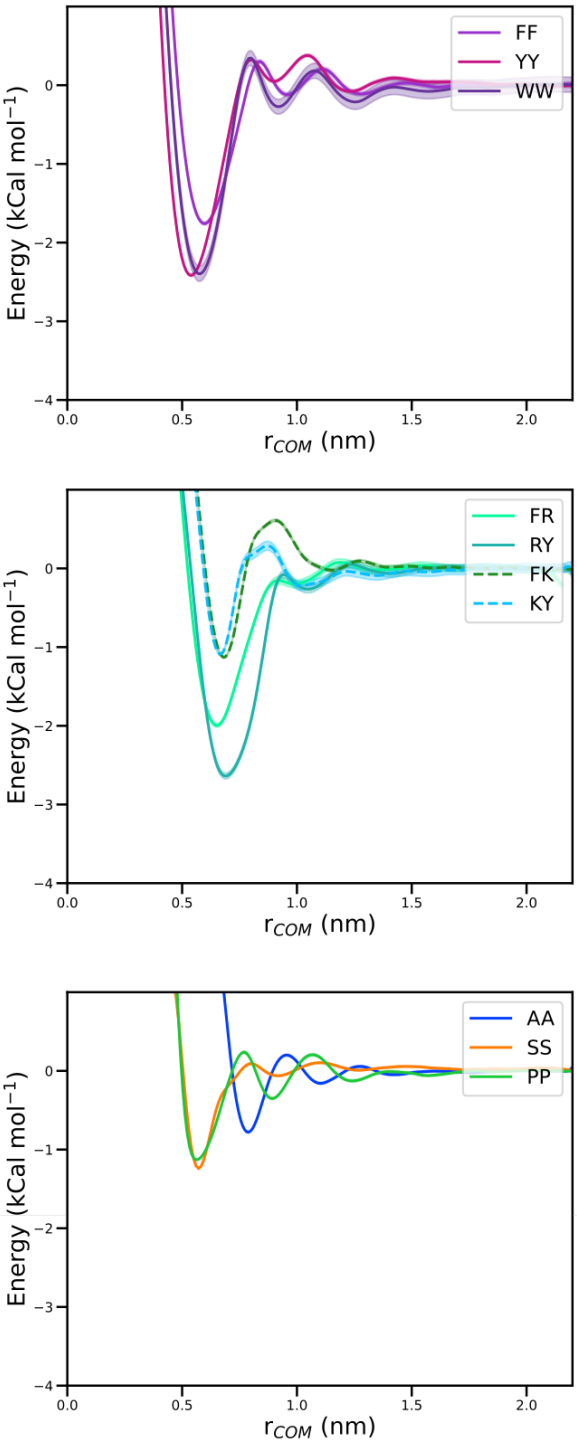


Fig. 5.3 **PMF curves at** 150 mM **NaCl salt concentration for π–π, cation–π and non-π-based interactions**. The curves are presented as a function of the centre-of-mass (COM) distance, with statistical errors (mean±st.deviation) represented as highlighted surface, computed via Bayesian bootstrapping using three independent simulations.

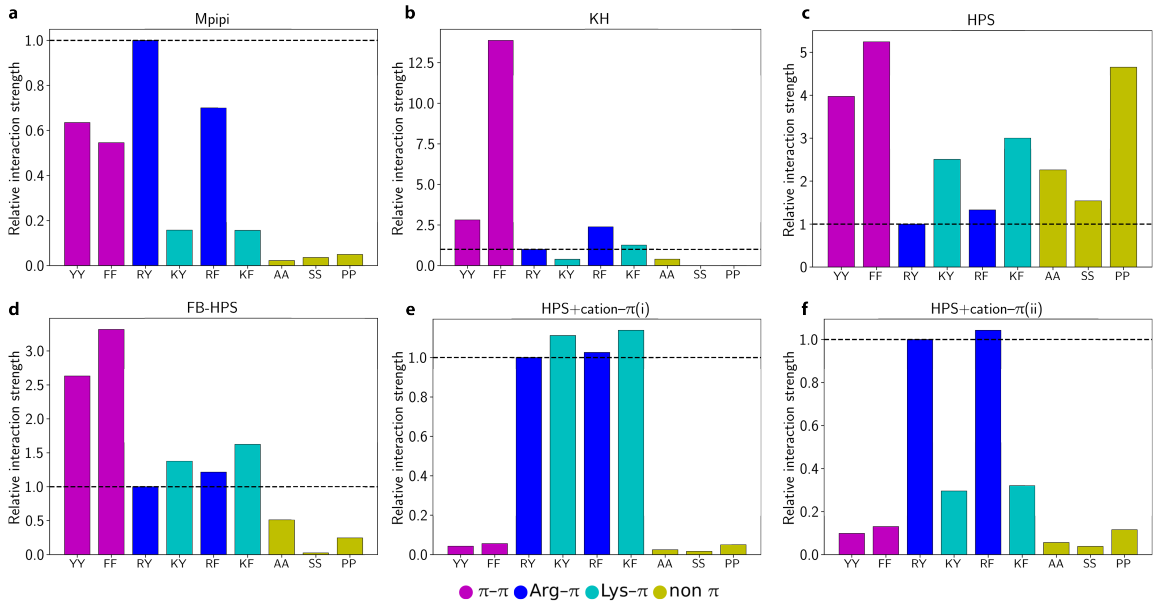


Fig. 5.4 **Relative contributions of π–π, cation–π and non-π-based interactions in different residue-level models.** a to f) Relative interaction strengths for select residue pairs in Mpipi, KH, HPS, FB-HPS, HPS+cation–π(i) and HPS+cation–π(ii) models. For each model, the data are normalized relative to the corresponding Arg–Tyr (RY) interaction. In each plot, a horizontal dashed line at the RY interaction strength is provided for comparison purposes. Aromatic π–π interactions are colored in magenta, Arg–π in blue, Lys–π in cyan, and non-π-based interactions in dark yellow.

5.3.2 Comparison of the relative contributions of π–π, cation–π and non-π-based interactions in residue-level models

To validate our model parameters, we first compare the Mpipi model with other residue-level coarse-grained models; in subsequent sections, we assess how well these models recapitulate experimental phase behavior. Specifically, we focus on the KH (Kim–Hummer) [89, 45], HPS (hydrophobicity scale) [45], and FB–HPS [40] models, as well as the HPS model with augmented cation–π interactions [schemes (i) and (ii)] [42]. Das et al. [42] recently provided a thorough comparison of the KH, HPS, HPS+cation–π(i), and HPS+cation–π(ii) models. Below, we briefly discuss the key features of these models and then evaluate them in terms of the balance of π–π, cation–π and non-π-based interactions.

As previously stated, the Mpipi model differs from other residue-level models in its parameterization of short-ranged interactions. In the KH model [89, 45], short-ranged

interactions (ε_{ij}) are determined by analyzing residue contact statistics [125] of exclusively folded proteins in the PDB. This model has successfully predicted LLPS propensities for DDX4 IDR variants (including the charge-scrambled, F→A, R→K variants compared to the wild-type IDR), described histone proteins in chromatin [55], and provided qualitative insights into the phase behavior of FUS and Laf-1 [45].

On the other hand, the HPS [45] model is widely used for studying biomolecular LLPS [45, 153, 130, 38, 157]. Short-ranged interactions in this model are based on the hydrophobicity scale of Kapcha and Rossky [85]. Each amino acid is assigned a λ_i value, which indicates its 'hydrophobicity', and residue–residue contacts (λ_{ij}) are determined by the arithmetic mean of the λ_i values of each residue [102]. Additionally, the model's absolute energy scale is optimised to recapitulate experimental radii of gyration (R_g) of an IDP subset.

Recently, Dannenhoffer-Lafage and Best [40] reparameterized the short-ranged interactions in the HPS model by employing machine-learning techniques. The model, termed FB-HPS, was optimized against experimental R_g of unfolded, phase-separating and intrinsically disordered proteins.

In another contribution, Das et al. [42] augmented the HPS model so as better to account for cation–π interactions. They presented two schemes: scheme (i), in which Arg/Lys–π interactions are scaled uniformly, and scheme (ii), where Arg/Lys–π interactions vary. Notably, the authors comment that despite these changes, the augmented models fail to capture fully the experimental LLPS propensities of their test set of proteins [42]. We have considered both the two augmented models, termed HPS+cation–π(i) and HPS+cation–π(ii), in this study to achieve a more complete view of how cation–π interactions contribute to biomolecular LLPS.

During our benchmarking, we analyzed the contributions of select – and non--based interactions of various residue-level coarse-grained models. The interaction strengths, i.e. free energies, were obtained by computing the integral of the well depths of the short-ranged potential. Fig. 5.4 summarizes our results.

Our results indicate that in the Mpipi, KH, and FB-HPS models, aromatic residue pairs (magenta bars in Fig. 5.4a,b,d) are considerably stronger than residue pairs not involving π

contacts (dark yellow bars in Fig. 5.4a,b,d). This suggests that YY and FF act as stickers, while AA, SS, and PP behave as spacers in these models, in accordance with the stickers-and-spacers framework. Notably, in the FB-HPS model, glycine has an interaction strength that is stronger than even the aromatic residues due to its strong backbone π–π contacts. However, mutational studies have shown that replacing Tyr with Gly significantly suppresses biomolecular LLPS.

Furthermore, our analysis of Tyr vs Phe interactions revealed that all previous models predict stronger Phe–Phe contacts than Tyr–Tyr interactions (Fig. 5.4). However, experimental evidence suggests that Phe–Phe contacts are weaker than Tyr–Tyr ones. Hence, we do not expect these models to accurately predict LLPS propensities for Tyr vs Phe mutations.

We also examined the relative strengths of cation–π interactions in the coarse-grained models. The HPS+cation–π(ii) model is most similar to our new model in terms of the relative contributions of Arg–π and Lys–π contacts. However, in both the HPS and FB-HPS models, Lys–π interactions are stronger than Arg–π interactions. Additionally, cation–π contacts in the HPS model are comparable to non-π-based interactions, and in the HPS+cation–π(ii) model, cation–π interactions dominate all other types of interactions. These findings suggest that the HPS model may overestimate LLPS propensities of proteins.

5.3.3 Estimating single-molecule radii of gyration

Certain single-molecule properties of proteins, such as the radius of gyration (R_g) in the context of coil-to-globule transitions, are often governed by similar driving forces as bulk liquid–liquid phase separation. Coiling transitions have therefore sometimes been used as a proxy for the upper critical solution temperature (T_c) of liquid–liquid phase separation [189]. Importantly, the strong correlation between single-molecule dimensions and T_c has been used as a target for optimizing coarse-grained LLPS models. Specifically, it is assumed that models that correctly reproduce experimental R_g of single proteins (i.e. at infinite dilution) should accurately predict LLPS propensities.

We tested whether Mpipi can recapitulate experimental R_g of IDPs (Fig. 5.5a). The set of IDPs has a net charge distribution ranging from -44e for ProTα to +16e for Ash1, where

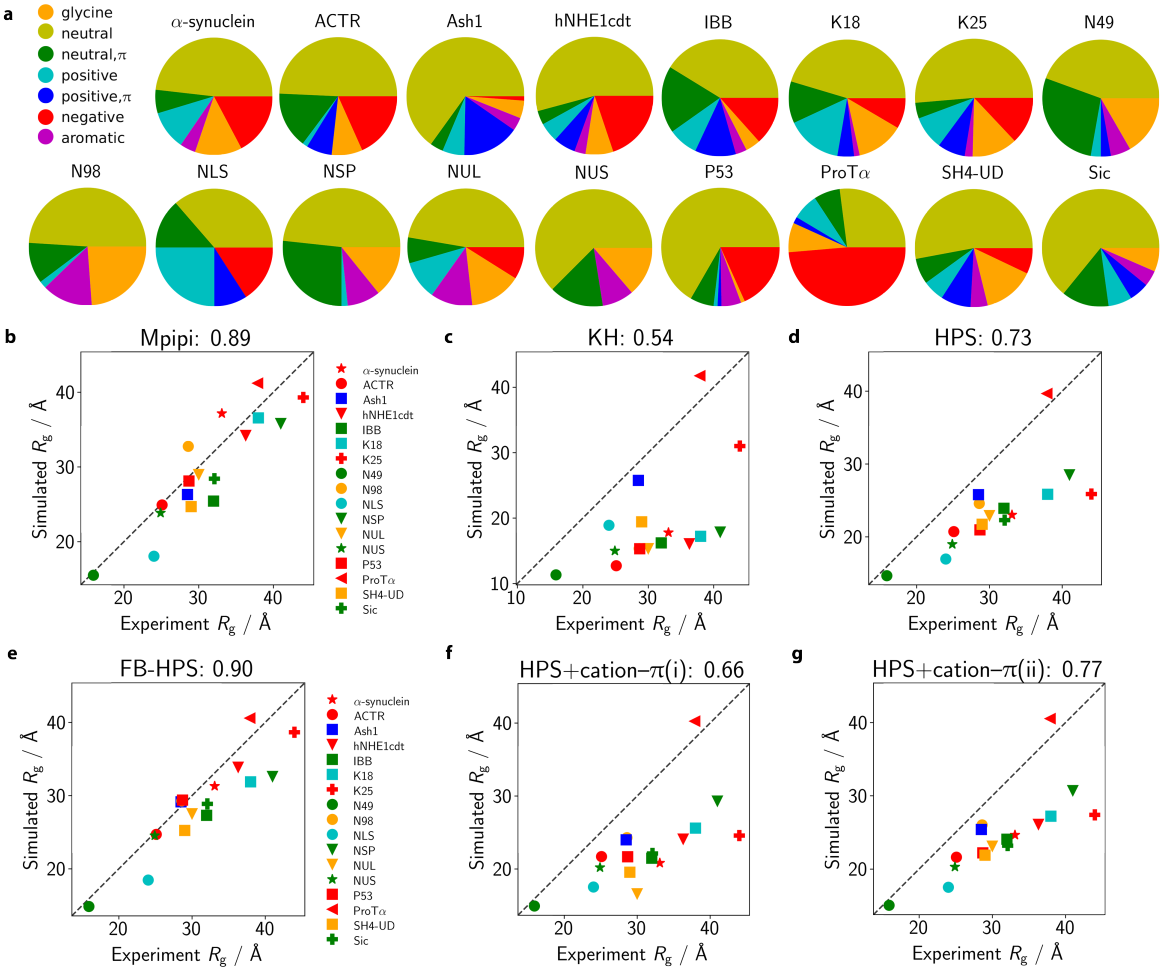


Fig. 5.5 **Comparison of single-molecule radii of gyration with experiment.** a Composition of simulated IDPs. We selected 17 IDPs for which experimental radii of gyration (R_g) data were available. We then assessed the composition of the IDPs in terms of the percentage of glycine (orange), neutral (dark yellow; no net charge at pH 7 and no π electrons in side-chain: A, C, I, L, M, P, S, T, V), neural with π (green; no net charge at pH 7 with π electrons in side chain: N, Q), positive (cyan; without π electrons in side-chain: K), positive with π (blue; with π electrons in side-chain: H, R), negative (red: D, E), aromatic (magenta: F, W, Y) residues. b–g Comparison of simulated and experiment R_g. Each protein is colored based on its dominant residue class (as categorized in a and excluding the 'neutral' class). The broken line represents the 'perfect fit' line. For each model, the Pearson correlation coefficient is reported in the respective figure title.

e is the elementary charge. Therefore, these proteins provide an indirect measure of how well electrostatic and short-ranged pairwise interactions are balanced in the coarse-grained models. Most proteins in our test set are largely comprised of neutral residues that lack π electrons in their side chains. Proteins amenable to single-molecule experiments are likely to have a high content of these neutral residues, which form weaker contacts and result in less aggregation/self-assembly. Apart from this class, the test set of proteins exhibits appreciable variation in protein composition.

In Figure 5.5(b-g), we observe a comparison between the simulated and experimental R_g values for each coarse-grained model. The protein's color is determined by the dominant residue class, as depicted in Fig. 5.5(a), while ignoring the neutral class. It is noteworthy that FB-HPS (Fig. 5.5(c)) achieves the closest match with the experiment, as it was designed to reproduce experimental R_g values. However, the Mpipi model, which was not parameterized on R_g data, performs almost as well as the FB-HPS model (see Fig. 5.5(b)). It is exciting to note that fits to bioinformatics data and atomistic PMFs appear to be physically sound, especially in capturing the single-molecule chain dimensions.

Although the HPS+cation–π(ii) model (Fig. 5.5g), the HPS model (Fig. 5.5d) and the HPS+cation–π(i) model (Fig. 5.5f) show reasonable agreement with experiment, they tend to predict more compact proteins than the actual observed ones. The KH model has the poorest agreement with experiment, probably due to the short-ranged pairwise interactions in the KH model being obtained from residue–residue contacts in folded/globular proteins, which may overestimate the relative strengths of such interactions.

5.3.4 Recapitulating the phase behavior of hnRNPA1 LCD variants

In an effort to evaluate Mpipi potential's ability to capture protein solution properties, we have analyzed the critical temperatures of various hnRNPA1-LCD variants. Recent experimental phase diagrams of these variants, as reported by Bremer et al. [29], align with our previously established experimental validation [106].

We first estimate the experimental critical temperatures [5.6a] of a range of hnRNPA1-LCD variants, following the nomenclature of Ref. 29. The sequences of these variants are

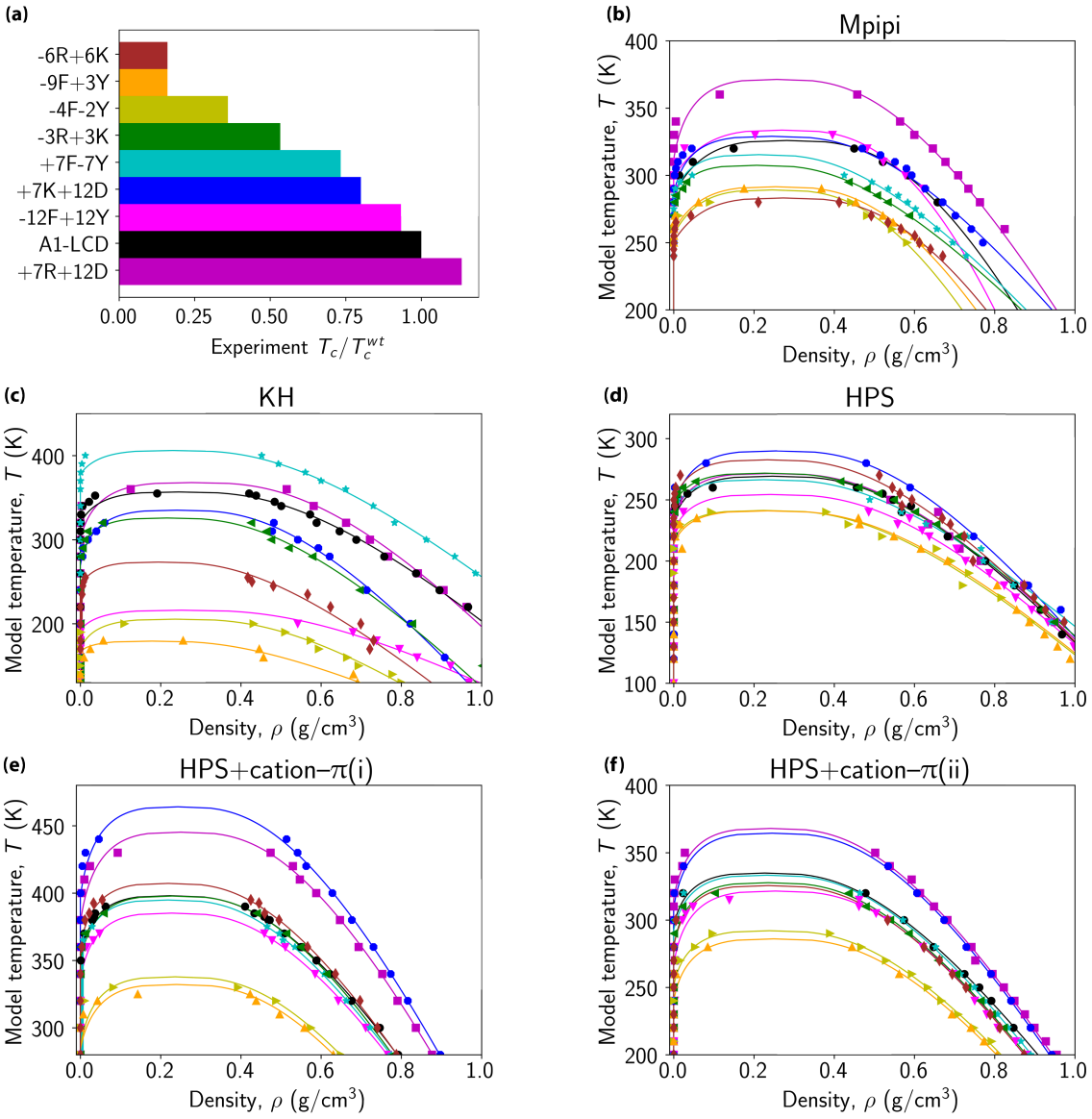


Fig. 5.6 **Recapitulating the phase behaviour of hnRNPA1 LCD variants.** a Nine variants of the hnRNPA1 LCD [including the wild-type LCD (A1-LCD)] are studied in this work, following the work of Bremer et al. [29] To estimate the experimental T_c, we referred to the phase diagrams reported in Ref. 29. The color assigned to each variant in panel a is consistent throughout all the remaining panels Panels b to g display the phase diagrams for each hnRNPA1 LCD variant, obtained through direct-coexistence simulations utilizing the Mpipi, KH, HPS, FB-HPS, HPS+cation–π(i) and HPS+cation–π(ii) models, respectively. Estimation of critical points of phase diagrams is described in Chapter 3. Curves are derived from empirical fits of the data to Eqs (3.13) and (3.14).

given in A.1, We show the computed phase diagrams corresponding to all the models studied in 5.6b–g, and the correlation between simulation and experimental values in 5.6h–m.

All models considered display a positive gradient in the fitted linear regression, indicating that they capture some of the underlying physics. However, the Pearson correlation coefficient shows significant variation across the models. The Mpipi model outperforms the other models by a considerable margin, demonstrating the effectiveness of its parameterisation for both single-molecule properties and bulk behaviour as shown in Figure 5.5.

On the other hand, the FB-HPS model, which mainly was parameterised on R_g data, performs predictably well in predicting the radius of gyration, but only shows marginal improvement in predicting the phase behaviour compared to the underlying HPS potential. This intriguing finding suggests that although R_g properties do correlate with phase behaviour in experimental observations, there are many parameterisations of coarse-grained models that are able to capture one property but not the other.

Removing certain degrees of freedom in modelling protein systems leads to an approximation of interaction energies, which can be considered as approximate free energies.

However, these models may not accurately represent the behaviour of protein systems outside the temperature range they were designed for. Despite this limitation, the fact that these models were parameterized to mimic protein behaviour at room temperature allows us to compare the experimental and simulation temperature scales. It is important to note that there may be some discrepancies in the results due to the limitations of the models. To this end, we can consider the dotted lines shown in 5.6b–g and the difference in slope between the black linear fits shown in 5.6h–m and the $y = x$ lines shown in red. These demonstrate that, at least within the range of experimental data available, the Mpipi model and the HPS+cation–π(ii) model give good predictions for hnRNPA1-LCD, while the HPS and FB-HPS models in particular have all predicted critical temperatures within a much narrower range than experiment.

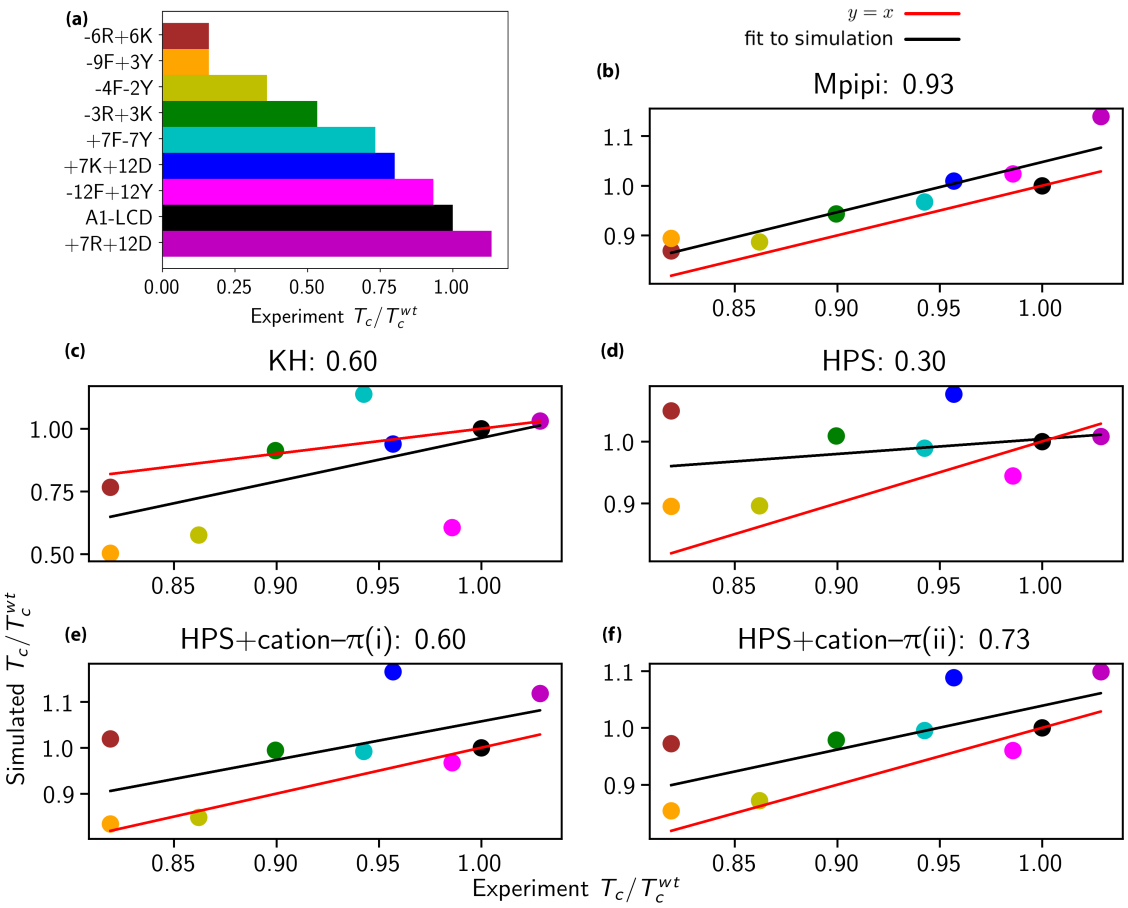


Fig. 5.7 **Simulated versus experimental T_c of hnRNPA1 variants across different LLPS models.** (a) Computed T_c relative to the critical temperature of the wild type (T_c^{WT}) for all hnRNPA1 variants. (b–f)Simulated critical temperature T_c relative to the critical temperature of the wild type (T_c^{WT}) shown against the corresponding experimental value. The Pearson correlation coefficient is provided for each model above each graph.

5.3.5 Probing the LLPS propensities of further proteins

In order to probe the model's transferability, we test its performance for other well-studied IDRs of RNA-binding proteins. In particular, we compute phase diagrams for the prion-like domain (PLD) of fused in sarcoma (FUS) protein, three variants of the arginine/glycine-rich (RGG) domain of LAF-1 (5.8a), four variants of the DDX4 NTD (5.8b), and G3BP1 dimer. For FUS PLD and LAF-1 RGG, experimental critical temperatures are not yet available for direct comparison; however, in vitro studies, including fluorescence microscopy and

temperature-dependent turbidity measurements, provide strong evidence for the relative LLPS propensities of these proteins [144, 157].

In addition, the ABSINTH (self-Assembly of Biomolecules Studied by an Implicit, Novel, Tunable Hamiltonian) [171] potential, which is known to reproduce well experimental conformational ensembles of IDRs, is employed here to obtain estimates of T_c for these proteins. Specifically, using ABSINTH, we compute the temperature for single-molecule collapse (T_θ), which can be used to infer experimental critical temperatures. For example, the critical temperatures of several proteins are well estimated by their corresponding collapse temperatures computed with ABSINTH [119, 189]. However, beyond probing single-molecule properties, ABSINTH is computationally expensive and therefore not applicable to multicomponent LLPS systems.

For the LAF-1 RGG, in vitro studies suggest that the relative ordering of T_c for the WT domain and its mutants is WT>Y→F>R→K [157]. The Mpipi model correctly predicts this trend (5.8). Moreover, the critical temperature of LAF-1 RGG (WT) obtained via the Mpipi model (330 K) coincides with the ABSINTH estimate ($T_\theta \approx 330$ K).

We also employed Mpipi to compute the phase diagram for the FUS PLD (green khaki curve in 5.8a) and estimated the temperature for single-molecule collapse via ABSINTH. Here, our critical temperature estimate (340 K) is 8 K higher than T_θ ($\sim$332 K). Besides an abundance of Tyr residues, about 65 % of FUS PLD is composed on Ser, Gln and Gly residues. As pointed out earlier, these residues are commonly classified as spacers. Thus, we speculate that the discrepancy between T_c and T_θ may suggest that the interactions involving these residues may be too strong within the model. We plan to interrogate this point further as more data become available.

We also compute phase diagrams for four variants of the DDX4 NTD (5.8b). Specifically, we assess the phase behaviour of the WT IDR, the charged-scrambled (CS) variant, a variant where Phe is replaced by Ala (F→A), and one where Arg residues are substituted by Lys (R→K). Brady and colleagues [26] have thoroughly characterised the LLPS propensities of the DDX4 NTD variants. They concluded that, although scrambling charges (i.e. the CS variant) reduce the propensity for LLPS of DDX4 NTD, the F→A and R→K mutations

result in the IDR not exhibiting phase separation at all the conditions tested [26]. Our computed phase diagrams agree qualitatively with these experimental predictions [26], with T_c decreasing in the order WT>CS$\gg$F$\rightarrow$A>R$\rightarrow$K.

In experiments, the highest temperature at which LLPS is observed for the WT and CS variants differs by approximately 30 K at 100 mM salt, whilst in our model, parameterised at 150 mM salt, the predicted critical temperatures for the WT and CS variants differ by 6 K (5.8b). From our simulations, we obtain a lower critical temperature for the CS variant than the WT IDR; however, the difference in critical temperatures is significantly smaller in the simulations than in the *in vitr*o study. Since our coarse-grained beads are isotropic, i.e. they do not explicitly account for orientation-dependent interactions that are likely to be important in charged residues, the effects of charge segregation are less pronounced within our potential. To capture these effects better, it may, in the future, be helpful to include a degree of anisotropic character for some interactions.

These results suggest that the Mpipi model might be underestimating the effects of electrostatics.

In addition to the previous IDRs, we calculate phase diagrams for the full-length FUS protein (FUS WT) and four additional FUS variants, whose protein sequence is provided in A.1. Wang *et al.* used fluorescence imaging to determine the saturation concentration of various FUS mutants [174]. Compared to the WT protein, the 27R and PLD 6D mutants had lower saturation concentrations than the LLPS protein, suggesting that these mutations increase the propensity of LLPS. In contrast, both WT protein-containing mutants PLD Y$\rightarrow$F and RBD R$\rightarrow$G showed higher saturation levels than the WT protein. We calculated the critical temperatures for LLPS for each variant of FUS and found that the order of T_c was consistent with the propensity in vitro to LLPS. That is, the order of T_c is consistent with the propensity to LLPS. We lack extensive experimental phase diagrams for G3BP1, unlike the other proteins we have mentioned, making it difficult to compare simulation results.

Our model indicates that this protein undergoes phase separation at temperatures below 311 K, but its condensates have a structure different from that of the others. Though our trajectories show liquid droplets clearly separated from the light phase, there are visible

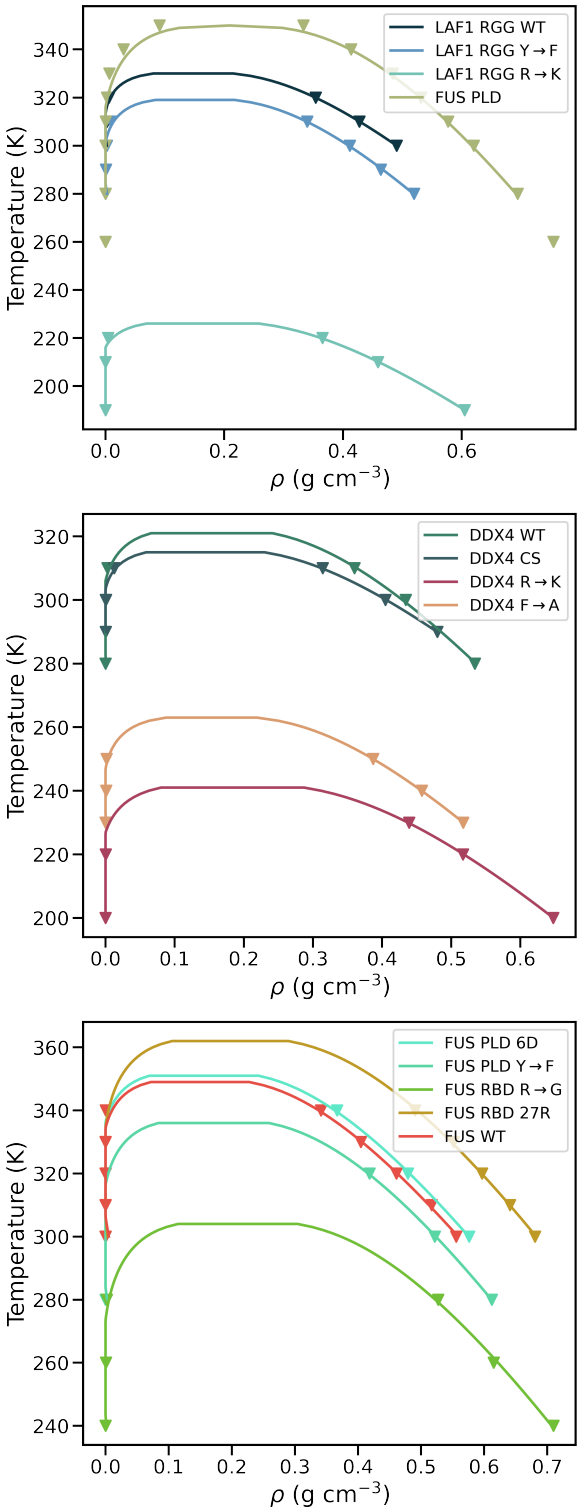


Fig. 5.8 **Phase diagrams of further testing systems, namely, from top to bottom: Laf-1 RGG and FUS PLD, DDX4 and FUS variants.**

holes inside them, even well below the critical temperature. Our simulations suggest that intermolecular interactions largely stem from the EV region, which contains negatively charged amino acids (see Figure A.2). The interactions between the EV and PR regions are also key in bringing the proteins together. However, EV regions are not conducive to phase separation, and we believe electrostatic repulsion causes said inhomogeneities in the condensates, as shown in Figure A.2a. Experiments carried out by Yang *et al*. [182] determined that the globular NTF2 domain is most likely the main driver of LLPS of the protein, while EV regions show an auto-inhibitory effect. Further experimental data is necessary to evaluate the validity of our simulation results, but this could also represent gaps in our CG model.

Assuming Brownian motion [138] our systems, we can extract the diffusion coefficient D from its relation to the Mean Squared Displacement of the system from its original position:

$$\text{MSD} \equiv 2nDt \tag{5.6}$$

, where n is the number of dimensions of the system and t refers to the time. In our case, we studied the diffusion in 3–D. We calculated the MSD from NPT simulations instead, in order to avoid introducing errors due to the presence the interface in the direct coexistence simulations. We obtained the diffusion coefficients from slope for the MSD as a function of time for each protein.

The full wild-type FUS has a diffusion coefficient of $5.49{\cdot}10^{-6}\,\text{cm}^2\,\text{s}^{-1}$, three orders of magnitude higher than the experimentally determined FUS [31]. The simulated diffusion coefficient of low complexity domain of FUS was $2.45{\cdot}10^{-6}\,\text{cm}^2\,\text{s}^{-1}$, yet higher than the experimental value [130] by three orders of magnitude. These results, however, do make qualitative sense, as the PLD is the key interacting region in FUS. Therefore, the diffusion of full FUS is mainly affected by that of the PLD, which is captured by our simulations.

Wild-type DDX4 had a diffusion coefficient of $1.31{\cdot}10^{-5}\,\text{cm}^2\,\text{s}^{-1}$ in our coarse-grained simulations, compared to the experimental value of $7.5{\cdot}10^{-9}\,\text{cm}^2\,\text{s}^{-1}$ [26].

LAF-1 RGG had a simulated diffusion coefficient of $4.10{\cdot}10^{-5}\,\mathrm{cm^2\,s^{-1}}$, almost eight orders of magnitude higher than the experimental value [158]. We obtained $\mathrm{D}=6.88{\cdot}10^{-7}\,\mathrm{cm^2\,s^{-1}}$ for G3BP1 dimer with the Mpipi model. The orders-of-magnitude difference in diffusion between experiments and simulations are to be expected due to coarse-graining, as well as the use of Langevin dynamics, which presents low friction.

5.4 Discussion

Our approach demonstrates the potential of coarse-grained multiscale models, which are derived from a combination of atomistic PMF calculations, bioinformatics data, and experimental observations. This approach proves to be robust in investigating the link between subtle chemical changes in biomolecules and their emergent collective behaviour. The Mpipi model is particularly promising as it can predict both single-molecule radii of gyration and the collective behaviour of proteins in solution, making it a valuable tool in designing experiments and gaining physical insights into LLPS at the microscopic scale.

That being said, the Mpipi is still not a perfect model for all kinds of proteins. The balance between interactions is affected by conflicting evidence on the relative ordering of strength. For example, in the work of Bremer and colleagues [29], Arg–π contacts are determined to be weaker than the corresponding π–π ones, while our work and the work of Wang et al. [174] appear to favour the view that Arg–aromatic interactions are stronger than analogous aromatic–aromatic ones. Furthermore, the data suggest that, in some cases, the precise ordering of these interactions within coarse-grained models is fundamental to recapitulate the observed behaviours, while in other cases an approximate ordering could suffice. For example, in FUS protein we find that the LLPS propensities of the full-length protein versus its PLD domain are highly sensitive to the relative ordering of Arg–π and corresponding π–π contacts [94], whilst our current benchmarks reveal that for the A1-LCD variants, all models that favour Arg–π and π–π contacts over the non-π-based contacts achieve a high Pearson correlation, regardless of the precise ordering of Arg–π and the π–π contacts.

However, the Mpipi model seems to struggle to capture the phase behaviour of proteins whose LLPS are driven by predominantly electrostatic interactions, such as G3BP1. In these cases, the associated error is significantly higher than for the rest of proteins, but still in the error range of the other mentioned coarse-grained models.

Hence, the ordering of these and other interaction strengths is likely to be context specific, and a system-specific coarse-graining strategy may be necessary to achieve satisfactory agreement with experimental data in some cases. Consequently, one set of measurements, be it experiments or simulations, will be unlikely to yield a completely accurate representation.

Our work assumes that the LLPS propensities of biomolecules can be captured by pairwise amino-acid interactions, which can be an oversimplification of the highly collective behaviour driving LLPS. This approximation allows us to create a transferable coarse-grained model that can capture qualitative and quantitative trends for phase-separating systems, especially those characterized in vitro. However, in crowded intracellular environments, three- and higher-body energy terms may contribute significantly to the overall macroscopic behaviour. Cooperative interactions can reshape the phase boundaries of LLPS systems. Therefore, it's important to carefully consider the contribution of such cooperative interactions in intracellular LLPS systems. As discussed above, interaction energies in coarse-grained potentials are effective free energies, and they should, in principle, depend on environmental factors, such as temperature and pH.

In particular, since we have not considered explicit protein–solvent interactions, the solubility of all proteins studied increases with increasing temperature, even when other effects, such as dominant hydrogen bonding at low temperatures, could result in significantly different phase behaviour as the temperature is lowered. Salt concentration is also critical in modulating pairwise interactions and driving reentrant LLPS behaviour [94].

Our approach adds to the set of rigorous tools that are helping to achieve a predictive quantitative description of the influence of amino-acid sequence in biological phase behaviour. Alongside experimental advances, theoretical work, and other computational approaches, the Mpipi model can help discover the molecular mechanisms underpinning phase separation

and provide a new biophysical understanding of how biomolecular condensates are formed, sustained and regulated.

Chapter 6

Design of a sequence-based LLPS model that represents the asymmetry in electrostatic interactions between proteins

In the previous chapter, we presented the Mpipi framework, a coarse-grained potential to study the phase behaviour of tens to hundreds of proteins via computer simulations. The Mpipi model showed great agreement with experimental data and proved that tackling pairwise residue–residue interactions is a key, yet simple approach to develop accurate models to study biological LLPS. In this chapter, we present a more advanced version of Mpipi, which we name Mpipi Recharged, as a model that describes interactions between amino acids to higher detail, and provides a more explicit and precise representation of electrostatic interactions. The latter is introduced as a means to provide a more exact depiction of interactions between charged residues, due to their asymmetry observed via atomistic simulations. The Mpipi Recharged shows enhanced agreement with experimental data at the single–molecule level and in the bulk and aims to be the 'go-to' model to study protein LLPS at physiological conditions.

Contents

6.1 Preamble

This chapter has been conceived as a result of collaborative work, which would not have been possible without the invaluable contributions of each member, who brought their unique expertise and dedication to the project.

As the first co-author in this work, I conceived the project and developed the parameterisation of the coarse-grained model. I also carried out the atomistic simulations and analysis required to create and adjust the terms of the potential and its parameters and carried out the computation of single-molecule radii of gyration of IDPs to validate the model.

Dr. Andres Tejedor performed most of the validation of Mpipi Recharged through the computation of phase diagrams of hnRNPA1, and is, at the time of submission, further validating the model on a wider set of proteins out of the scope of this work. Simulations of DDX4 NTD variants were carried out by both Dr. Tejedor and I. Julia Maristany carried out

finite size analysis of the model and Prof. Jorge Espinosa contributed with his expertise in computational simulations of LLPS.

Eukaryotic cells represent a complex system in which membrane-bound organelles and membraneless organelles (MLOs) collaborate to efficiently carry out the biochemical reactions necessary for sustaining life while being temporally and spatially segregated. MLOs are known to be generated from LLPS of macromolecules such as proteins and nucleic acids [78, 15]. Functional biostructures, which are also referred to as biocondensates, have been the subject of extensive research in recent years. Although the complete physicochemical grammar that governs this phenomenon is not yet fully comprehended, a significant number of driving forces of liquid-liquid phase separation (LLPS) have been elucidated. The mechanisms behind LLPS are complex and involve various factors such as protein-protein interactions, molecular crowding, and post-translational modifications (PTMs). A comprehensive understanding of LLPS and its underlying driving forces is crucial for the development of targeted therapies for various diseases caused by pathologic LLPS [172].

For a comprehensive understanding of the fundamental principles governing biological LLPS, it is essential to conduct both experimental and computational studies. Experimental studies provide valuable insights into the physical and chemical properties of biomolecules involved in LLPS, while computational studies help to elucidate the underlying mechanisms and predict the behaviour of these systems. The combination of these approaches allows for a holistic comprehension of complex biological systems with LLPS. Therefore, having accurate computational models to simulate LLPS is crucial in achieving this goal.

In chapter 5 we presented the Mpipi framework, comprising a coarse-grained potential that makes it possible to simulate tenths to hundreds of proteins and study their phase behaviour at computationally reasonable costs. While this model proved highly accurate and transferable for a wide array of proteins undergoing LLPS, this work aimed to further improve the model by adjusting the parameters of the energy terms and fine-tuning the sizes of the beads and interaction terms. More specifically, we aimed to improve the description of long-range electrostatic interactions, which the original Mpipi showed higher errors in its predictions of highly charged protein sequences.

In this work, we introduce the Mpipi Recharged model, a further iteration of the Mpipi model, where we improve the description of the pairwise interactions by fine-tuning the parameterisation of the pairwise Wang–Frenkel terms, and provide a customisable treatment of the electrostatic interactions at the pairwise level. The latter is introduced to describe the nuances in charged interactions, observed through atomistic simulations to be asymmetric when comparing typically 'attractive' and 'repulsive' interactions.

6.2 Methods

6.2.1 Atomistic PMF calculations

All-atom PMF calculations are performed on a wide range of amino acid pairs, using the same method as at the development of the first Mpipi model in Chapter 5. The amino acid pair set is extended to include repulsive interactions between anions (EE, ED and DD) and cations (RR, KR and KK). All residue pairs are simulated at a salt concentration of 150 mM NaCl. Na^+ and Cl^- ions and water solvent are modelled using the JC-SPC/E-ion/TIP4P/2005 parameterisation from Benavides *et al*. [19].

6.2.1.1 Umbrella Sampling

The interaction between each dimer is probed with umbrella sampling following the methodology described in Chapter 5.2.During production, we limit the movement of heavy atoms by holding them in place with 1 $J\,mol^{-1}\,pm^{-2}$ restraints perpendicular to the pull direction. We also use a harmonic umbrella potential with a pulling spring constant of k=6 $J\,mol^{-1}\,pm^{-2}$ to constrain the distance between interacting pairs, and a 2 fs integration time step.

To conduct each dimer simulation, we use 34 to 40 windows spaced at 50 pm intervals between 0 nm to 2 nm for each umbrella sampling run. As mentioned in the Methods Chapter 5.2, we simulated each window for 10 ns and conducted three independent simulations for each system. The PMF curves were obtained from the last 9 ns of the simulations via WHAM [96] coupled with Bayesian bootstrapping analysis.

6.2.2 Mpipi recharged

The residue-level coarse-grained model is built upon the aforementioned Mpipi models for liquid-liquid phase separation (see the previous Chapter 5). The total energy of the system in this model is the sum of pairwise bonded, non-bonded and electrostatic terms, as shown in equation 5.1. A harmonic potential represents covalent bonds, and pairwise non-bonded interactions are computed via the Wang–Frenkel potential [176].

Electrostatic interactions are modelled via the Yukawa potential [185] instead of the Debye Huckel screening potential we used in the original Mpipi model. The Yukawa potential and the Debye-Hückel potential are both used to describe screened electrostatic interactions, but they are applied in different contexts and have distinct characteristics [185, 44]. The Yukawa potential is a more general form of the screened Coulomb potential. The Yukawa potential has a parameter, which we denote as A, that determines the range of the interaction. By adjusting this parameter, one can model a variety of interaction ranges, making it versatile for different systems. In the Mpipi Recharged model, it concedes every charged residue pair an effective charge, as shown below:

$$E_{\text{electrostatic}} = \sum_i \sum_j \frac{A_{i,j}}{r} \exp(-\kappa r), \qquad r < r_c \tag{6.1}$$

where κ, which we set at 0.126 Å^{-1} is the Debye screening length and $A_{i,j}$ is the Yukawa parameter, in units of e^2 kCal mol^{-1} Å^{-1}, which we specify for each pairwise charged interaction.

6.2.2.1 Refitting of Wang-Frenkel μ parameter and effective interaction lengths σ

The Wang-Frenkel potential is one of the advanced potentials that describe pairwise interactions, especially in the context of complex molecular and ionic systems. One of the terms that play a pivotal role in this potential is the μ term.

The μ term in the Wang–Frenkel potential modulates short-range interactions between particles based on their geometric and electronic structures, in addition to electrostatic and van der Waals forces [175]. The μ term represents these more complex effects, ensuring that

Residue Pair	Charges	A (e^2 kCal mol^{-1} Å^{-1})
D–R	- +	-4.929075
E–R	- +	-4.929075
D–K	- +	-4.337586
E–K	- +	-4.337586
D–D	- -	4.00000
E–E	- -	4.00000
E–D	- -	4.00000
R–R	+ +	4.00000
K–K	+ +	4.00000
K–R	+ +	4.00000
E–H	- +	-2.4645375
D–H	- +	-2.4645375
K–H	+ +	1.4787225
R–H	+ +	1.55655
H–H	+ +	1.0377

Table 6.1 **Possible combinations of charged residue pairs, their charges, and corresponding assigned Yukawa parameters.**

the potential is sensitive to the specific characteristics of the interacting entities rather than treating them as simple spherical particles.

One of the key consequences of this term is the modification of the repulsion or attraction forces at short distances. Traditional potentials can often predict too strong of a repulsion or attraction in certain cases because they do not consider the subtle effects that arise due to the specific nature of the interacting entities. The term acts as a corrective mechanism for these inaccuracies. This can lead to better predictions of material properties, phase transitions, and system behaviours and more accurate results in computational studies of chemical reactions.

At very short distances, where particles come close enough for strong repulsion due to overlapping electron clouds, the term with the larger exponent (i.e., $r^{-\mu}$) will dominate. As μ increases, the potential at very short distances will become steeper and more repulsive. This means particles will experience a stronger force pushing them apart if they come too close.

At intermediate distances, where van der Waals or dispersion attractions dominate, the term with the smaller exponent (usually r^{-6} in Lennard-Jones-like potentials) plays a

more significant role. As μ increases, the attractive well, representing the balance between repulsion and attraction, will become narrower and potentially deeper. Both terms tend to zero at large distances, and the influence of μ diminishes. However, the rate at which the potential approaches zero with increasing distance becomes faster as μ increases. A higher

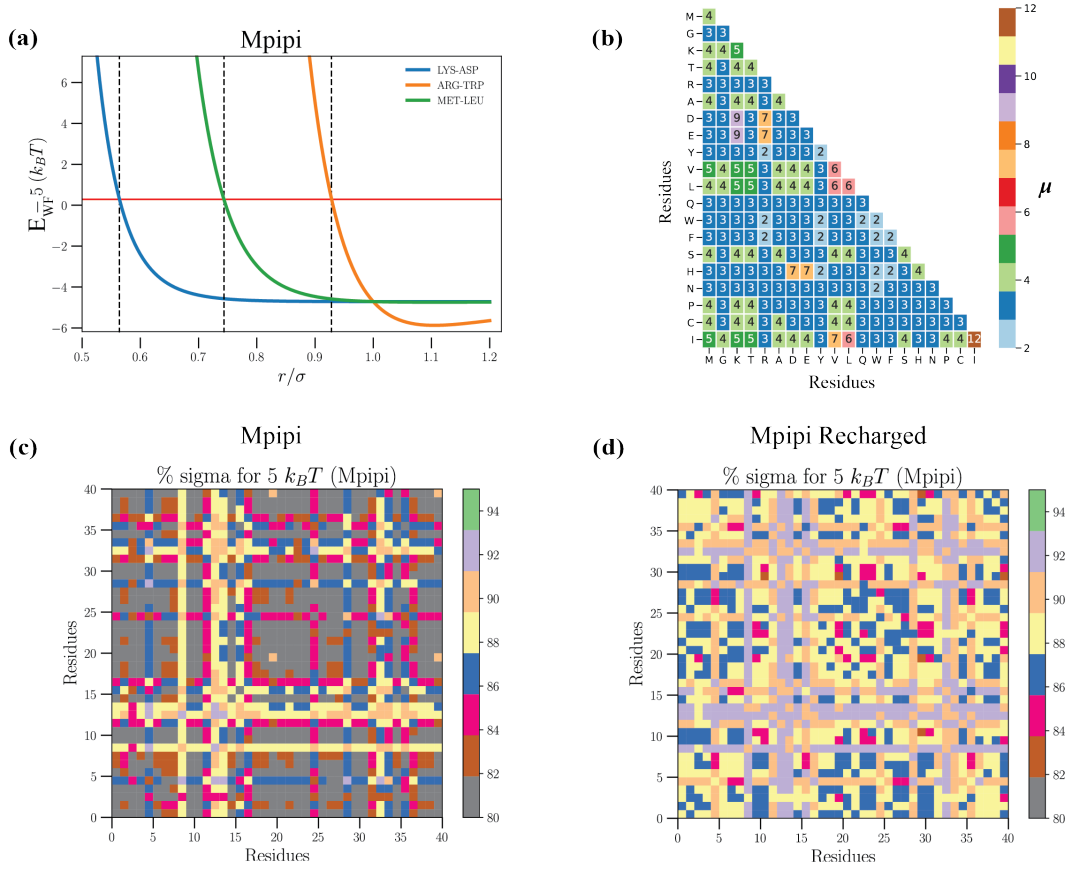


Fig. 6.1 **Shifting of effective σ, bead sizes, via fine-tuning Wang–Frenkel μ exponent.** (a) Wang–Frenkel potential of the Mpipi model vs r/σ with a shift down of 5 k_BT to get the 'real' or effective σ, shown by vertical black dashed lines. Examples computed for pairs Lys–Asp (KE), Arg–Trp (RW) and Met–Leu (ML). (b) New values of μ exponents in Wang–Frenkel pairwise potential in Mpipi Recharged, capturing more accurate effective σ in non-bonded pairwise interactions. Percentual decay of effective σ in the original Mpipi model is shown in (c), where nearly half of the interaction pairs show at least a 10% lower σ_{eff} respective to σ_{real}. After shifting up the μ exponent and changing the ε of pairwise interactions accordingly, the Mpipi recharged model recovers σ_{eff} values closer to the parameterisation values (d). The residue numbering follows the order shown in Table 6.2, and residues indexed 20 to 40 refer to the same residues but in globular domains.

μ value makes the pairwise interactions more specific and short-ranged. The particles are more strongly repelled at short distances and prefer a specific intermediate distance where the attraction is optimised. This can affect phase behaviour and properties like viscosity as the particles become more sensitive to specific spatial arrangements relative to each other.

The original Mpipi model (presented in Ch. 5) was parameterised with $\mu = 2$ except for the pairs I–I and V–I, where $\mu_{I-I} = 11$ and $\mu_{V-I} = 4$. The implication of such low values is that at shorter distances, the repulsive term of the Wang–Frenkel potential is weakened, therefore the minimum of the potential, which should match with the pairwise distance of the interaction pair σ, at which the change of the potential energy curve changes. σ is also referred to as bead or particle size. The effective value of the depth of the WF well is, therefore, shortened, which might lead to residue beads overlapping each other. Such unphysical effects are undesirable and might be a source of significant noise in the predictive capabilities of Mpipi.

Figures 6.1a and c depict the difference between σ and the effective σ_{eff} and the extent of said decrease in effective interaction distance in the Mpipi model. With approximately half of the pairwise interactions showing at least a 10% reduction in effective bead size, the issues that might arise from bead overlapping are non-negligible.

For this reason, in order to avoid bead overlap, we corrected the values of μ in pairwise parameters, followed by ε values accordingly as well, in Mpipi recharged. In this new model, the values of μ are shifted up by a value, η, determined by the ratio between μ and μ_{eff}, below a tolerance of 90%. The shift in μ, η, was determined according to the percentual decrease of σ. In other words, the exponent μ is increased inversely with σ_{eff}, as follows:

σ_{eff}/σ	η
0.5 to 0.6	6
0.6 to 0.7	4
0.7 to 0.8	2
0.8 to 0.9	1

The value of the μ exponents are shown in Figure 6.1b and, together with adjusted values of ε for pairwise WF interactions, are shown in Table 6.2. In this new parameterization, the

Table 6.2 **Wang–Frenkel parameters for the Mpipi Recharged potential. For each residue, two lines are provided. The row highlighted in red lists $\varepsilon/\mathrm{kJ\,mol^{-1}}$ and the row highlighted in orange lists μ. All charged amino acids have $\nu = 1$ and $q = \pm 1e$, as appropriate, except H, which has $q = 0.5e$, where e is the elementary charge.**

	M	G	K	T	R	A	D	E	Y	V	L	Q	W	F	S	H	N	P	C	I
M	0.074	0.123	0.071	0.066	0.288	0.083	0.128	0.135	0.373	0.044	0.049	0.214	0.492	0.351	0.093	0.359	0.208	0.108	0.100	0.040
	4.000	3.000	4.000	4.000	3.000	4.000	3.000	3.000	3.000	5.000	4.000	3.000	3.000	3.000	4.000	3.000	3.000	4.000	4.000	5.000
G		0.173	0.123	0.116	0.335	0.132	0.176	0.183	0.420	0.094	0.099	0.264	0.539	0.397	0.143	0.409	0.258	0.158	0.149	0.089
	3.000	4.000	3.000	3.000	3.000	3.000	3.000	3.000	4.000	4.000	3.000	3.000	3.000	3.000	3.000	3.000	3.000	3.000	3.000	4.000
K			0.046	0.067	0.217	0.085	0.001	0.001	0.184	0.044	0.049	0.224	0.198	0.206	0.096	0.166	0.218	0.112	0.103	0.039
			5.000	4.000	3.000	4.000	9.000	9.000	3.000	5.000	5.000	3.000	3.000	3.000	4.000	3.000	3.000	4.000	4.000	5.000
T				0.059	0.280	0.075	0.121	0.128	0.365	0.037	0.041	0.206	0.485	0.343	0.085	0.352	0.201	0.100	0.092	0.032
				4.000	3.000	4.000	3.000	3.000	3.000	5.000	5.000	3.000	3.000	3.000	4.000	3.000	3.000	4.000	4.000	5.000
R					0.151	0.296	0.007	0.007	0.804	0.259	0.264	0.423	0.934	0.717	0.306	0.208	0.417	0.320	0.313	0.255
					3.000	3.000	7.000	7.000	2.000	3.000	3.000	3.000	2.000	2.000	3.000	3.000	3.000	3.000	3.000	3.000
A						0.091	0.136	0.144	0.381	0.053	0.058	0.223	0.501	0.358	0.102	0.368	0.217	0.117	0.108	0.048
						4.000	3.000	3.000	3.000	4.000	4.000	3.000	3.000	3.000	4.000	3.000	3.000	3.000	4.000	4.000
D							0.138	0.143	0.413	0.099	0.104	0.263	0.529	0.391	0.147	0.008	0.258	0.161	0.153	0.095
							3.000	3.000	3.000	4.000	4.000	3.000	3.000	3.000	3.000	7.000	3.000	3.000	3.000	4.000
E								0.149	0.420	0.107	0.111	0.271	0.536	0.398	0.154	0.008	0.265	0.168	0.160	0.103
								3.000	3.000	4.000	4.000	3.000	3.000	3.000	3.000	7.000	3.000	3.000	3.000	4.000
Y									0.646	0.345	0.349	0.505	0.793	0.623	0.391	0.642	0.500	0.405	0.397	0.340
									2.000	3.000	3.000	3.000	2.000	2.000	3.000	2.000	3.000	3.000	3.000	3.000
V										0.015	0.019	0.184	0.464	0.322	0.064	0.330	0.179	0.078	0.070	0.010
										6.000	6.000	3.000	3.000	3.000	4.000	3.000	3.000	4.000	4.000	7.000
L											0.024	0.189	0.469	0.327	0.068	0.335	0.183	0.083	0.075	0.015
											6.000	3.000	3.000	3.000	4.000	3.000	3.000	4.000	4.000	6.000
Q												0.354	0.625	0.482	0.233	0.499	0.348	0.248	0.240	0.180
												3.000	2.000	3.000	3.000	3.000	3.000	3.000	3.000	3.000
W													0.803	0.770	0.511	0.763	0.620	0.525	0.517	0.460
													2.000	2.000	3.000	2.000	2.000	3.000	3.000	3.000
F														0.600	0.368	0.620	0.477	0.382	0.375	0.318
														2.000	3.000	2.000	3.000	3.000	3.000	3.000
S															0.112	0.379	0.227	0.127	0.119	0.059
															4.000	3.000	3.000	3.000	3.000	4.000
H																0.052	0.494	0.393	0.385	0.325
																4.000	3.000	3.000	3.000	3.000
N																	0.343	0.242	0.234	0.174
																	3.000	3.000	3.000	3.000
P																		0.142	0.134	0.074
																		3.000	3.000	4.000
C																			0.126	0.066
																			3.000	4.000
I																				0.006
																				12.000

effective bead size σ_{eff} is higher than in its predecessor and kept the ratio $\sigma_{\text{eff}}/\sigma$ above 0.8. Due to this shift in μ exponents, the ε values of those $i-j$ interaction pairs also required to be readjusted. As a result, the Mpipi Recharged model shows a wider range of non-bonded interaction energies. The comparison can be seen in Figure 6.3 c and d.

6.3 Results

6.3.1 Refitting of electrostatic terms to represent asymmetry in charged-charged interactions

Another one of the issues raised on Mpipi was its transferability to proteins with predominantly charged–charged interactions. In such cases, the Mpipi model could not perform at a high level of accuracy as with other kinds of proteins, with other predominant kinds of interactions other than cation–π and π–π interactions.

During the development of Mpipi Recharged, we updated the proxy used to calculate relative interaction strength from atomistic PMF curves. While the integral of the attractive portion of the PMF curve allowed for an accurate model of LLPS simulation, we hypothesized that it could be describing torsional and steric effects further from the plain interaction strength. To account for this, we decided to use the depth of the well as a metric, considering only the interaction strength of a pair. When it comes to repulsive interactions, we calculate the point at which the exponential repulsive term of the curves ends.

Upon closer examination of our PMF calculations, we realized that electrostatic interactions are not symmetric. Figures 6.2 a, c, and e show that while interactions between cationic residues are mostly repulsive, anion-anion interactions exhibit slightly attractive potential wells. Even within the same group, there are notable differences: K-R and K-K interactions remain repulsive, whereas R-R interactions are weakly attractive. We do not completely understand this phenomenon but it could be due to the charge delocalization in Arginine residues and other structural properties that promote R-R interactions. As mentioned in the previous section, in this revamped version of Mpipi, we made a significant modification:

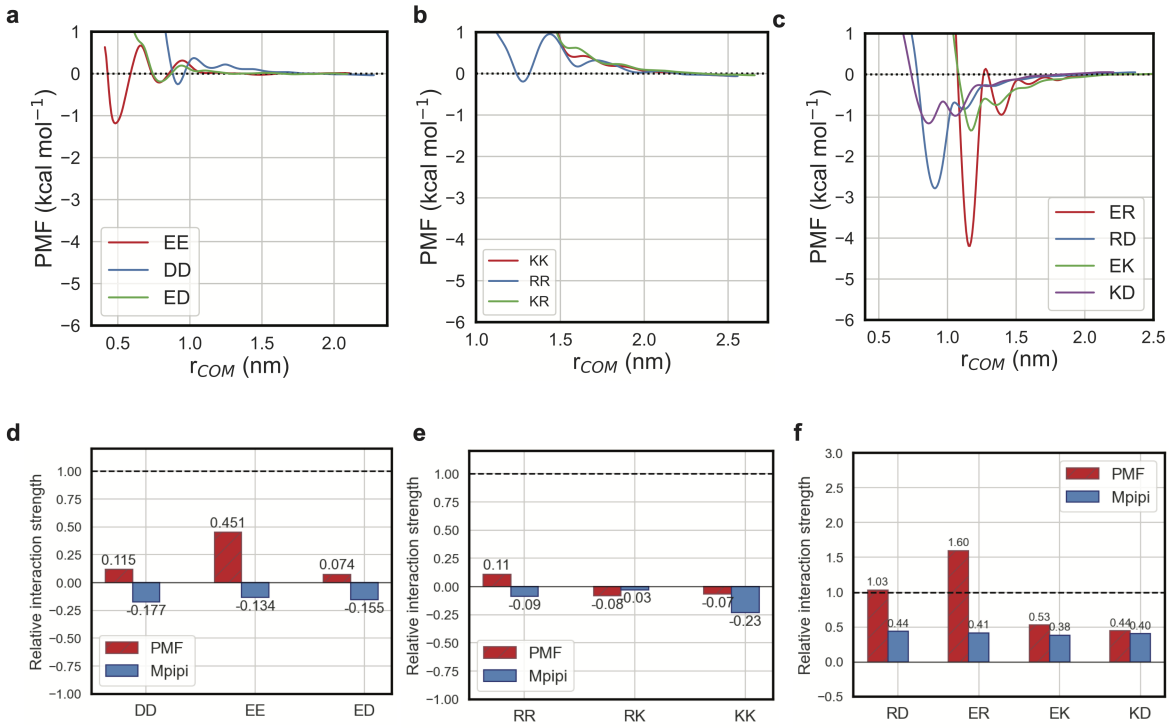


Fig. 6.2 **Asymmetry of non-bonded interactions between charged residues, from atomistic PMF calculations (top) *versus* parameterisation of original Mpipi model (bottom)**. The PMF curves obtained from Umbrella Sampling simulations point out the asymmetric nature of interactions between samely-charged residues (A and B). Interactions between anionic residues (A) are, surprisingly, slightly attractive, while ones between cationic residues are mainly repulsive, as expected, with the exception of R–R interaction (B). The Mpipi model, which accounts for electrostatics with a fixed-charge Debye-Huckel Coulomb term, is not able to capture this phenomenon. Furthermore, it also significantly underestimates the interaction strengths of oppositely charged residues, more specifically E–R and D–R interactions (C and F).

replacing the Debye-Hückel electrostatic potential with the Yukawa potential. This decision was driven by the need for a more flexible and adaptable framework for describing non-bonded interactions within LLPS-driving biomolecular systems (see Figures 6.3a and b). It provides the capability to adjust and optimize interaction energies individually (Figure 6.3b), ensuring that our model is not only more adaptable but also more reflective of the nuanced behaviours we have observed in atomistic simulations and that we aim to capture.

On a similar note, the original Mpipi underestimated the contribution of charged–charged interactions on phase separation. This could easily be explained by the differences between

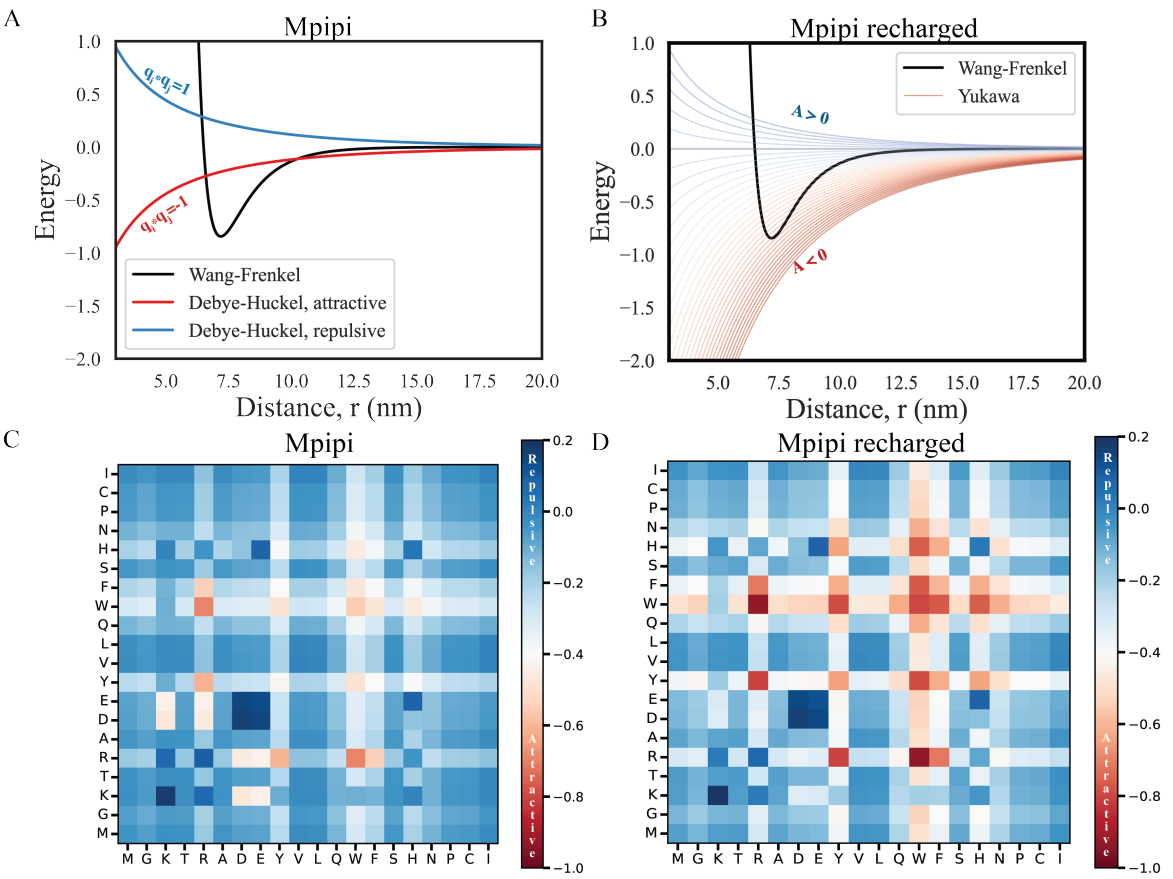


Fig. 6.3 **Comparison of potential energy curves and pairwise interaction energies**. (A) Wang–Frenkel potential energy curve plotted against distance with the inclusion of a Debye–Hückel term for electrostatic interactions. (B) Wang–Frenkel potential energy curve combined with a Yukawa tunable potential for electrostatics. The value of the Yukawa parameter A (see Equation 6.1) determines the range of electrostatic contributions. (C) and (D) are heatmaps representing non-bonded pairwise interaction energies, computed by summing electrostatic and Wang-Frenkel interaction energies, for the respective models, Mpipi and Mpipi Recharged, presented in (A) and (B).

the relative interaction strengths calculated from PMF curves and the Mpipi parameterisation for oppositely charged residues. In our new parameterisation, we took advantage of the flexibility of the Yukawa potential and fine-tuned the electrostatic energies of individual charged interactions. In other words, the repulsions between anion-anion residues were tuned down while R–E and R–D interactions were enhanced.

6.3.2 Calculation of radii of gyration of a set of intrinsically disordered proteins

Transitions between compact and extended states can be indicated by changes in the R_g value over time, providing insight into the protein's conformational flexibility. When using coarse-grained modelling, it is essential to ensure that the simulated R_g values align with experimental observations. In order to test the ability of the new parameterisation to recapitulate single-molecule conformational transitions, we computed the R_g of the same IDPs as in Chapter 5.

Some state-of-the-art computational models like the CALVADOS model from Tesei *et al*. [164, 163]; which has been parameterised to recapitulate experimental data of single-molecule properties of a vast set of IDPs, would very likely score very highly on computing the R_g of our choice of IDPs. However, a high accuracy in single–IDP R_g does not necessarily guarantee the accuracy of the phase diagrams for bulk solutions of the same IDP.

Figure 6.4 shows the radii of gyration for the same set of IDPs as in Chapter 5 with the Mpipi Recharged model. We observed a slight improvement on the accuracy of the calculations, as well as the correlation coefficient for Mpipi Recharged from the original Mpipi. We compared these results to computed R_g measurements for the CALVADOS M2 model [163, 164], which contains a machine-learned force field optimised to fit single-molecule R_g and NMR data of IDPs. As expected, the CALVADOS M2 model recapitulates the radii of gyration of our set of IDPs with a high level of accuracy (see Figure A.3). Interestingly, the Mpipi Recharged model also shows great agreement with experimental data even though its parameterisation was carried out using atomistic PMF calculations at the residue level only.

Overall, single–molecule coil-to-globule transitions are pivotal in their subsequent phase behaviour. It is assumed that models that can predict experimental R_g measurements will likely predict LLPS propensities at higher accuracy levels. Therefore, even though Mpipi Recharged is not parameterised to reproduce single–molecule radii of gyration of IDPs, its

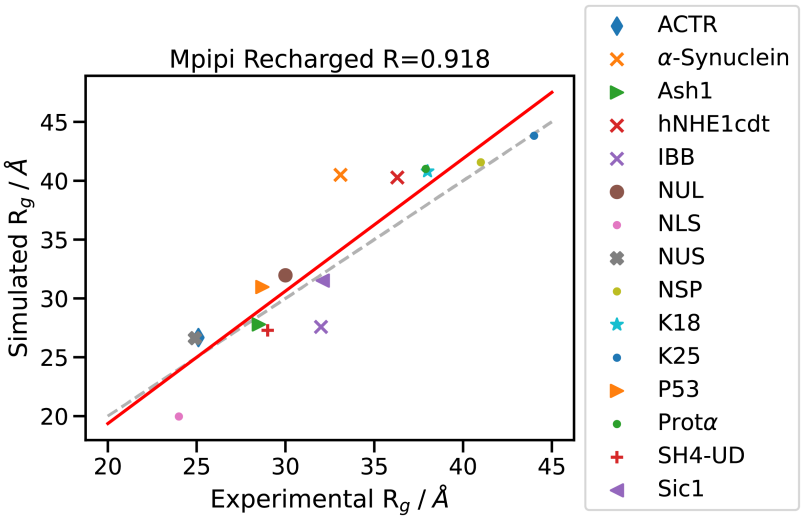


Fig. 6.4 **Comparison of single-molecule radii of gyration of a set of IDPs with experiments**. All calculations were carried out at 300 K and at Debye lengths equivalent to concentrations of NaCl specified in Table A.1, matching the conditions of experiments.

ability to recapitulate the experimental data for this set of IDPs is a reasonable indicator that the interactions of the model are well balanced.

6.3.3 Recapitulating LLPS propensities of hnRNPA1 variants

Following the same proxy as in the evaluation of the original Mpipi model, we studied the phase behaviour of the wild-type (WT) and other eight variants of protein hnRNPA1, whose critical temperatures have been measured in experiments by Bremer *et al*. [29]. These measurements were later validated following the method of Lichtinger *et al*. [106].

The variants of hnRNPA1 span a wide range of mutations in order to assess the ability of the model to capture the relative strengths of different kinds of interactions. We show the correlation between computed and experimental critical temperatures for the original Mpipi and the Mpipi Recharged model in Figure 6.5.

Both models show extraordinary agreement with the experimental phase behaviour and indicate a Pearson correlation coefficient almost identical. We observe a slight improvement in the Mpipi Recharged model to capture the underlying physics underpinning LLPS of these

variants due to the improved treatment of effective interaction distances and a wider range of interaction strengths.

It is also important to mention that the critical points for variants 7K12D and 7R12D were underestimated in the original work of Bremer *et al*., as shown in SI Fig. 2 of Ref. 29; therefore, we consider these two variants to contain a higher error. Although we hypothesize that the Mpipi Recharged model could get a better correlation with experiments if these two variants were not taken into account, its current ability to recapitulate experimental phase diagrams is unparalleled and there is little room for improvement at a reasonable cost and without adding extra terms to the potential.

6.3.4 Validation of charged interactions of Ddx4 NTD variants

In order to further validate the transferability of the model, we computed the phase diagrams for DDX4 NTD variants, such as the wild-type, and specific mutations, including the charge-scrambled (CS), FtoA and RtoK variants, as studied by Brady *et al*. [26].

As depicted by the binodals in Figure 6.6, the more detailed treatment of electrostatics in Mpipi Recharged allows us to obtain better agreement with experimental binodals for DDX4 NTD variants. The difference in critical temperature between the wild type ($T_c = 325.4$) and CS ($T_c = 303.5$) variants is widened compared to the original Mpipi. Mpipi Recharged is able to capture to some extent the effect of charged blocks, or the lack of thereof, in the phase behaviour of DDX4 NTD, even though the potential does not explicitly account for charge patches in the sequence.

6.4 Discussion

In this chapter, we have presented Mpipi Recharged, a coarse-grained model to study protein LLPS that shows promising ability to recapitulate experimental radii of gyration measurements of IDPs and bulk phase diagrams of proteins containing IDR and globular domains. In the context of its predecessor, the Mpipi model, the Mpipi Recharged model brings a more detailed treatment of pairwise interactions, by matching the exponents μ in

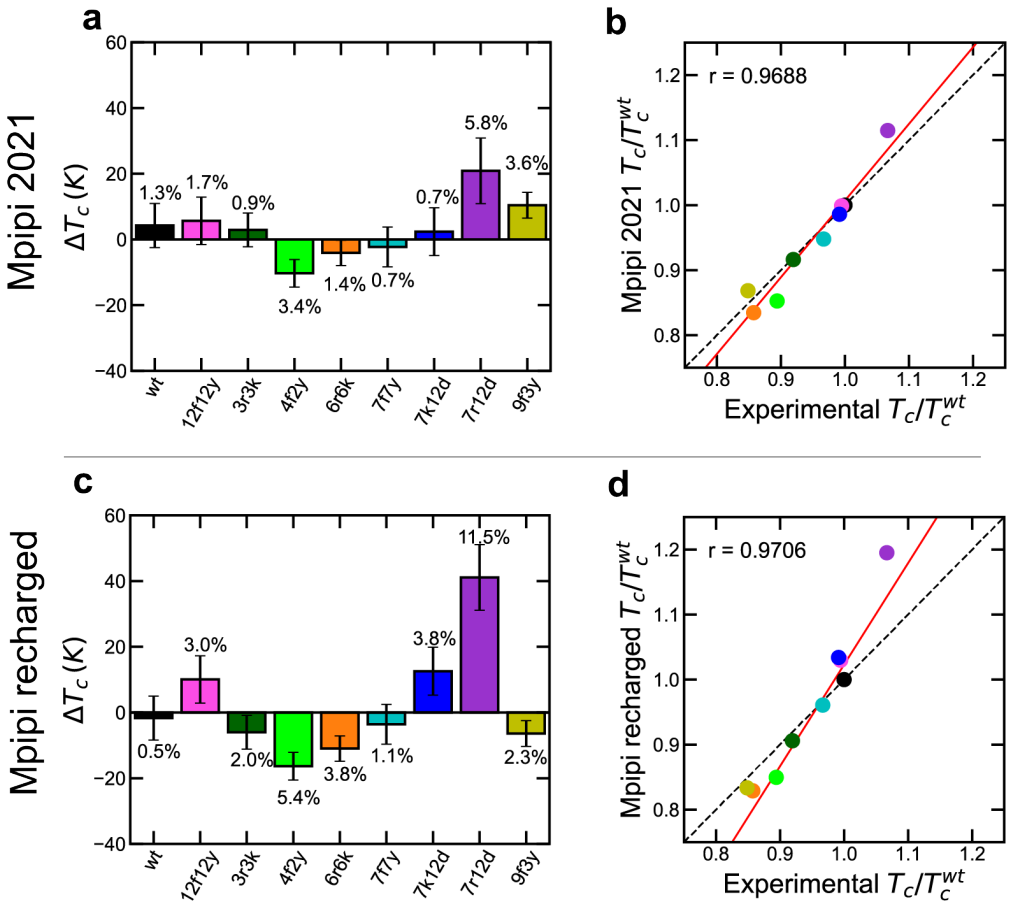


Fig. 6.5 **Critical temperatures of hnRNPA1 variants using the original Mpipi model (top) and Mpipi Recharged (bottom)**. We computed the phase diagrams of nine hnRNPA1 variants, including the wild-type. The barplots in (a) and (c) show the raw difference in critical temperature of the variants with the wild-type, and its corresponding percentual difference annotated on their corresponding bar. For each model, they are followed by respectively (b) and (d), which show the correlation between the computed T_c relative to the wild-type (T_c^{wt}), and their experimental counterpart. The black dashed line represents an ideal 1-to-1 linear correlation between computed and experimental values, while the solid red lines show the correlation curve obtained for both models using linear regression.

the Wang–Frenkel potential to the desired interaction length, σ. This enhancement, coupled with the expansion of relative interaction strengths, has resulted in our simulations aligning more closely with validation experimental data.

Furthermore, our atomistic-resolution simulations of pairwise interactions have allowed us to pinpoint the asymmetry of typically 'repulsive' interactions between charged residue

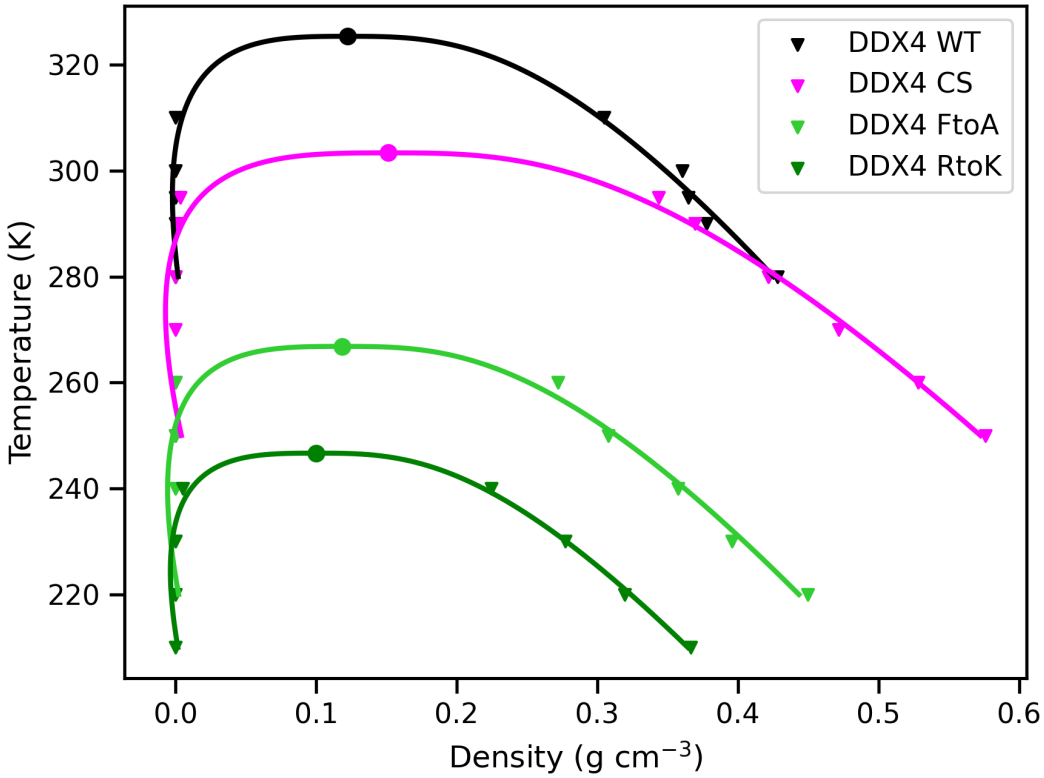


Fig. 6.6 **T–ρ phase diagrams of DDX4 NTD variants at 150 mM NaCl, using Mpipi Recharged.**

pairs. The introduction of the pair-specific treatment of electrostatics through the Yukawa potential confers the model with greater accuracy and flexibility. In particular, we postulate that this additional flexibility of the electrostatic potential term can help us consider further protein–solvent interactions beyond physiological conditions.

However, it is essential to acknowledge certain areas where the model could benefit from further refinement. A salient example is the challenge of residue patterning, especially charge patterning. As emerging experimental studies highlight the pivotal role of patterning in phase behavior, addressing this aspect becomes paramount. Introducing many-body terms to the potential energy function might offer a solution to the intricate task of amino acid patterning.

Overall, our model proves that the physicochemical grammar driving LLPS is mainly dependent on pairwise amino acid interactions since our model shows great predictive power

to simulate LLPS near physiological conditions. To further enhance its applicability across diverse conditions, such as varying temperatures or pH levels, or to incorporate higher-order interactions, we anticipate that machine learning methodologies would be instrumental.

Chapter 7

Effect of Mg^{2+} ions in protein-protein interactions

LLPS in proteins is a phenomenon of growing interest in cellular biology, playing a key role in the formation of biomolecular condensates. The presence of ions, particularly magnesium or Mg^{2+}, can significantly influence this process, yet its mechanistic understanding remains limited. This chapter introduces a novel coarse-grained model designed to elucidate the role of magnesium ions in protein LLPS. The model was meticulously parameterized using high-resolution atomistic simulations, complemented by a comprehensive meta-analysis of magnesium-mediated contacts sourced from the PDB. This dual approach ensured a robust and data-driven foundation for the model. In the absence of extensive quantitative experimental data, our model successfully recapitulated the qualitative phase behaviour of several intranuclear proteins, notably MED1, BRD4, Nanog, and orthologues of DDX4 and DDX3. This achievement underscores the model's potential in predicting and understanding LLPS behaviours in a range of protein systems.

Contents

7.1 Introduction

Biomolecular condensates are ubiquitous multi-component compartments inside cells thought to be sustained and segregated in space simply by the physical chemistry of phase transitions. The idea of intracellular organisation via phase separation emphasises that the seemingly complex properties of multi-component compartments can be explained by the interactions among their biomolecular components and, very importantly, the features of the microenvironment (e.g., salt, pH, temperature). Among the many ions found inside Eukaryotic cells that can regulate the physicochemical properties of condensates, magnesium ions are one of the most abundant (≈20-100 mM [127]) and fascinating. Indeed, the concentration of free Magnesium ions inside cells is highly dynamic and regulated by its complex interaction with ATP [16].

Magnesium is an essential element that guarantees the correct functioning of cells. At the same time, its advantageous physicochemical properties have favoured the incorporation of

this ion into a wide array of biological functions. To this day, magnesium has been observed to participate in a myriad of cellular processes, spanning energy metabolism, nucleic acid synthesis, protein synthesis, signal transduction and many more [23, 178, 43].

Magnesium is a divalent cation characterised by specific physicochemical features: it has a small ionic radius and is consequently endowed with a high charge density [178]. Magnesium has been shown to be an indispensable cofactor in nucleic acid systems [43] The cation not only can screen the electrostatic repulsion among the negatively charged phosphate groups of nucleotides but also establishes associative interactions through both water-mediated and direct binding [43]. Indeed, the coordination of a single magnesium ion to water occurs in an octahedral conformation, with the magnesium ion in the centre, and causes a slower water exchange as compared to other metal cations [178]. Such water molecules are more mobile and can move around the magnesium ion, allowing it to interact more readily with other molecules. These key features aid RNA and DNA to fold into the tertiary structure they require for proper functioning [108, 100, 79, 141].

Magnesium ions have been suggested to also play a key role in regulating the compartmentalization of biomolecules inside cells via the formation of liquid droplets. These droplets, also known as membrane-less organelles or biomolecular condensates, are not enclosed by a membrane but are instead stabilized by the interactions between the biomolecules, usually proteins and nucleic acids, that make up the droplets, and are brought together through a process termed liquid-liquid phase separation [27, 14]. These biomolecular condensates have been demonstrated to play a critical role in cellular function, namely reaction regulation, macromolecular transport, signalling and stress response, along with others. Surprisingly, biomolecular codensates generated from intrinsically disordered proteins (IDPs) have been linked to cellular dysfunction and disease, as precursors of neurodegenerative diseases such as Alzheimer's, amyotrophic lateral sclerosis (ALS), Parkinson's disease, and frontotemporal dementia [172].

From a chemical standpoint, there is plenty of experimental evidence [174] that the interactions between biomolecules driving the generation of phase-separated liquid droplets are based on multivalent, weak and transient interactions. Such molecular interactions mainly

span hydrophobic, electrostatic contacts between the positively charged amino acids in IDR domains (arginine, lysine, histidine) and the negatively charged amino acids(glutamate, aspartate) or the phosphate group in nucleic acids, hydrogen bonds between polar residues, and $\pi - \pi$ and cation–π interactions. Biomolecules or groups of biomolecules with higher contribution to droplet formation that self-associate are classified as 'scaffolds', while 'clients' are thought to play an accommodating function and are recruited by the former [15].

Another critical insight that has emerged from research in this field regarding the formation and behaviour of biomolecular condensates is that they are highly dynamic and regulated by a complex network consisting of an interplay of biochemical and biophysical factors, rather than merely component sequences. For example, recent studies have shown that the properties of the surrounding environment, such as pH [2, 80], salt concentration [94], and the presence of other molecules, can dramatically affect the behaviour of biomolecular condensates.

In the cell nucleus, most Mg^{2+} ions are chelated with ATP. The hydrolysis of the ATP-Mg complex leads to the release of free Mg^{2+} ions to the medium [114]. In *in vitro* experiments have revealed that these free Mg^{2+} ions play a crucial role in chromatin condensation, phase separation of double-stranded DNA, single-stranded DNA, nucleosomes and chromatin, and liquid-to-gel transition of model nucleolus condensates [79, 52, 112, 135, 66].

In addition to this, a recent study by Wright et al. [180] proposes that the phase separation dynamics occurring in the nucleus may be regulated by the interplay between the ATP and free Mg^{2+}, where their concentration balance or their levels are regulated by the chelation of Mg^{2+} to ATP and its release. The study suggests that the presence of free Mg^{2+}ions in the nucleus is necessary to drive the generation of intranuclear condensates.

Despite the numerous studies highlighting the prevalence and role of Mg^{2+} cations in nucleic systems, the contribution of different metal ions in protein dynamics and bulk behaviour has been sparse and has just recently caught up. Experimental evidence points to the need for high concentrations of crowders to induce LLPS of various nuclear proteins, such as MED1, BRD4 [154], as well as proteins undergoing LLPS in the presence of divalent cations such as Ca^{2+}.

Here, we hypothesize that Mg^{2+}ions have a stronger 'kosmotropic' effect on proteins than that of sodium - following the Hofmeister series [142, 84] as well as a higher ion screening capability can significantly alter the interaction network of proteins and, hence, their phase behaviour. In this chapter, we carry out umbrella sampling simulations to calculate the potential of mean force of pairwise protein-protein interaction in the presence of Mg^{2+}ions. By comparing the interaction energies of amino acid pairs in the presence of NaCl and $MgCl_2$, we observe that the salt present in the solvent does indeed significantly affect the interaction strength of charged–charged pairs and, to a lesser extent, interactions involving π–electron-containing residues. Interestingly, not only the presence of the divalent cation but its coordination with water molecules affects the strength at which residue pairs interact. Indeed, we find that Mg^{2+}ions, when pairing to residues through its inner coordination layer, shows a more significant contribution to the net interaction than when mediated through a water layer.

Guided by atomistic calculations of free energies of interactions between amino acids in the presence of Mg^{2+} and insights from magnesium-mediated contacts in the Protein Data Bank, we set out to design a coarse-grained model to capture the bulk behaviour of nuclear condensates by optimising the protein-protein interaction energies in an Mg^{2+}-containing solvent.

7.2 Methods

7.2.1 Atomistic PMF calculations

All-atom PMF calculations are carried out on a vast set of amino-acid pairs, following the same proxy as mentioned in the previous two chapters. In this case, the amino-acid pair set is extended to also anion–anion and cation–cation charged interactions, namely EE, ED and DD, and RR, KR and KK. All the residue pairs are simulated with different salts and different concentrations: at 100 mM and 150 mM of NaCl, at 50 mM, 75 mM and 100 mM of $MgCl_2$, and in absence of ions after net charge neutralisation. Sodium-based salt environments are modelled via the JC-SPC/E-ion/TIP4P/2005 parameterisation [19], while

magnesium in solution is modelled via the *microMg* model from Grotz *et al.* [70, 69] and TIP4P/2005 water [19]. The model is capable of replicating a wide range of experiments, such as measuring the free energy of solvation, coordinating with water and determining its geometry, self-diffusion, and the activity derivative in a solution. Although the magnesium ion parameters were initially created to work with force field AMBER *ff14sb*, there were no significant variations observed in the potential of mean force (PMF) while using the parameters on *ff03ws*, as demonstrated in Figure 7.1.

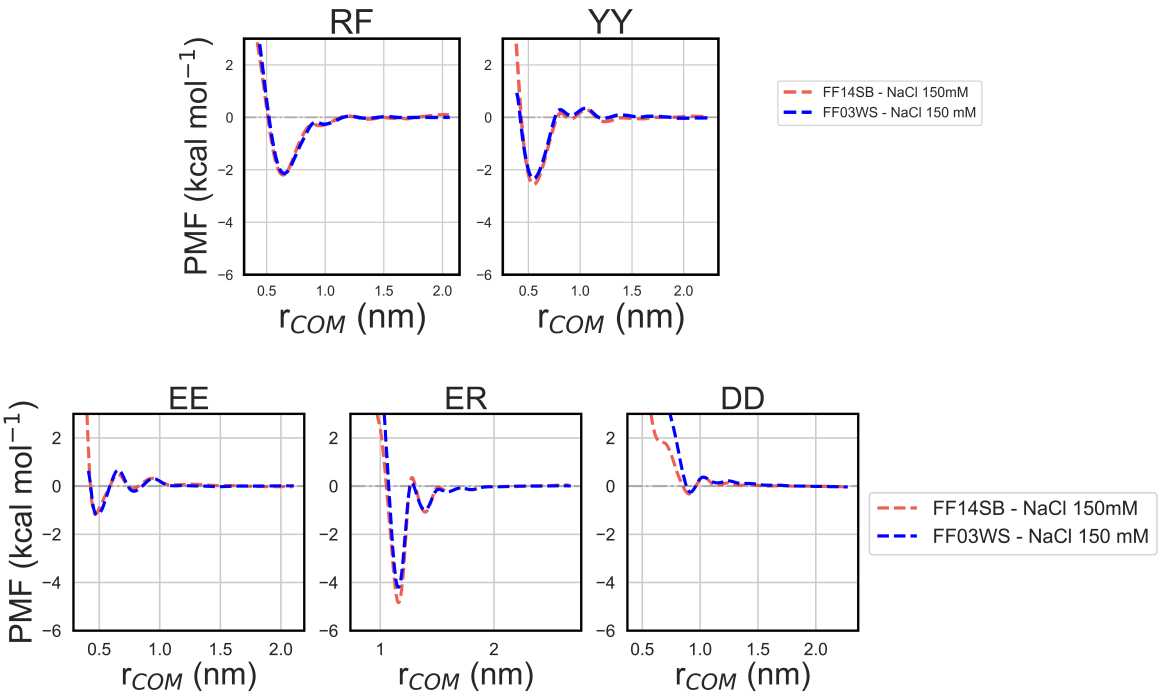


Fig. 7.1 **Comparison of PMF curves obtained for aromatic (top) and charged (bottom) residue-residue pairs with *ff14sb* and *ff03ws*, at** 150 mM **NaCl.**

7.2.1.1 Ad-hoc configurations of magnesium-bonded protein–protein interactions

The water-ion exchange rate in most known solvent and Mg^{2+} force fields is generally over the nanosecond range [115, 104, 6, 70], hence longer than the timescale of PMF simulations. During the simulation setup steps, however, ions are added by replacing water molecules at random locations. This means it would essentially be quite unlikely to sample configurations

where Mg^{2+} ions are bound to the amino-acid side-chain through its first coordination layer. Additionally, we also run several sets of simulations at increasing concentration levels of $MgCl_2$, upon addition of an *ad-hoc* Mg^{2+} ion in-between the residue's sidechains in order to sample interactions mediated by the inner coordination layer.

7.2.1.2 Umbrella Sampling

For the umbrella sampling calculation, we utilized approximately 40 windows, depending on the pair, with a distance between each pair ranging from 0.1 nm to 2 2 nm. Before starting the production runs, we minimized each system with a force tolerance of 500 kJ mol^{-1} nm^{-1} and applied positional restraints of 200,000 kJ mol^{-1} nm^{-2} in each dimension to prevent conformational changes to the starting configuration. As mentioned earlier in Chapter 4, these are within the standard values typically used for these constants. In the simulations, including an *ad-hoc* Mg^{2+} ion, we also applied a force of 10,000 kJ mol^{-1} nm^{-2} to restrain the ion prior to the production runs in all directions and of 1,000 kJ mol^{-1} nm^{-2} in the directions perpendicular to the pulling force on umbrella sampling simulations.

We used the NPT ensemble for production runs, and the chain-wise COM distance is restrained with a harmonic umbrella potential with a force constant of $k = 6000$ kJ mol^{-1} nm^{-2} (see Figure 4.2). The heavy atoms are restrained in the directions perpendicular to the pulling, with positional restraints of 1,000 kJ mol^{-1} nm^{-2}. We applied periodic boundary conditions (PBC), and the electrostatics were computed using Particle-Mesh Ewald [41] summations with a Coulomb cutoff of 0.9 nm. We run three 30 ns–long MD simulations for each calculation with an integration step of 10 fs for each window.

The potential of mean force for each pair at a certain salt concentration is calculated with WHAM from umbrella sampling simulations. The first 1000 ps were discarded to ensure the analysis is carried out on equilibrated trajectories. The statistical uncertainties were evaluated using the Bayesian bootstrapping method [152] using 100 bootstrapping steps.

7.2.2 MagPi coarse–grained model to study protein LLPS in the presence of Mg^{2+}ions

The residue-level coarse-grained model, which we term MagPi, is built upon the aforementioned Mpipi and Mpipi Recharged models for liquid-liquid phase separation (see Chapters 5 and 6). The total energy of the system is the cumulative sum of pairwise bonded, non-bonded and electrostatic terms. Covalent bonds are represented by a harmonic potential, and pairwise non-bonded interactions are computed via the Wang–Frenkel potential [176].

Electrostatic interactions are modelled via the Yukawa potential [185] which provides greater flexibility than a Debye-Huckel Coulombic term, since it concedes every charged residue pair to be assigned an efective charge through the Yukawa parameter A, as shown below:

$$E_{\text{electrostatic}} = \sum_i \sum_j \frac{A_{i,j}}{r} \exp(-\kappa r), \qquad r < r_c \tag{7.1}$$

where κ, which we set at 0.126 Å^{-1} is the Debye screening length and $A_{i,j}$ is the Yukawa parameter, in units of e^2 kCal mol^{-1} Å^{-1}, which we specify for each pairwise charged interaction. The specific Yukawa parameters for charged residue-residue interactions are shown in Table 7.1 The cut-off r_c is set at 35 Å.

7.2.3 Meta-analysis of magnesium-mediated protein contacts

In order to gain a better understanding of how magnesium affects protein interactions and confirm the results obtained from PMF calculations, we decided to examine the contacts made by this divalent cation in chemical structures stored in the Protein Data Bank. We selected PDB structures that included one or more magnesium ions as a ligand and filtered them based on taxonomy (using their Uniprot ID as reference) in order to reduce sequence bias. In total, we analyzed a total of 6,422 structures.

We analyzed the distance between amino acids and Mg^{2+} ions within a 10 Å radius, taking into account the center-of-mass distance and minimum distance. To ensure accuracy, we only considered contacts between residues that were at least five indexes apart, to avoid

Residue Pair	Charges	A (e^2 kCal mol^{-1} Å^{-1})
D–R	- +	-4.4361675
E–R	- +	-4.4361675
D–K	- +	-3.90382739996
E–K	- +	-3.90382739996
D–D	- -	-1.78752
E–E	- -	-3.31968
E–D	- -	-2.5536
R–R	+ +	4.00000
K–K	+ +	4.00000
K–R	+ +	4.00000
E–H	- +	-2.4645375
D–H	- +	-2.4645375
K–H	+ +	1.4787225
R–H	+ +	1.4787225
H–H	+ +	1.0377

Table 7.1 **Possible permutations of charged residue-residue interaction pairs, their individual charges, and corresponding assigned Yukawa parameters.**

counting contacts between adjacent residues in the same chain. Furthermore, we also calculated the relative occurrence of naturally-occurring amino acids.. All these were carried out using GROMACS 2019.3 [20] and BioPython [37] software packages.

7.3 Results

7.3.1 Estimating the effect of magnesium ions in protein-protein interactions

From our PMF curves, we can observe that the mediation of the divalent cation on residue-wise interactions is electrostatic in nature. However, its ability to module the interaction strength is not only limited to residues containing a point charge or a *zwitterion*, but also to non-polar residues with a π orbital with delocalised charge.

Interactions between uncharged residues such as Phe–Phe, Ala–Ala and Pro–Pro are lightly enhanced with NaCl up to 150 mM (see Figures 7.2, A.4). Interestingly, except for the

case of Trp–Trp, NaCl and $MgCl_2$ seem to have opposing effects on π–π interactions. While the energy of the interaction of Tyr–Tyr is reduced by NaCl, the PMF curves of systems with Mg^{2+} show a deepening of the well. We hypothesise this could be due to the hydroxyl (-OH) group of the side chains, to which the divalent cation can bind more favourably, hence stabilising the pairwise interaction. Cation–π interactions do not seem to be affected by salt at the concentrations sampled.

Although the attractive interaction energies at short inter-molecular distances among cation–anion amino acid pairs decrease monotonically with increasing concentrations of Na^+ and hydrated Mg^{2+}, notably smaller concentrations of the latter are needed to induce such decrease. This trend seems to be conserved, at least in the case of NaCl, at notably higher concentrations of 1.5 and 3 M, as reported by Kreiner et al. [94].

Strikingly, the opposite effect is observed in the case of same-charge amino acid pairs. The interactions, which remain predominantly repulsive in the absence of salt, are significantly screened, and even transformed into attractive in the presence of Mg^{2+}. The enhancing effect, nonetheless, is significantly stronger when mediated by the divalent ion. The effect is most notable for interactions between acidic residues (Asp–Asp and Glu–Glu, in Figure 7.2) become highly attractive, even when mediated through the second hydration shell of magnesium.

When the first hydration shell of the divalent cation is free to contact directly the side chains of protein residues, their contribution to the energy of the interaction between amino acids is considerably more robust. The introduction of a single Mg^{2+} cation between the side-chains of a positively charged residue and an aromatic one shows a stabilisation of 0.4644 and 0.8782 kcal mol^{-1} for Arg–Phe and Arg–Tyr, respectively. The sp^2 hybridisation of the guanidino group of arginine might be the key factor in such stabilisation, as compared to Lys–Phe and Lys–Tyr interactions, which remain unaffected by the presence of magnesium at any concentration or configuration tested.

Figure 7.3 reveals similar trends for π–π interactions. While the effects of magnesium and sodium ions are negligible for Trp–Trp, an *ad-hoc* magnesium ion increases the energy of Tyr–Tyr interaction by about 0.453 kcal mol^{-1}, while 150 mM of NaCl reduces the

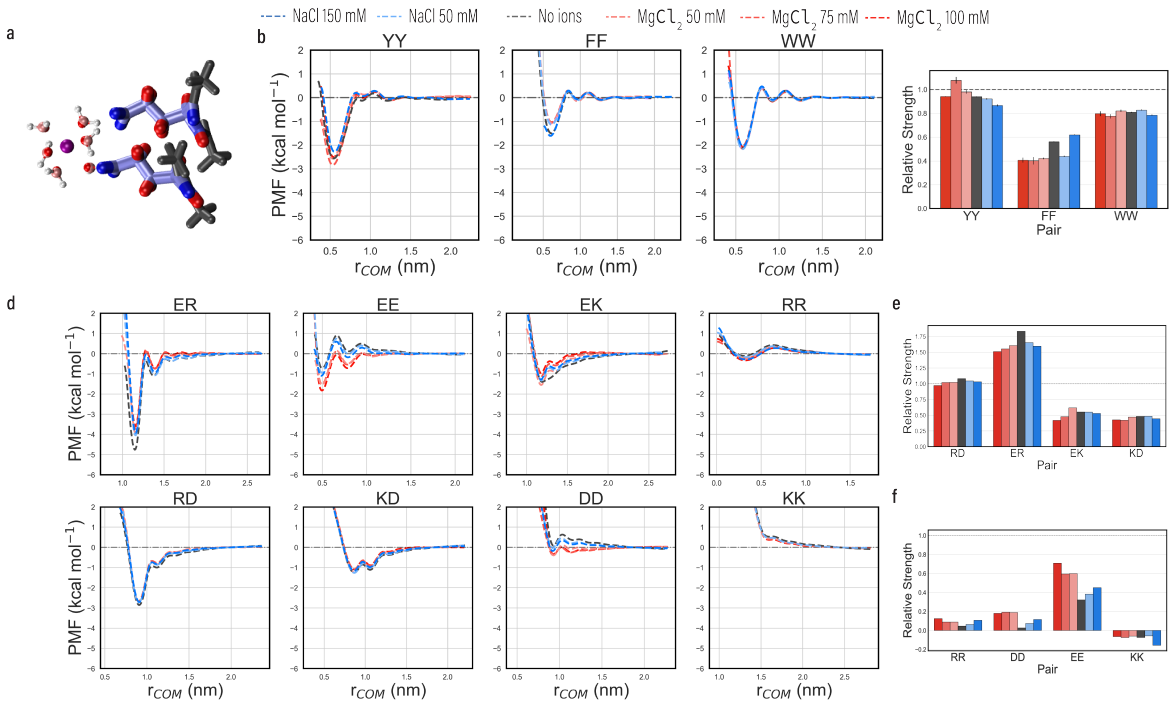


Fig. 7.2 **Outer-shell mediated PMF calculations**. (a) Umbrella Sampling simulations of selected amino acids are carried out using explicit solvent and ions at an all-atom resolution. In the simulation box, ions are randomly introduced to the simulation box. During the simulations, Mg^{2+} ions coordinate with six water molecules in octahedral symmetry to mediate the pairwise interactions. The PMF curves of π–π interactions are shown in (b) at different concentrations of NaCl (blue), $MgCl_2$ (red), and without salt (dark grey). The relative strengths of π–π interactions in different salts and salt concentrations are computed using the well depth of the deepest peak in the PMF curve and then normalized by the value of said peak of the RY interaction with no ions (horizontal dashed line) as shown in (c). The PMF curves at different salts of charged pairs are shown in (d), and the relative strengths of interactions between oppositely charged amino acids and amino acids with the same charge are shown in (e) and (f), respectively.

interaction energy by 0.20 kcal mol^{-1}. The interaction energy of Phe–Phe with an *ad-hoc* magnesium ion follows a similar trend to the one with the hydrated ion, thus decreasing slightly with Mg^{2+} while increasing with Na^{+}

The phenomena observed in electrostatic interactions coordinated to Mg^{2+} ions through a water molecule are intensified when the divalent ion is directly bound to the charged side chains. Commonly electrostatic repulsive interactions become highly attractive in a switch-like manner. In the neutralised systems with no extra ions, the interaction energies, computed from the well depth of the PMF curve, for Glu–Glu and Asp–Asp were -0.89 and

-0.074 kcal mol^{-1}. Upon addition of a single magnesium ion, the energy of the interaction increases to -3.30 kcal mol^{-1} and -1.05 kcal mol^{-1} for Glu–Glu and Asp–Asp, respectively. Attractive interactions between cationic and anionic residues are very sensitive to screening and get diminished upon the addition of magnesium. A similar trend was reported by Krainer et al. [94] at high concentrations of NaCl, but our PMF curves show that a single Mg^{2+}ion reduces the interaction energy by almost an order of magnitude.

A striking key feature we observe through the PMF curves of these *ad-hoc* systems is that a single magnesium ion seems to be sufficient to completely transform the strength of associative and repulsive interactions involving charged amino acids, with additional ions making a less significant contribution. This could explain the millimolar Mg^{2+} concentrations needed in experiments to induce LLPS in nuclear proteins and could be due to a saturation effect. Further addition of ions to the medium can further alter the interaction between amino acid pairs, but not as dramatically as when the first divalent cation is introduced.

These unexpected results can most certainly provide a thermodynamical explanation for the role of Mg^{2+} ions in stabilising protein structures, more specifically those inside the nucleus. Experimental data has pointed out that nuclear proteins contain a higher than-average number of aromatic and acidic residues, which can be related to the switch-like behaviour of Asp and Glu in the presence of magnesium.

7.3.2 Designing a coarse-grained model for Mg^{2+}–driven biomolecular LLPS

Calculations of the potential of mean force are powerful in determining the thermodynamic aspects of pairwise residue interactions and the impact of different salt concentrations and ions in the environment. However, due to the need to keep computing costs low, these calculations can only run for a limited number of steps and timescales, and for small systems. As a result, they have limited ability to reveal the impact of collective phenomena, for instance, the impact of amino acid interactions in biomolecular structures and LLPS. It should be noted that the formation of biomolecular condensates is a complex process, influenced by

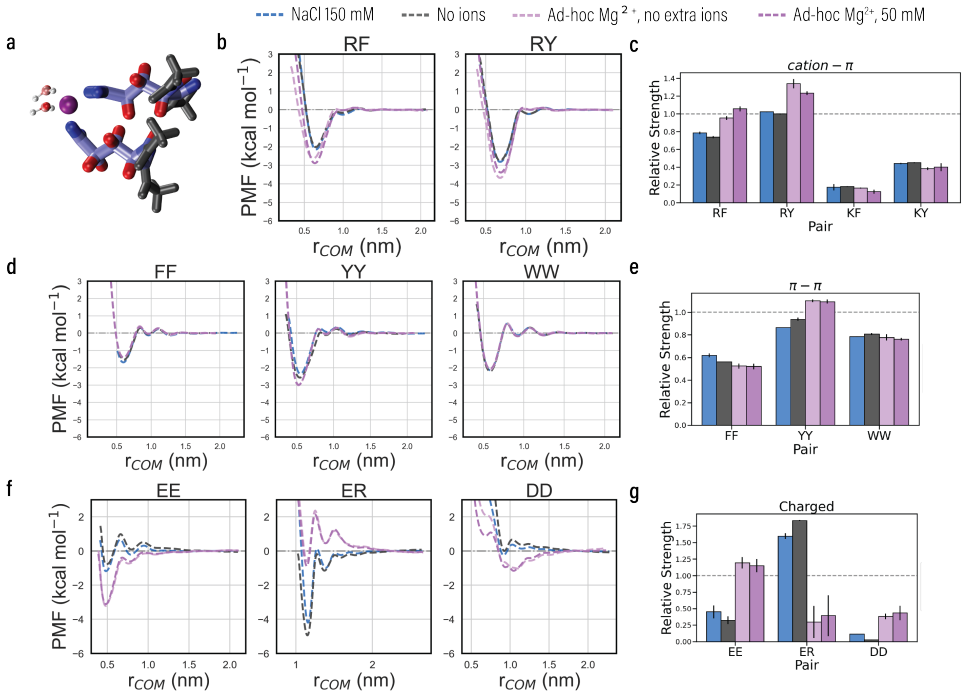


Fig. 7.3 **Inner–shell mediated PMF calculations**. (a) In the PMFs of magnesium-mediated interactions in the inner shell, a single Mg^{2+}ion is positioned between the side chains of an amino acid pair. Additional ions are introduced randomly in the simulation box at 75 mM $MgCl_2$. (b) PMF curves demonstrate cation-π interactions with NaCl (blue), various concentrations of $MgCl_2$ (purple), and without extra salt during neutralization (dark grey). (c) The relative strengths of cation-π interactions in different salt concentrations are computed using the well depth of the deepest peak in the PMF curve and normalized by the value of the RY interaction peak with no ions (horizontal dashed line). (d) PMF curves at different salts of π–π pairs. (e) Relative strengths of interactions between aromatic amino acids. (F) PMF curves of samely charged pairs Glu–Glu and Asp–Asp, and oppositely charged pair Glu–Arg. (g) Relative strengths of interactions between charged pairs Glu–Glu, Glu–Arg, Asp–Asp.

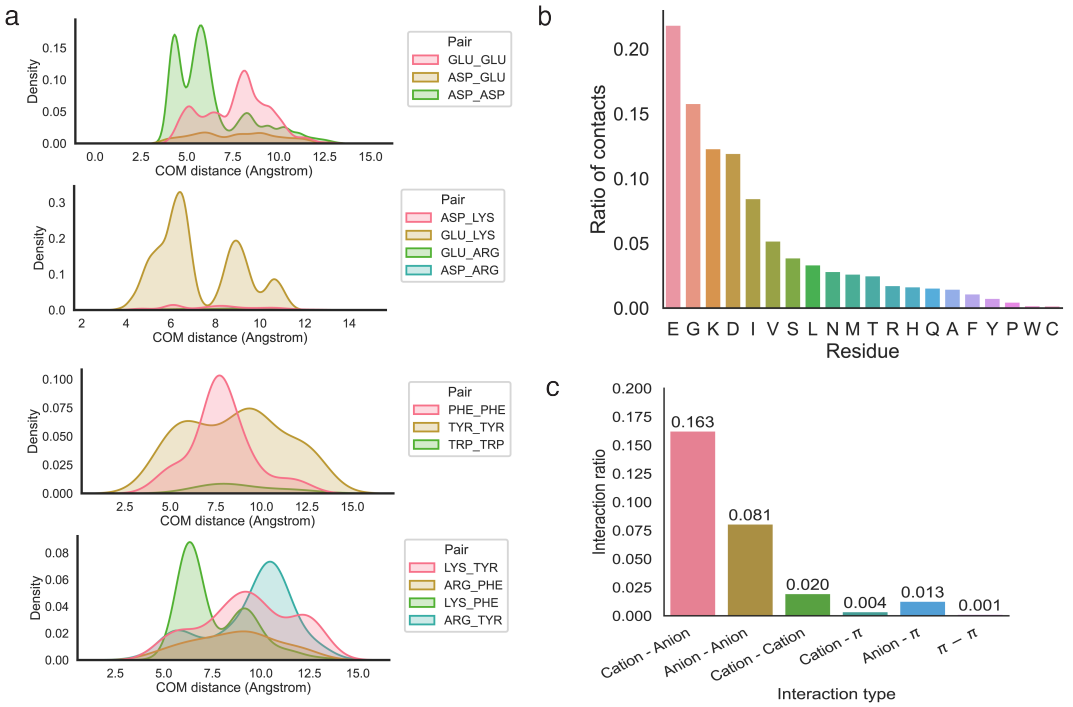


Fig. 7.4 **Meta-analysis of Mg^{2+} dominated contacts on PDB data**: (a) Kernel Density Estimation (KDE) as a function of centre-of-mass distance for Mg^{2+}2–mediated interactions, by groups, from top to bottom: anion-anion, anion-cation,π–π and cation–π. . (b) Mg^{2+}–mediated contact interaction frequencies for each residue type, (c) The ratio of protein residuewise interactions, per type.

a network of forces that collectively encourage the co-localisation of biomolecules while separating them from the cytoplasm or nucleoplasm.

The meta-analysis of PDB structures highlights the role of Asp and Glu as major magnesium ion binders (see Figure 7.4). Our analysis shows that interactions between negatively charged residues make up 8.1% of those mediated by magnesium ions, with aspartic residues being the most prevalent in those interactions. Interactions between oppositely charged residues are the next most abundant, and lysines seem to be favoured over arginine residues. Arginine-based electrostatic attractive interactions are originally more favourable from an energetic point of view (see PMFs in the presence of NaCl or PMFs of neutral systems in Figure 7.2), but arginine occupies more volume, and the charge is less localised. We

speculate that the forces that bring these magnesium ions into the moieties might find it less costly, collectively, to do so, favouring Lysine-based contacts over Arginine ones.

Although cation-π and π-π contacts are prevalent throughout the proteome, our bioinformatic analyses show that only a small percentage of them are facilitated by the divalent cation. On an interesting note, our PMF curves, in agreement with the contact analysis, show that Mg^{2+} rarely mediate tryptophan-containing contacts.

The abundance of aliphatic residues such as glycine, leucine, isoleucine, and valine in magnesium-mediated contacts may be due to their high occurrence in protein sequences and their ability to aid in more energetic interactions as a 'client', which hence explains their high occurrence in our contact analysis, as shown in Figure 7.4b.

Overall, our analysis highlights the important role of magnesium in regulating electrostatic interactions, particularly between oppositely charged residues. By using these insights, we were able to parameterize our Mg^{2+}-containing coarse-grained model using energy calculations and bioinformatics data to gain further insight into which residue-residue interactions are mediated by divalent salts in protein systems.

The first and main variation of pairwise interactions in our model is the replacement of repulsive electrostatic forces between acidic residues for weak attractive forces by adjusting the Yukawa electrostatic term parameter A for each E–E, E–D and D–D interaction. All-atom PMF calculations of these pairs in the presence of magnesium ions show that interaction strength in these pairs is stronger for glutamic acid than for aspartic acid. Therefore E–E > E–D > D–D, which is represented by the absolute value of parameter A (see Table 7.1). Acidic residues are very common inside the cell nucleus and in the absence of magnesium ions or crowing agents, their pairwise repulsive electrostatics might explain their difficulty in forming distinct droplets *in vitro*, as shown in Ref. Sabari et al. [154] Figure 4c in, for instance, the case of MED1-IDR and BRD4-IDR, and nucleolar proteins NPM1 [122] and Fib1 [183].

In addition, we observed in our atomistic simulations that magnesium significantly weakens electrostatic interactions between oppositely charged residues since cationic residues compete with magnesium ions for bonding with acidic residues. This effect is negligible for

Na^{+} ions. Hence, our model also includes this effect, represented by a lowered electrostatic constant, customised to follow the trend E–R > D –R » E–K > D–K (see Table 7.1).

7.3.3 Validation of model on LLPS of intranuclear proteins

To confirm the accuracy of our model, we performed direct-coexistence simulations on MED1-IDR, BRD4-IDR, and the C-terminal domain (CTD) of Nanog. Although there is a lack of detailed quantitative phase diagrams for these proteins, previous studies have shown that they need a significant amount of PEG to create droplets in vitro. Furthermore, their complete sequence has been observed to form super-enhancer droplets within the cell nucleus [154]. Additionally, acidic residues play a crucial role as LLPS-driving dominant forces [155] in intranuclear biocondensates.

We also carried out DC simulations (Methodology available in Chapter 5.2) assessing the phase behaviour of these proteins in the presence of sodium chloride at physiological conditions as a reference for further validation of our model. For this task, we used the Mpipi Recharged model, also developed in Chapter 6, due to its proven accuracy, its well-balanced electrostatic, cation–π and π–π interactions and its similar energy scale to our model. Moreover, although Mpipi Recharged has been parameterised to represent the interactions at physiological conditions of NaCl only, its Yukawa potential proves as an advantageous tool to fine-tune the electrostatic contributions only upon addition of Mg^{2+}ions to the new model.

Our simulations indicate that, as we expected, none of the proteins tested can undergo phase separation at room temperature without magnesium ions or crowding agents, as shown by dashed lines in Figure 7.5b and c. However, when Mg^{2+} ions are present in the implicit solvent, the interaction network of these proteins changes, which enhances their ability to drive LLPS. The three IDRs simulated with our MagPi model show critical temperatures that are significantly higher than room temperature. Interestingly, the addition of magnesium makes MED1 IDR have a higher critical point than BRD4 IDR.

In order to further probe the qualitative accuracy of the model, we probed the LLPS propensities of a set of DDX4 orthologues and paralogous DDX3 orthologues, namely

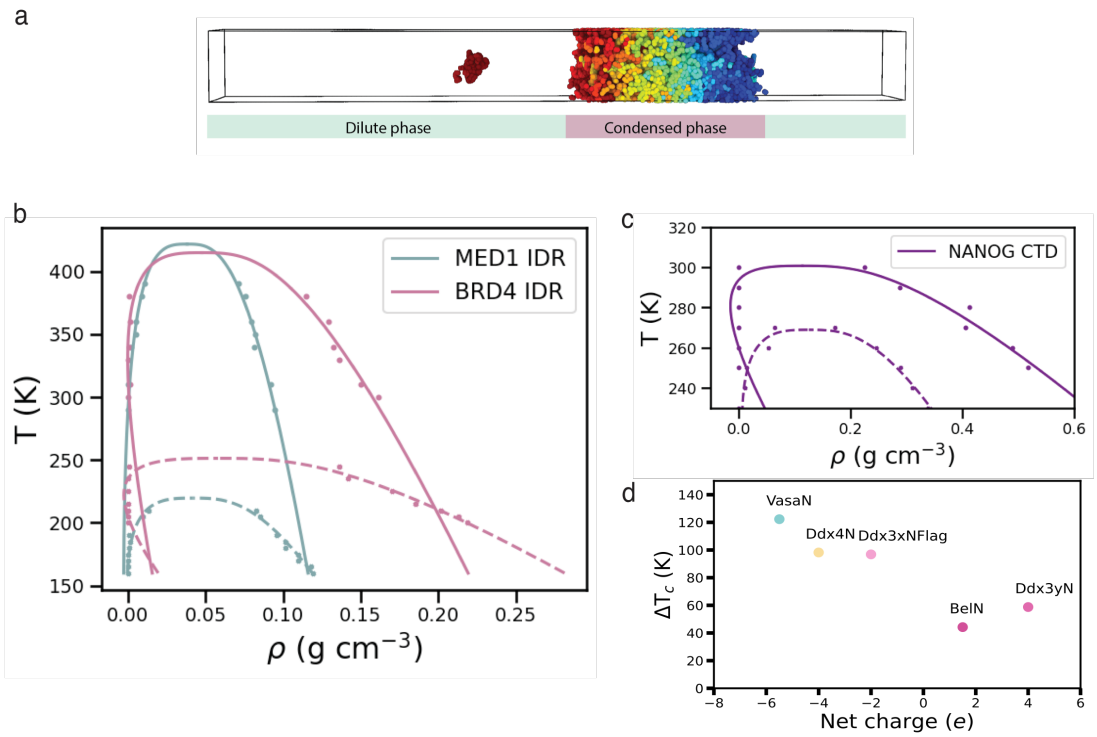


Fig. 7.5 **Predicting LLPS propensities of intranuclear proteins with and without Mg^{2+}.** (a) Snapshot of a direct coexistence simulation. 100 protein molecules/chains were used in simulations. The colour code follows the chain number. (b) phase diagrams of the intrinsically disordered domains of nuclear mediator complex proteins MED1 (teal) and BRD4 (pink). PDs with continuous lines represent the phase behaviour of MED1 and BRD4 when the solvent contains Mg^{2+} ions, while dashed lines are computed at physiological concentrations of NaCl only. (c) Phase diagrams of the C-terminal domain of Nanog protein with (continuous line) and without (dashed) magnesium–mediated interactions. (d) Increase of T_c of DDX4 and DDX3 orthologues as a function of net charge.

DDX4N, BelN, DDX3yN, DDX3xNFlag and VasaN (sequences can be found in Appendix A.1). Crabtree and colleagues [39] reported that upon the addition of millimolar concentrations of calcium chloride, the ability to go under phase transitions for these constructs is not only enhanced but also correlated with their net charge. Calcium ions, although less prevalent inside the cell, have, like magnesium, a charge of $+2$ and coordinates with six water molecules in octahedral geometry on its first hydration layer. Although the effect of Ca^{2+} on residue–residue interactions seems to be milder than that of Mg^{2+}, Ca^{2+} has also been reported to interact with COO^{-} groups of acidic residues in calcium-binding sites [132].

We henceforth hypothesize that Mg^{2+} ions might show a similar effect on inducing LLPS of intranuclear proteins at significantly lower required concentrations of the ion.

Similarly to the trends observed by Crabtree *et al*. [39] for DDX4 and DDX3 orthologues in the presence of 3 mM $CaCl_2$, our model is able to capture the equivalent trends of difference in critical temperature upon the addition of magnesium ions. As shown in Figure 7.5d, DDX4 orthologs with negative net charge tend to phase separate at significantly higher temperatures than their counterparts with higher / more positive net charge when Mg^{2+} ions are in the media. Although the experimental trends reported by Crabtree and colleagues [39] are in the range between 0 to around 0 to 10 °C degrees difference in transition temperature with calcium chloride, we expected these differences to be overall larger for magnesium chloride due to its higher ionic strength and small ionic radius. We observed differences one order of magnitude larger in our magnesium-induced LLPS simulations.

Further experimental work is needed to verify the numerical accuracy of these differences. However, recovering the trends in critical temperature and net charge are clear indicators that our magnesium model is quite reliable for studying the phase behaviour of proteins in more realistic physiological conditions.

7.4 Discussion

As mentioned at the beginning of this chapter, the cytoplasm and nucleoplasm of eukaryotic cells are highly complex solutions, where a wide array of ions, namely Na^{+}, K^{+}, Mg^{2+}, etc; are flowing in and out while participating in a myriad of biochemical processes. Most biomolecular simulation models, however, have been designed to account for sodium-only-based solvents, and yet the effect of other polyvalent cations in biomolecular interactions is to be fully understood.

In this chapter, we have developed a novel residue-level coarse-grained model for LLPS transferrable to intranuclear environments and biomolecules. Our model, which we termed MagPi, aimed to address the limitations of existing models by incorporating the unique effects of Mg^{2+} on LLPS, more specifically in intranuclear systems, generally more crowded than

cytoplasmic ones. Throughout this section, we will highlight the key findings, implications, and limitations of our model, as well as potential future directions for further research.

Our coarse-grained model has effectively captured the intricate interactions between biomolecular components and Mg^{2+} ions, allowing us to explore LLPS behaviour in realistic intranuclear physiological conditions. Our observations indicate that the presence of Mg^{2+} ions has a significant impact on the phase separation process, resulting in enhanced liquid-liquid demixing behaviour and the emergence of distinct phase-separated droplets in distinct intranuclear proteins that are known to phase separate in vivo but fail to do so in the absence of high amounts of crowding agents (*i.e.* PEG, dextran or Ficoll) or micromolar concentrations of magnesium salts. This is due to the enhanced interactions between negatively charged residues in the presence of Mg^{2+}ions, which neutralises the repulsions between the negative charges, hence turning the interaction attractive.

Our results demonstrate that Mg^{2+}ions can notably alter the extension and strength of interactions between amino acids, and very much so in the case of electrostatic interactions between charged pairs. Due to its small size and high charge, Mg^{2+}ions, especially when binding directly to a charged residue

The MagPi model was validated against experimental observations of the intrinsically disordered regions of proteins MED1 and BRD4, and the C-terminal IDR of Nanog, whose LLPS behaviour has been studied experimentally. Our model is able to recapitulate the phase behaviour of these IDPs. These do not form distinct droplets at room temperature, and form a well-mixed solution. These proteins contain a significant number of charged residues, many of which are negatively charged. Due to the ampholytic nature of these sequences, the electrostatic repulsions between samely charged residues are too ubiqutuous for associative phase separation to take place. Our Mg^{2+}–model can predict that these proteins undergo LLPS with minimal concentrations of Mg^{2+}ions due to mainly the reduced repulsions between acidic residues. Our model is also able to recapitulate the dependency of phase transition temperature with net charge, with greater changes in T_c as protein sequences have, in total, more negatively charged residues, as shown by the simulations of DDX4 and DDX3 variants.

Furthermore, our model was further put into application to test the effect of Mg^{2+}ions in phase separation of Polycomb proteins, namely CBX2 and RING1B, and stoichiometric mixtures of them in the paper "**Principles of assembly and regulation of condensates of Polycomb repressive complex 1 through phase separation**", which will be published in *Cell Reports*, but pre-print is available in Brown et al. [30]. Our model not only recapitulated experimental findings, such as the requirement of both proteins of micromolar concentrations of Mg^{2+}in order to form liquid droplets, but also recapitulated the reentrant behaviour of CBX2 and RING1B mixtures in the presence of Mg^{2+}ions, with its highest critical temperature at 1:1 stoichiometric ratio.

The implications of this work are multifaceted and contribute to a deeper understanding of biological LLPS in the context of polyvalent ions also present in MLO-forming systems. Firstly, our model provides valuable insights into the regulatory role of Mg^{2+} ions in cellular processes involving LLPS, shedding light on their potential influence on cellular organization and compartmentalization.

While our coarse-grained model demonstrates promising results, it is essential to acknowledge its limitations. One such limitation is the simplification of the molecular interactions within the system, which could overlook certain finer details crucial for accurately capturing all aspects of LLPS in biological environments, such as the lack of conformational information due to the absence of torsional and dihedral terms in the model's potential. Additionally, the parameterisation of the model is based on atomistic studies and bioinformatics data collected at room temperature. While our model can predict that LLPS will happen more favourably for intranuclear proteins in presence of Mg^{2+}ions, the furthest away from room temperature that these predictions are made, the higher the error will be. For this reason, further work is required to elucidate how interactions change with increasing temperatures.

Additionally, the absence of certain biological components or specific cellular factors in our model may impact the realism and complexity of the observed LLPS behaviour. Future work should aim to refine the model by including a more comprehensive representation of relevant biomolecules, like RNA or DNA. The parameterisation of these might be highly challenging, but will open up the door for many more studies, due to the experimentally

probed importance of Mg^{2+}ions in structural compaction and organisation of RNA and DNA molecules.

Furthermore, investigations into the molecular basis of Mg^{2+}-mediated effects on LLPS can provide a deeper understanding of the underlying mechanisms governing this phenomenon. Experimental validation and cross-validation with other computational models will be critical in confirming the robustness and reliability of our findings.

In conclusion, our coarse-grained model represents a significant step forward in studying LLPS in biological systems under the influence of Mg^{2+}ions. By shedding light on the role of Mg^{2+}in LLPS and providing insights into the formation and behaviour of phase-separated droplets, our work contributes to the broader understanding of the complex organization and regulation of biological processes. This research has the potential to impact various fields, including cellular biology, biophysics, and even drug development, where controlling LLPS could be of therapeutic interest. However, further experimental investigations are necessary to fully elucidate the intricacies of Mg^{2+}-mediated LLPS and bridge the gap between our computational model and real-world biological systems.

Chapter 8

Discussion and future work

In this thesis, we have developed multiscale coarse-grained models to investigate liquid-liquid phase separation phenomena of proteins, spanning systems with sizes from single molecule to hundreds of molecules. LLPS plays a pivotal role in the generation of membraneless organelles, which are crucial for various cellular functions [78]. The development and analysis of these models have provided valuable insights into the underlying mechanisms that govern LLPS, emphasizing the importance of multiscale modelling and the profound influence of residue-wise interactions on the bulk behaviour of droplet-forming proteins.

With our Mpipi and Mpipi Recharged multiscale residue-level models, we have been able to investigate the following:

- The importance of π-π and cation-π interactions as main driving forces of LLPS. Experimental studies and bioinformatics analysis [170, 174] had previously reported that aromatic and positively charged residues play a major role in the formation of biocondensates. The computational models that were available at the time of the development of Mpipi, which had been parameterised according to residue hydrophobicity, electrostatics, or to mimic experimental single-molecule properties, had not considered the correct balance of such interactions [45, 148, 164]. Through a detailed fine-tuning methodology including atomistic simulations and bioinformatics data, both Mpipi and Mpipi Recharged were able to recapitulate the phase behaviour of a wide set of proteins, such as FUS, hnRNPA1, *Laf1* and Ddx4, and several variants.

- The contribution of electrostatic interactions on LLPS, specifically on highly charged proteins. The PMF curves demonstrate that the "type" of interaction, be it cation-π, electrostatic, *etc*., is not sufficient to modulate the interaction affinities between amino acids. Indeed, the structural properties of amino acids determine the ordering of the interaction strengths. For instance, arginine residues participate in stronger interactions than lysines. The same can be said about glutamic *versus* aspartic acid. Particularly, the Mpipi Recharged model takes care of electrostatic contributions on a residue-by-residue basis, due to the flexibility of the Yukawa term that describes the charged long-range interactions.

- Since both models have been parameterised from all-atom calculations and represent residue–residue contact affinities also observed in microscopic and crystallographic data, Mpipi and Mpipi recharged are able to recapitulate the experimental radii of gyration of a set of IDPs with different net charges with high accuracy.

- The models in this thesis were set up at a residue-level resolution, which allowed us to study the effect of sequence mutations on the macromolecular phase behaviour of the proteins we tested, and reproduced experimental data with great accuracy. This is a non-trivial feature, that makes our models extremely powerful and useful to uncover molecular mechanisms explaining the phase behaviour of biomolecular systems. With more and more collaborative efforts in the field carrying out experimental and computational studies, one could argue that sequence specificity is one of the keys to bridging the gaps between theory and experiment. Furthermore, sequence-level resolution might be a powerful tool since it can aid in screening relevant proteins prior to experiments.

  Furthermore, this grants the models with high transferability to other protein systems. For instance, in an unprecendent study of 100+ variants across multiple phase-separating proteins, Maristany *et al*. [117] were able to elucidate mathematical expressions that associate a protein's sequence to its critical temperature using the Mpipi model.

Our investigation of the implication of Mg^{2+}ions in regulating amino acid interactions provided a strong base on top of which to build the Mg^{2+}–mediated LLPS model. This model, albeit still qualitative in nature due to scarce quantitative experimental data, is able to simulate phase transitions of intranuclear proteins in environments more similar to that inside of the cell nucleus.

We observed the dependency of LLPS on protein net charge that is captured by salt-screened electrostatic repulsion, even when assuming a uniform dielectric constant throughout the two-phase system.

This work demonstrates the effectiveness of "bottom-up" modelling, particularly all-atom simulations like PMF calculations, in fueling the parameterization process for simplified models of LLPS. The use of potential of mean force calculations to parameterize coarse-grained potentials enables the development of models that are both computationally efficient and physically meaningful, capturing the essential interactions and behaviours of the underlying system and enabling accurate and broadly applicable simulations.

These results do not rule out the influence of other factors in inducing LLPS, as could be pH, conformational changes at the molecular level, *etc*. However, introducing this in a simplified model is not a trivial task and would be resource intensive.

Biological LLPS is heavily influenced by the patterning of residues, especially within intrinsically disordered regions of proteins. The specific sequence and arrangement of these residues can modulate the phase behaviour of proteins [107]. For instance, a combination of charged and aromatic residues can promote phase separation under certain conditions while determining the strength and specificity of protein-protein interactions at the same time [119, 124]. Moreover, droplet patterning aids in selectively recruiting molecules for functional specificity in membraneless organelles.

One current limitation of our coarse-grained models is the loss of conformational dynamics data due to the exclusion of torsional and dihedral terms in the force field. Although *a priori* this means our model is faster and more efficient, it makes it impossible to recover information on the structural transitions that proteins might undergo inside condensates. The development of a coarse-grained model for LLPS that possesses backmapping abilities would

be a significant advancement in the field. Backmapping refers to the ability to revert the simplified CG representation back to its original atomistic detail. This capability would allow researchers to transition seamlessly between the macroscopic phase behaviour observed in the CG model and the detailed molecular interactions at the atomistic level.

By emphasizing single-molecule conformations in a CG model with backmapping abilities, researchers can gain insights into how individual molecular states contribute to the collective behavior observed in LLPS, such as coil-to-globule transitions [189], IDR expansion [64], as well as β-sheet formation, which is of significant importance in pathological LLPS. This would provide a more comprehensive understanding of the factors that regulate phase separation at both the individual molecule and system-wide levels.

On the other hand, in biological systems, where electrostatic interactions are crucial, the adaptability of polarisable force fields, which allow atomic charges to adapt to their local environment, is particularly important. Traditional force fields with fixed charges are inadequate in capturing the nuanced interplay of forces in complex biochemical environments. Polarisable force fields can accurately model electrostatic interactions by adapting to changing conditions. This adaptability is especially essential when studying LLPS, where subtle shifts in molecular interactions can lead to significant changes in phase behavior. Polarisable force fields are ideally suited to model the inherent heterogeneity of biological LLPS, with their ability to adapt to different molecular environments and account for the unique electrostatic signature of each molecule. The addition of polarisability into force fields is not an easy task and has not caught up yet in the field of LLPS, but we strongly believe that it would be a necessary step towards obtaining representations of the phase behaviour of biomolecules in the closest conditions possible as in *in vivo* experiments.

Several biologically relevant and multiphase condensates have emerged recently, and these condensates have been suggested to have distinct architectures. Therefore, computer models capable of reproducing the ubiquitous properties of these condensates, which result from the diversity of biomolecular components, are highly desirable. By including these effects in computational models, we can better understand the driving forces behind intracellular LLPS and the resulting material properties. However, none of the current models

presented in this work is capable of taking this kind of phase behaviour into account. An approach that could be used to enable successful simulations of a broader range of proteins would be similar to that in Ref. 46 for the HPS-KR model, which includes a clear temperature dependence of the interaction forces.

Turning now to future endeavours, one of the primary areas of focus in our projected future work is the development of a coarse-grained force field tailored specifically for proteins and RNA that undergo LLPS. Even though the Mpipi model originally included a parameterisation for protein–RNA interactions [83] based on combination rules from protein–protein RNA–RNA interactions, further work not included in this thesis points that combination rules fail to capture these heterotypic interactions correctly. Parameters for protein–RNA hybrid systems are being developed at the moment using better tailored approaches.

In tandem with the development of the coarse-grained force field, we aim to investigate the role of magnesium ions in mediating interactions within biomolecular complexes. Magnesium ions are known to play a crucial role in stabilizing various biomolecular structures and facilitating interactions. Our study will encompass protein-RNA, RNA-RNA, and DNA interactions, providing a comprehensive overview of how magnesium ions influence the stability, conformation, and dynamics of these systems. Given the ubiquity of magnesium ions in biological processes, understanding their role at a molecular level can shed light on various cellular mechanisms, from transcription and translation to DNA replication and repair.

In conclusion, the work presented in this thesis and our future work aims to bridge the gap between atomistic and molecular structures and macroscopic behavior in biomolecular systems. By developing a specialized coarse-grained force field for proteins and nucleic acids undergoing LLPS and studying the role of various salts in biomolecular interactions, we hope to provide a holistic understanding of the forces that govern the formation of MLOs and cellular processes. This knowledge, in turn, has the potential to drive innovations in both basic science and therapeutic applications.

References

[1] (2005). The Steepest Descent Method. In *Nonlinear Optimization with Financial Applications*, pages 51–64. Springer US, Boston, MA.

[2] Adame-Arana, O., Weber, C. A., Zaburdaev, V., Prost, J., and Jülicher, F. (2020). Liquid phase separation controlled by ph. *Biophysical Journal*, 119(8):1590–1605.

[3] Alberti, S., Gladfelter, A., and Mittag, T. (2019). Considerations and challenges in studying liquid-liquid phase separation and biomolecular condensates.

[4] Alexander, P. A., He, Y., Chen, Y., Orban, J., and Bryan, P. N. (2007). The design and characterization of two proteins with 88identity but different structure and function. *Proceedings of the National Academy of Sciences*, 104(29):11963–11968.

[5] Allen, M. P. and Tildesley, D. J. (2017). *Computer simulation of liquids*. Oxford university press.

[6] Allnér, O., Nilsson, L., and Villa, A. (2012). Magnesium ion–water coordination and exchange in biomolecular simulations. *Journal of Chemical Theory and Computation*, 8(4):1493–1502.

[7] Ambadipudi, S., Biernat, J., Riedel, D., Mandelkow, E., and Zweckstetter, M. (2017). Liquid–liquid phase separation of the microtubule-binding repeats of the alzheimer-related protein tau. *Nature Communications*, 8(1).

[8] Andrew, C. D., Bhattacharjee, S., Kokkoni, N., Hirst, J. D., Jones, G. R., and Doig, A. J. (2002). Stabilizing Interactions between Aromatic and Basic Side Chains in {α}-Helical Peptides and Proteins. Tyrosine Effects on Helix Circular Dichroism. *Journal of the American Chemical Society*, 124(43):12706–12714.

[9] Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, 181(4096):223–230.

[10] Araki, K., Yagi, N., Nakatani, R., Sekiguchi, H., So, M., Yagi, H., Ohta, N., Nagai, Y., Goto, Y., and Mochizuki, H. (2016). A small-angle X-ray scattering study of alpha-synuclein from human red blood cells. *Scientific Reports*, 6(1):30473.

[11] Arbesú, M., Maffei, M., Cordeiro, T. N., Teixeira, J. M. C., Pérez, Y., Bernadó, P., Roche, S., and Pons, M. (2017). The Unique Domain Forms a Fuzzy Intramolecular Complex in Src Family Kinases. *Structure*, 25(4):630–640.

[12] Babinchak, W. M., Haider, R., Dumm, B. K., Sarkar, P., Surewicz, K., Choi, J.-K., and Surewicz, W. K. (2019). The role of liquid–liquid phase separation in aggregation of the tdp-43 low-complexity domain. *Journal of Biological Chemistry*, 294(16):6306–6317.

[13] Bairoch, A. and Boeckmann, B. (1992). The swiss-prot protein sequence data bank. *Nucleic Acids Research*, 20(suppl):2019–2022.

[14] Banani, S. F., Lee, H. O., Hyman, A. A., and Rosen, M. K. (2017). Biomolecular condensates: Organizers of cellular biochemistry.

[15] Banani, S. F., Rice, A. M., Peeples, W. B., Lin, Y., Jain, S., Parker, R., and Rosen, M. K. (2016). Compositional Control of Phase-Separated Cellular Bodies. *Cell*, 166(3):651–663.

[16] Banci, L. (2013). *Metallomics and the cell*. Springer.

[17] Baul, U., Chakraborty, D., Mugnai, M. L., Straub, J. E., and Thirumalai, D. (2019). Sequence Effects on Size, Shape, and Structural Heterogeneity in Intrinsically Disordered Proteins. *The Journal of Physical Chemistry B*, 123(16):3462–3474.

[18] Bayly, C. I., Cieplak, P., Cornell, W., and Kollman, P. A. (1993). A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *The Journal of Physical Chemistry*, 97(40):10269–10280.

[19] Benavides, A. L., Aragones, J. L., and Vega, C. (2016). Consensus on the solubility of NaCl in water from computer simulations using the chemical potential route. *The Journal of Chemical Physics*, 144(12):124504.

[20] Berendsen, H., van der Spoel, D., and van Drunen, R. (1995). Gromacs: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications*, 91(1-3):43–56.

[21] Berry, J., Weber, S. C., Vaidya, N., Haataja, M., and Brangwynne, C. P. (2015). Rna transcription modulates phase transition-driven nuclear body assembly. *Proceedings of the National Academy of Sciences*, 112(38):E5237–E5245.

[22] Best, R. B., Zheng, W., and Mittal, J. (2014). Balanced Protein–Water Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association. *Journal of Chemical Theory and Computation*, 10(11):5113–5124.

[23] Black, C., Huang, H.-W., and Cowan, J. (1994). Biological coordination chemistry of magnesium, sodium, and potassium ions. protein and nucleotide binding sites. *Coordination Chemistry Reviews*, 135-136:165–202.

[24] Boehning, M., Dugast-Darzacq, C., Rankovic, M., Hansen, A. S., Yu, T., Marie-Nelly, H., McSwiggen, D. T., Kokic, G., Dailey, G. M., Cramer, P., Darzacq, X., and Zweckstetter, M. (2018). RNA polymerase II clustering through carboxy-terminal domain phase separation. *Nature Structural and Molecular Biology*, 25(9):833–840.

[25] Boire, A., Sanchez, C., Morel, M.-H., Lettinga, M. P., and Menut, P. (2018). Dynamics of liquid-liquid phase separation of wheat gliadins. *Scientific Reports*, 8(1).

[26] 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., and Kay, L. E. (2017). Structural and hydrodynamic properties of an intrinsically disordered region of a germ cell-specific protein on phase separation. *Proceedings of the National Academy of Sciences of the United States of America*, 114:E8194–E8203.

[27] Brangwynne, C. P., Eckmann, C. R., Courson, D. S., Rybarska, A., Hoege, C., Gharakhani, J., Julicher, F., and Hyman, A. A. (2009). Germline P Granules Are Liquid Droplets That Localize by Controlled Dissolution/Condensation. *Science*, 324(5935):1729–1732.

[28] Brangwynne, C. P., Mitchison, T. J., and Hyman, A. A. (2011). Active liquid-like behavior of nucleoli determines their size and shape in Xenopus laevis oocytes. *Proceedings of the National Academy of Sciences of the United States of America*, 108(11):4334–4339.

[29] Bremer, A., Farag, M., Borcherds, W. M., Peran, I., Martin, E. W., Pappu, R. V., and Mittag, T. (2021). Deciphering how naturally occurring sequence features impact the phase behaviors of disordered prion-like domains. *bioRxiv*, page 2021.01.01.425046.

[30] Brown, K., Chew, P. Y., Ingersoll, S., Espinosa, J. R., Aguirre, A., Kutateladze, T., Guevara, R. C., and Ren, X. (2022). *Principles of Assembly and regulation of condensates of Polycomb repressive complex 1 through phase separation*.

[31] Burke, K. A., Janke, A. M., Rhine, C. L., and Fawzi, N. L. (2015). Residue-by-residue view of in vitro fus granules that bind the c-terminal domain of rna polymerase ii. *Molecular Cell*, 60(2):231–241.

[32] Bussi, G. and Laio, A. (2020). Using metadynamics to explore complex free-energy landscapes. *Nature Reviews Physics*, 2(4):200–212.

[33] Cai, S., Zhang, C., Zhuang, Z., Zhang, S., Ma, L., Yang, S., Zhou, T., Wang, Z., Xie, W., Jin, S., and et al. (2023). Phase-separated nucleocapsid protein of sars-cov-2 suppresses cgas-dna recognition by disrupting cgas-g3bp1 complex. *Signal Transduction and Targeted Therapy*, 8(1).

[34] Chen, L. and Feany, M. B. (2005). -synuclein phosphorylation controls neurotoxicity and inclusion formation in a drosophila model of parkinson disease. *Nature Neuroscience*, 8(5):657–663.

[35] Choi, J. M., Dar, F., and Pappu, R. V. (2019). Lassi: A lattice model for simulating phase transitions of multivalent proteins. *PLoS Computational Biology*, 15.

[36] Choi, J.-M., Holehouse, A. S., and Pappu, R. V. (2020). Physical Principles Underlying the Complex Biology of Intracellular Phase Transitions. *Annual Review of Biophysics*, 49(1).

[37] Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. (2009). Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423.

[38] Conicella, A. E., Dignon, G. L., Zerze, G. H., Schmidt, H. B., D'Ordine, A. M., Kim, Y. C., Rohatgi, R., Ayala, Y. M., Mittal, J., and Fawzi, N. L. (2020). Tdp-43 -helical structure tunes liquid–liquid phase separation and function. *Proceedings of the National Academy of Sciences*, 117:5883–5894.

[39] Crabtree, M. D., Holland, J., Kompella, P., Babl, L., Turner, N., Baldwin, A. J., and Nott, T. J. (2020). Repulsive electrostatic interactions modulate dense and dilute phase properties of biomolecular condensates. *bioRxiv*.

[40] Dannenhoffer-Lafage, T. and Best, R. B. (2021). A Data-Driven Hydrophobicity Scale for Predicting Liquid–Liquid Phase Separation of Proteins. *The Journal of Physical Chemistry B*, 125(16):4046–4056.

[41] Darden, T., York, D., and Pedersen, L. (1993). Particle mesh ewald: Annlog(n) method for ewald sums in large systems. *The Journal of Chemical Physics*, 98(12):10089–10092.

[42] Das, S., Lin, Y.-H., Vernon, R. M., Forman-Kay, J. D., and Chan, H. S. (2020). Comparative Roles of Charge, {π}, and Hydrophobic Interactions in Sequence-Dependent Phase Separation of Intrinsically Disordered Proteins. *Proceedings of the National Academy of Sciences*.

[43] de Baaij, J. H., Hoenderop, J. G., and Bindels, R. J. (2015). Magnesium in man: Implications for health and disease. *Physiological Reviews*, 95(1):1–46.

[44] Debye, P. and Hückel, E. (1923). Zur Theorie der Elektrolyte. I. Gefrierpunktserniedrigung und verwandte Erscheinungen. *Phys. Z.*, 24:185–206.

[45] Dignon, G. L., Zheng, W., Kim, Y. C., Best, R. B., and Mittal, J. (2018). Sequence determinants of protein phase behavior from a coarse-grained model. *PLOS Computational Biology*, 14:e1005941.

[46] Dignon, G. L., Zheng, W., Kim, Y. C., and Mittal, J. (2019). Temperature-Controlled Liquid–Liquid Phase Separation of Disordered Proteins. *ACS Central Science*, 5(5):821–830.

[47] Ditlev, J. A., Case, L. B., and Rosen, M. K. (2018). Who's in and who's out—compositional control of biomolecular condensates. *Journal of molecular biology*.

[48] Doi, M. (2017). *Soft matter physics*. Oxford University Press.

[49] Dubreuil, B., Matalon, O., and Levy, E. D. (2019). Protein Abundance Biases the Amino Acid Composition of Disordered Regions to Minimize Non-functional Interactions. *Journal of Molecular Biology*, 431(24):4978–4992.

[50] Dyson, H. J., Wright, P. E., and Scheraga, H. A. (2006). The role of hydrophobic interactions in initiation and propagation of protein folding. *Proceedings of the National Academy of Sciences*, 103(35):13057 LP – 13061.

[51] Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1 – 26.

[52] Engelhardt, M. (2004). Condensation of chromatin in situ by cation-dependent charge shielding and aggregation. *Biochemical and Biophysical Research Communications*, 324(4):1210–1214.

[53] Espinosa, J. R., Garaizar, A., Vega, C., Frenkel, D., and Collepardo-Guevara, R. (2019). Breakdown of the law of rectilinear diameter and related surprises in the liquid-vapor coexistence in systems of patchy particles. *Journal of Chemical Physics*, 150.

[54] Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., and Pedersen, L. G. (1995). A smooth particle mesh Ewald method. *The Journal of Chemical Physics*, 103(19):8577–8593.

[55] Farr, S. E., Woods, E. J., Joseph, J. A., Garaizar, A., and Collepardo-Guevara, R. (2020). Nucleosome plasticity is a critical element of chromatin liquid–liquid phase separation and multivalent nucleosome interactions. *bioRxiv*, page 2020.11.23.391599.

[56] Feng, Z., Chen, X., Wu, X., and Zhang, M. (2019). Formation of biological condensates via phase separation: Characteristics, analytical methods, and physiological implications. *Journal of Biological Chemistry*, 294(40):14823–14835.

[57] Feric, M., Vaidya, N., Harmon, T. S., Mitrea, D. M., Zhu, L., Richardson, T. M., Kriwacki, R. W., Pappu, R. V., and Brangwynne, C. P. (2016). Coexisting liquid phases underlie nucleolar subcompartments. *Cell*, 165(7):1686 – 1697.

[58] Fisher, R. S. and Elbaum-Garfinkle, S. (2020). Tunable multiphase dynamics of arginine and lysine liquid condensates. *Nature communications*, 11(1):4628.

[59] Fossat, M. J., Zeng, X., and Pappu, R. V. (2021). Uncovering Differences in Hydration Free Energies and Structures for Model Compound Mimics of Charged Side Chains of Amino Acids. *The Journal of Physical Chemistry B*, 125(16):4148–4161.

[60] Frenkel, D. and Smit, B. (2002). *Understanding Molecular Simulation: From Algorithms to Applications*, volume 1 of *Computational Science Series*. Academic Press, San Diego, second edition.

[61] Fuertes, G., Banterle, N., Ruff, K. M., Chowdhury, A., Mercadante, D., Koehler, C., Kachala, M., Estrada Girona, G., Milles, S., Mishra, A., Onck, P. R., Gräter, F., Esteban-Martín, S., Pappu, R. V., Svergun, D. I., and Lemke, E. A. (2017). Decoupling of size and shape fluctuations in heteropolymeric sequences reconciles discrepancies in SAXS vs. FRET measurements. *Proceedings of the National Academy of Sciences*, 114(31):E6342 LP – E6351.

[62] Gall, J. G. (2003). The centennial of the cajal body. *Nature Reviews Molecular Cell Biology*, 4(12):975–980.

[63] Gallivan, J. P. and Dougherty, D. A. (1999). Cation-interactions in structural biology.

[64] Garaizar, A., Sanchez-Burgos, I., Collepardo-Guevara, R., and Espinosa, J. R. (2020). Expansion of Intrinsically Disordered Proteins Increases the Range of Stability of Liquid–Liquid Phase Separation. *Molecules*, 25(20):4705.

[65] Gibbs, J. W., Bumstead, H. A., and Gibbs, V. N. R. (1906). *The scientific papers of J. Willard Gibbs*. Longmans, Green and Co.

[66] Gibson, B. A., Doolittle, L. K., Schneider, M. W., Jensen, L. E., Gamarra, N., Henry, L., Gerlich, D. W., Redding, S., and Rosen, M. K. (2019). Organization of chromatin by intrinsic and regulated phase separation. *Cell*, 179(2).

[67] Gilson, M. K. and Zhou, H.-X. (2007). Calculation of protein-ligand binding affinities. *Annual Review of Biophysics and Biomolecular Structure*, 36(1):21–42.

[68] Gomes, G.-N. W., Krzeminski, M., Namini, A., Martin, E. W., Mittag, T., Head-Gordon, T., Forman-Kay, J. D., and Gradinaru, C. C. (2020). Conformational Ensembles of an Intrinsically Disordered Protein Consistent with NMR, SAXS, and Single-Molecule FRET. *Journal of the American Chemical Society*, 142(37):15697–15710.

[69] Grotz, K. K., Cruz-León, S., and Schwierz, N. (2021). Optimized magnesium force field parameters for biomolecular simulations with accurate solvation, ion-binding, and water-exchange properties. *Journal of Chemical Theory and Computation*, 17(4):2530–2540.

[70] Grotz, K. K. and Schwierz, N. (2021). Optimized magnesium force field parameters for biomolecular simulations with accurate solvation, ion-binding, and water-exchange properties in spc/e, tip3p-fb, tip4p/2005, tip4p-ew, and tip4p-d. *Journal of Chemical Theory and Computation*, 18(1):526–537.

[71] Hamelryck, T., Borg, M., Paluszewski, M., Paulsen, J., Frellsen, J., Andreetta, C., Boomsma, W., Bottaro, S., and Ferkinghoff-Borg, J. (2010). Potentials of mean force for protein structure prediction vindicated, formalized and generalized. *PLoS ONE*, 5(11).

[72] Han, T. W., Kato, M., Xie, S., Wu, L. C., Mirzaei, H., Pei, J., Chen, M., Xie, Y., Allen, J., Xiao, G., and McKnight, S. L. (2012). Cell-free formation of RNA granules: Bound RNAs identify features and components of cellular assemblies. *Cell*, 149(4):768–779.

[73] Handwerger, K. E., Cordero, J. A., and Gall, J. G. (2005). Cajal bodies, nucleoli, and speckles in the xenopus oocyte nucleus have a low-density, sponge-like structure. *Molecular Biology of the Cell*, 16(1):202–211.

[74] Harmon, T. S., Holehouse, A. S., Rosen, M. K., and Pappu, R. V. (2017). Intrinsically disordered linkers determine the interplay between phase separation and gelation in multivalent proteins. *eLife*, 6.

[75] He, Y., Rozak, D. A., Sari, N., Chen, Y., Bryan, P., and Orban, J. (2006). Structure, dynamics, and stability variation in bacterial albumin binding modules: implications for species specificity. *Biochemistry*, 45(33):10102–10109.

[76] Hess, B., Bekker, H., Berendsen, H. J., and Fraaije, J. G. (1997). Lincs: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry*, 18(12):1463–1472.

[77] Holcomb, C. D., Clancy, P., and Zollweg, J. A. (1993). A critical study of the simulation of the liquid-vapour interface of a lennard-jones fluid. *Molecular Physics*, 78(2):437–459.

[78] Hyman, A. A. and Simons, K. (2012). Beyond Oil and Water-Phase Transitions in Cells. *Science*, 337(6098):1047–1049.

[79] Inoue, S., Sugiyama, S., Travers, A. A., and Ohyama, T. (2006). Self-assembly of double-stranded dna molecules at nanomolar concentrations. *Biochemistry*, 46(1):164–171.

[80] Iserman, C., Desroches Altamirano, C., Jegers, C., Friedrich, U., Zarin, T., Fritsch, A. W., Mittasch, M., Domingues, A., Hersemann, L., Jahnel, M., and et al. (2020). Condensation of ded1p promotes a translational switch from housekeeping to stress protein production. *Cell*, 181(4).

[81] Ishov, A. M., Sotnikov, A. G., Negorev, D., Vladimirova, O. V., Neff, N., Kamitani, T., Yeh, E. T., Strauss, J. F., and Maul, G. G. (1999). Pml is critical for nd10 formation and recruits the pml-interacting protein daxx to this nuclear structure when modified by sumo-1. *Journal of Cell Biology*, 147(2):221–234.

[82] Jiang, W., Phillips, J. C., Huang, L., Fajer, M., Meng, Y., Gumbart, J. C., Luo, Y., Schulten, K., and Roux, B. (2014). Generalized scalable multiple copy algorithms for molecular dynamics simulations in {NAMD}. *Comput. Phys. Commun.*, 185(3):908–916.

[83] Joseph, J. A., Espinosa, J. R., Sanchez-Burgos, I., Garaizar, A., Frenkel, D., and Collepardo-Guevara, R. (2021). Thermodynamics and kinetics of phase separation of protein-RNA mixtures by a minimal model. *Biophysical Journal*.

[84] Kang, B., Tang, H., Zhao, Z., and Song, S. (2020). Hofmeister series: Insights of ion specificity from amphiphilic assembly and interface property. *ACS Omega*, 5(12):6229–6239.

[85] Kapcha, L. H. and Rossky, P. J. (2014). A simple atomic-level hydrophobicity scale reveals protein interfacial structure. *Journal of Molecular Biology*, 426(2):484–498.

[86] Kästner, J. (2011). Umbrella sampling. *WIREs Computational Molecular Science*, 1(6):932–942.

[87] Kato, M., Han, T. W., Xie, S., Shi, K., Du, X., Wu, L. C., Mirzaei, H., Goldsmith, E. J., Longgood, J., Pei, J., Grishin, N. V., Frantz, D. E., Schneider, J. W., Chen, S., Li, L., Sawaya, M. R., Eisenberg, D., Tycko, R., and McKnight, S. L. (2012). Cell-free formation of RNA granules: Low complexity sequence domains form dynamic fibers within hydrogels. *Cell*, 149(4):753–767.

[88] Kawahata, I., Finkelstein, D. I., and Fukunaga, K. (2022). Pathogenic impact of -synuclein phosphorylation and its kinases in -synucleinopathies. *International Journal of Molecular Sciences*, 23(11):6216.

[89] Kim, Y. C. and Hummer, G. (2008). Coarse-grained Models for Simulations of Multiprotein Complexes: Application to Ubiquitin Binding. *Journal of Molecular Biology*, 375(5):1416–1433.

[90] Kirkwood, J. G. (1935). Statistical mechanics of fluid mixtures. *The Journal of Chemical Physics*, 3(5):300–313.

[91] Kjaergaard, M., Nørholm, A., Hendus–Altenburger, R., Pedersen, S. F., Poulsen, F. M., and Kragelund, B. B. (2010). Temperature-dependent structural changes in intrinsically disordered proteins: Formation of α–helices or loss of polyproline II? *Protein Science*, 19(8):1555–1564.

[92] Klein, I., Boija, A., Afeyan, L., Hawken, S., Fan, M., Taatjes, D., Chakraborty, A., Sharp, P., Chang, Y. T., Hyman, A., and et al. (2021). Partitioning of cancer therapeutics in nuclear condensates. *Journal of Clinical Oncology*, 39(15$_s$*uppl*) : 3131˘3131.

[93] Koningsveld, R., Stockmayer, W. H., and Nies, E. (2008). *Polymer phase diagrams: A textbook*. Oxford Univ. Press.

[94] Krainer, G., Welsh, T. J., Joseph, J. A., St George-Hyslop, P., Hyman, A. A., Collepardo-Guevara, R., Alberti, S., and Knowles, T. P. (2021). Reentrant liquid condensate phase of proteins is stabilized by hydrophobic and non-ionic interactions. *Biophysical Journal*, 120(3).

[95] Kumar, K., Woo, S. M., Siu, T., Cortopassi, W. A., Duarte, F., and Paton, R. S. (2018). Cation– interactions in protein–ligand binding: Theory and data-mining reveal different roles for lysine and arginine. *Chemical Science*, 9(10):2655–2665.

[96] Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H., and Kollman, P. A. (1992). The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of computational chemistry*, 13(8):1011–1021.

[97] Laflamme, G. and Mekhail, K. (2020). Biomolecular condensates as arbiters of biochemical reactions inside the nucleus. *Communications Biology*, 3(1).

[98] Laghmach, R., Malhotra, I., and Potoyan, D. A. (2022). Multiscale modeling of protein-rna condensation in and out of equilibrium. *Methods in Molecular Biology*, page 117–133.

[99] Langevin, P. (1908). Sur la théorie du mouvement brownien. *CR Acad. Sci.*, 146(530-533).

[100] Latham, J. A. and Cech, T. R. (1989). Defining the inside and outside of a catalytic rna molecule. *Science*, 245(4915):276–282.

[101] Le Ferrand, H., Duchamp, M., Gabryelczyk, B., Cai, H., and Miserez, A. (2019). Time-resolved observations of liquid–liquid phase separation at the nanoscale using in situ liquid transmission electron microscopy. *Journal of the American Chemical Society*, 141(17):7202–7210.

[102] Li, H., Tang, C., and Wingreen, N. S. (1997). Nature of driving force for protein folding: A result from analyzing the statistical potential. *Physical Review Letters*, 79(4):765–768.

[103] 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., and Rosen, M. K. (2012). Phase transitions in the assembly of multivalent signalling proteins. *Nature*, 483:336–340.

[104] Li, P., Song, L. F., and Merz, K. M. (2014). Parameterization of highly charged metal ions using the 12-6-4 lj-type nonbonded model in explicit water. *The Journal of Physical Chemistry B*, 119(3):883–895.

[105] Li, S., Yoshizawa, T., Yamazaki, R., Fujiwara, A., Kameda, T., and Kitahara, R. (2021). Pressure and temperature phase diagram for liquid–liquid phase separation of the rna-binding protein fused in sarcoma. *The Journal of Physical Chemistry B*, 125(25):6821–6829.

[106] Lichtinger, S. M., Garaizar, A., Collepardo-Guevara, R., and Reinhardt, A. (2020). Targeted modulation of protein liquid-liquid phase separation by evolution of amino-acid sequence. *bioRxiv*.

[107] Lin, Y. H., Brady, J. P., Chan, H. S., and Ghosh, K. (2020). A unified analytical theory of heteropolymers for sequence-specific phase behaviors of polyelectrolytes and polyampholytes. *Journal of Chemical Physics*, 152(4):45102.

[108] Lindahl, T., Adams, A., and Fresco, J. R. (1966). Renaturation of transfer ribonucleic acids through site binding of magnesium. *Proceedings of the National Academy of Sciences*, 55(4):941–948.

[109] Lindt, J. v., Bratek-Skicki, A., Pakravan, D., Den Bosch, L. V., Maes, D., and Tompa, P. (2019). *A generic approach for studying the kinetics of liquid-liquid phase separation under near-native conditions*.

[110] Linsenmeier, M., Hondele, M., Grigolato, F., Secchi, E., Weis, K., and Arosio, P. (2022). Dynamic arrest and aging of biomolecular condensates are modulated by low-complexity domains, rna and biochemical activity. *Nature Communications*, 13(1).

[111] Liu, H., Fu, H., Shao, X., Cai, W., and Chipot, C. (2020). Accurate description of cation interactions in proteins with a nonpolarizable force field at no additional cost. *Journal of Chemical Theory and Computation*, 16(10):6397–6407.

[112] Liu, W., Samanta, A., Deng, J., Akintayo, C. O., and Walther, A. (2022). Mechanistic insights into the phase separation behavior and pathway-directed information exchange in all-dna droplets. *Angewandte Chemie*, 134(45).

[113] Luo, H., Lee, N., Wang, X., Li, Y., Schmelzer, A., Hunter, A. K., Pabst, T., and Wang, W. K. (2017). Liquid-liquid phase separation causes high turbidity and pressure during low ph elution process in protein a chromatography. *Journal of Chromatography A*, 1488:57–67.

[114] Maeshima, K., Matsuda, T., Shindo, Y., Imamura, H., Tamura, S., Imai, R., Kawakami, S., Nagashima, R., Soga, T., Noji, H., and et al. (2018). A transient rise in free mg2+ ions released from atp-mg hydrolysis contributes to mitotic chromosome condensation. *Current Biology*, 28(3).

[115] Mamatkulov, S. and Schwierz, N. (2018). Force fields for monovalent and divalent metal cations in tip3p water based on thermodynamic and kinetic properties. *The Journal of Chemical Physics*, 148(7).

[116] Mao, Y. S., Zhang, B., and Spector, D. L. (2011). Biogenesis and function of nuclear bodies. *Trends in Genetics*, 27(8):295–306.

[117] Maristany, M. J., Gonzalez, A. A., Collepardo-Guevara, R., and Joseph, J. A. (2023). *Universal predictive scaling laws of phase separation of prion-like low complexity domains*.

[118] Martin, E. W., Holehouse, A. S., Grace, C. R., Hughes, A., Pappu, R. V., and Mittag, T. (2016). Sequence Determinants of the Conformational Properties of an Intrinsically Disordered Protein Prior to and upon Multisite Phosphorylation. *Journal of the American Chemical Society*, 138(47):15323–15335.

[119] Martin, E. W., Holehouse, A. S., Peran, I., Farag, M., Incicco, J. J., Bremer, A., Grace, C. R., Soranno, A., Pappu, R. V., and Mittag, T. (2020). Valence and patterning of aromatic residues determine the phase behavior of prion-like domains downloaded from.

[120] Meng, Y., Sabri Dashti, D., and Roitberg, A. E. (2011). Computing alchemical free energy differences with hamiltonian replica exchange molecular dynamics (h-remd) simulations. *Journal of Chemical Theory and Computation*, 7(9):2721–2727.

[121] Milovanovic, D., Wu, Y., Bian, X., and De Camilli, P. (2018). A liquid phase of synapsin and lipid vesicles. *Science*, 361(6402):604–607.

[122] 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., and Kriwacki, R. W. (2018). Self-interaction of npm1 modulates multiple mechanisms of liquid–liquid phase separation. *Nature Communications*, 9(1).

[123] Mittag, T., Kay, L. E., and Forman-Kay, J. D. (2010). Protein dynamics and conformational disorder in molecular recognition. *Journal of molecular recognition*, 23(2):105–116.

[124] Miyagi, T., Yamazaki, R., Ueda, K., Narumi, S., Hayamizu, Y., Uji-i, H., Kuroda, M., and Kanekura, K. (2022). The patterning and proportion of charged residues in the arginine-rich mixed-charge domain determine the membrane-less organelle targeted by the protein. *International Journal of Molecular Sciences*, 23(14):7658.

[125] Miyazawa, S. and Jernigan, R. L. (1996). Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *Journal of Molecular Biology*, 256(3):623–644.

[126] Molliex, A., Temirov, J., Lee, J., Coughlin, M., Kanagaraj, A. P., Kim, H. J., Mittag, T., and Taylor, J. P. (2015). Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. *Cell*, 163(1):123–133.

[127] Moomaw, A. S. and Maguire, M. E. (2008). The unique nature of mg2+channels. *Physiology*, 23(5):275–285.

[128] Moser, J. J. and Fritzler, M. J. (2010). Cytoplasmic ribonucleoprotein (rnp) bodies and their relationship to gw/p bodies. *The International Journal of Biochemistry & Cell Biology*, 42(6):828 – 843.

[129] Mueller, F., Mazza, D., Stasevich, T. J., and McNally, J. G. (2010). Frap and kinetic modeling in the analysis of nuclear protein dynamics: What do we really know? *Current Opinion in Cell Biology*, 22(3):403–411.

[130] Murthy, A. C., Dignon, G. L., Kan, Y., Zerze, G. H., Parekh, S. H., Mittal, J., and Fawzi, N. L. (2019). Molecular interactions underlying liquid liquid phase separation of the FUS low complexity domain. *Nature Structural and Molecular Biology*, 26(7):637–648.

[131] Mylonas, E., Hascher, A., Bernadó, P., Blackledge, M., Mandelkow, E., and Svergun, D. I. (2008). Domain Conformation of Tau Protein Studied by Solution Small-Angle X-ray Scattering. *Biochemistry*, 47(39):10345–10353.

[132] Nara, M., Morii, H., and Tanokura, M. (2013). Coordination to divalent cations by calcium-binding proteins studied by ftir spectroscopy. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1828(10):2319–2327.

[133] Nedelsky, N. B. and Taylor, J. P. (2019). Bridging biophysics and neurology: Aberrant phase transitions in neurodegenerative disease. *Nature Reviews Neurology*, 15(5):272–286.

[134] Nilsson, D. and Irbäck, A. (2021). Finite-size shifts in simulated protein droplet phase diagrams. *The Journal of Chemical Physics*, 154(23):235101.

[135] Nishikawa, J.-i. and Ohyama, T. (2012). Selective association between nucleosomes with identical dna sequences. *Nucleic Acids Research*, 41(3):1544–1554.

[136] Nott, T. J., Craggs, T. D., and Baldwin, A. J. (2016). Membraneless organelles can melt nucleic acid duplexes and act as biomolecular filters. *Nature Chemistry*, 8(6):569–575.

[137] Nott, T. J. J. J., Petsalaki, E., Farber, P., Jervis, D., Fussner, E., Plochowietz, A., Craggs, T. D., Bazett-Jones, D. P. P. P., Pawson, T., Forman-Kay, J. D. D. D., and Baldwin, A. J. J. J. (2015). Phase Transition of a Disordered Nuage Protein Generates Environmentally Responsive Membraneless Organelles. *Molecular Cell*, 57(5):936–947.

[138] Nye, M. J. (1972). *Molecular reality: A perspective on the scientific work of Jean Perrin*. Macdonald.

[139] Pak, C. W. W., Kosno, M., Holehouse, A. S. S., Padrick, S. B. B., Mittal, A., Ali, R., Yunus, A. A. A., Liu, D. R. R., Pappu, R. V. V., and Rosen, M. K. K. (2016). Sequence Determinants of Intracellular Phase Separation by Complex Coacervation of a Disordered Protein. *Molecular Cell*, 63(1):72–85.

[140] Paloni, M., Bailly, R., Ciandrini, L., and Barducci, A. (2020). Unraveling molecular interactions in liquid–liquid phase separation of disordered proteins by atomistic simulations. *The Journal of Physical Chemistry B*, 124(41):9009–9016.

[141] Pan, J., Thirumalai, D., and Woodson, S. A. (1999). Magnesium-dependent folding of self-splicing rna: Exploring the link between cooperativity, thermodynamics, and kinetics. *Proceedings of the National Academy of Sciences*, 96(11):6149–6154.

[142] Parmar, A. S. and Muschol, M. (2009). Hydration and hydrodynamic interactions of lysozyme: Effects of chaotropic versus kosmotropic ions. *Biophysical Journal*, 97(2):590–598.

[143] Patel, A., Lee, H. O., Jawerth, L., Maharana, S., Jahnel, M., Hein, M. Y., Stoynov, S., Mahamid, J., Saha, S., Franzmann, T. M., Pozniakovski, A., Poser, I., Maghelli, N., Royer, L. A., Weigert, M., Myers, E. W., Grill, S., Drechsel, D., Hyman, A. A., and Alberti, S. (2015). A liquid-to-solid phase transition of the als protein fus accelerated by disease mutation. *Cell*, 162:1066–1077.

[144] Portz, B., Lee, B. L., and Shorter, J. (2021). Fus and tdp-43 phases in health and disease. *Trends in Biochemical Sciences*, 46(7):550–563.

[145] Qamar, S., Wang, G. Z., Randle, S. J., Ruggeri, F. S., Varela, J. A., Lin, J. Q., Phillips, E. C., Miyashita, A., Williams, D., Ströhl, F., Meadows, W., Ferry, R., Dardov, V. J., Tartaglia, G. G., Farrer, L. A., Schierle, G. S. K., Kaminski, C. F., Holt, C. E., Fraser, P. E., Schmitt-Ulms, G., Klenerman, D., Knowles, T., Vendruscolo, M., and George-Hyslop, P. S. (2018). Fus phase separation is modulated by a molecular chaperone and methylation of arginine cation- interactions. *Cell*, 173:720–734.e15.

[146] Rao, B. S. and Parker, R. (2017). Numerous interactions act redundantly to assemble a tunable size of p bodies innbsp; nbsp; nbsp; nbsp; nbsp; nbsp; saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences*, 114(45).

[147] Ray, S., Singh, N., Kumar, R., Patel, K., Pandey, S., Datta, D., Mahato, J., Panigrahi, R., Navalkar, A., Mehra, S., Gadhe, L., Chatterjee, D., Sawner, A. S., Maiti, S., Bhatia, S., Gerez, J. A., Chowdhury, A., Kumar, A., Padinhateeri, R., Riek, R., Krishnamoorthy, G., and Maji, S. K. (2020). {$${$α$}$}-Synuclein aggregation nucleates through liquid–liquid phase separation. *Nature Chemistry*, 12(8):705–716.

[148] Regy, R. M., Dignon, G. L., Zheng, W., Kim, Y. C., and Mittal, J. (2020). Sequence dependent co-phase separation of RNA-protein mixtures elucidated using molecular simulations. *bioRxiv*.

[149] Reichheld, S. E., Muiznieks, L. D., Keeley, F. W., and Sharpe, S. (2017). Direct observation of structure and dynamics during phase separation of an elastomeric protein. *Proceedings of the National Academy of Sciences*, 114(22).

[150] Rott, R., Szargel, R., Shani, V., Hamza, H., Savyon, M., Abd Elghani, F., Bandopadhyay, R., and Engelender, S. (2017). Sumoylation and ubiquitination reciprocally regulate -synuclein degradation and pathological aggregation. *Proceedings of the National Academy of Sciences*, 114(50):13176–13181.

[151] Rowlinson, J. S. and Widom, B. (2013). *Molecular theory of capillarity*. Courier Corporation.

[152] Rubin, D. B. (1981). The bayesian bootstrap. *The Annals of Statistics*, 9(1).

[153] Ryan, V. H., Dignon, G. L., Zerze, G. H., Chabata, C. V., Silva, R., Conicella, A. E., Amaya, J., Burke, K. A., Mittal, J., and Fawzi, N. L. (2018). Mechanistic View of hnRNPA2 Low-Complexity Domain Structure, Interactions, and Phase Separation Altered by Mutation and Arginine Methylation. *Molecular Cell*, 69(3):465–479.

[154] Sabari, B. R., Dall'Agnese, A., Boija, A., Klein, I. A., Coffey, E. L., Shrinivas, K., Abraham, B. J., Hannett, N. M., Zamudio, A. V., Manteiga, J. C., Li, C. H., Guo, Y. E., Day, D. S., Schuijers, J., Vasile, E., Malik, S., Hnisz, D., Lee, T. I., Cisse, I. I., Roeder, R. G., Sharp, P. A., Chakraborty, A. K., and Young, R. A. (2018). Coactivator condensation at super-enhancers links phase separation and gene control. *Science*, 361(6400).

[155] Sabari, B. R., Dall'Agnese, A., and Young, R. A. (2020). Biomolecular condensates in the nucleus. *Trends in Biochemical Sciences*.

[156] Schmidt, H. B., Barreau, A., and Rohatgi, R. (2019). Phase separation-deficient tdp43 remains functional in splicing. *Nature Communications*, 10(1).

[157] Schuster, B. S., Dignon, G. L., Tang, W. S., Kelley, F. M., Ranganath, A. K., Jahnke, C. N., Simpkins, A. G., Regy, R. M., Hammer, D. A., Good, M. C., and Mittal, J. (2020). Identifying sequence perturbations to an intrinsically disordered protein that determine its phase-separation behavior. *Proceedings of the National Academy of Sciences of the United States of America*, 117:11421–11431.

[158] Schuster, B. S., Good, M. C., and Hammer, D. A. (2018). Controllable protein phase separation and modular recruitment to investigate biochemical compartmentalization in membraneless organelles. *Biophysical Journal*, 114(3).

[159] Solomon, D. A., Smikle, R., Reid, M. J., and Mizielinska, S. (2021). Altered phase separation and cellular impact in c9orf72-linked als/ftd. *Frontiers in Cellular Neuroscience*, 15.

[160] Sprik, M. and Ciccotti, G. (1998). Free energy from constrained molecular dynamics. *The Journal of Chemical Physics*, 109(18):7737–7744.

[161] Srinivasan, J., Cheatham, T. E., Cieplak, P., Kollman, P. A., and Case, D. A. (1998). Continuum solvent studies of the stability of dna, rna, and phosphoramidatedna helices. *Journal of the American Chemical Society*, 120(37):9401–9409.

[162] Strom, A. R., Emelyanov, A. V., Mir, M., Fyodorov, D. V., Darzacq, X., and Karpen, G. H. (2017). Phase separation drives heterochromatin domain formation. *Nature*, 547(7662):241.

[163] Tesei, G. and Lindorff-Larsen, K. (2022). Improved predictions of phase behaviour of intrinsically disordered proteins by tuning the interaction range. *bioRxiv*.

[164] Tesei, G., Schulze, T. K., Crehuet, R., and Lindorff-Larsen, K. (2021). *Accurate model of liquid-liquid phase behaviour of intrinsically-disordered proteins from optimization of single-chain properties*.

[165] Thompson, A. P., Aktulga, H. M., Berger, R., Bolintineanu, D. S., Brown, W. M., Crozier, P. S., in 't Veld, P. J., Kohlmeyer, A., Moore, S. G., Nguyen, T. D., Shan, R., Stevens, M. J., Tranchida, J., Trott, C., and Plimpton, S. J. (2022). LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comp. Phys. Comm.*, 271:108171.

[166] Torrie, G. M. and Valleau, J. P. (1977). Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling.

[167] Uversky, V. N. (2017). Protein intrinsic disorder-based liquid–liquid phase transitions in biological systems: Complex coacervates and membrane-less organelles. *Advances in Colloid and Interface Science*, 239:97 – 114.

[168] Uversky, V. N., Gillespie, J. R., and Fink, A. L. (2000). Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins: structure, function, and bioinformatics*, 41(3):415–427.

[169] Vega, C., Sanz, E., Abascal, J. L. F., and Noya, E. G. (2008). Determination of phase diagrams via computer simulation: methodology and applications to water, electrolytes and proteins. *Journal of Physics: Condensed Matter*, 20(15):153101.

[170] Vernon, R. M., Chong, P. A., Tsang, B., Kim, T. H., Bah, A., Farber, P., Lin, H., and Forman-Kay, J. D. (2018). Pi-pi contacts are an overlooked protein feature relevant to phase separation.

[171] Vitalis, A. and Pappu, R. V. (2009). Absinth: A new continuum solvation model for simulations of polypeptides in aqueous solutions. *Journal of Computational Chemistry*, 30(5):673–699.

[172] Wang, B., Zhang, L., Dai, T., Qin, Z., Lu, H., Zhang, L., and Zhou, F. (2021). Liquid–liquid phase separation in human health and diseases. *Signal Transduction and Targeted Therapy*, 6(1).

[173] Wang, H., Yan, X., Aigner, H., Bracher, A., Nguyen, N. D., Hee, W. Y., Long, B. M., Price, G. D., Hartl, F. U., and Hayer-Hartl, M. (2019). Rubisco condensate formation by ccmm in -carboxysome biogenesis. *Nature*, 566(7742):131–135.

[174] 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., and Hyman, A. A. (2018). A molecular grammar governing the driving forces for phase separation of prion-like rna binding proteins. *Cell*, 174:688–699.e16.

[175] Wang, X., Ramírez-Hinestrosa, S., Dobnikar, J., and Frenkel, D. (2020). The Lennard-Jones potential: when (not) to use it. *Physical Chemistry Chemical Physics*, 22(19):10624–10633.

[176] Wang, X., Schwartz, J. C., and Cech, T. R. (2015). Nucleic acid-binding specificity of human FUS protein. *Nucleic acids research*, 43(15):7535–7543.

[177] Wells, M., Tidow, H., Rutherford, T. J., Markwick, P., Jensen, M. R., Mylonas, E., Svergun, D. I., Blackledge, M., and Fersht, A. R. (2008). Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. *Proceedings of the National academy of Sciences*, 105(15):5762–5767.

[178] Wolf, F. I. and Cittadini, A. (2003). Chemistry and biochemistry of magnesium. *Molecular Aspects of Medicine*, 24(1):3–9.

[179] Wong, C. F. and McCammon, J. (2003). Protein simulation and drug design. *Protein Simulations*, page 87–121.

[180] Wright, R. H., Dily, F. L., and Beato, M. (2019). Atp, mg2+, nuclear phase separation, and genome accessibility. *Trends in Biochemical Sciences*, 44:565–574.

[181] Wu, X., Cai, Q., Shen, Z., Chen, X., Zeng, M., Du, S., and Zhang, M. (2019). Rim and rim-bp form presynaptic active-zone-like condensates via phase separation. *Molecular Cell*, 73(5).

[182] Yang, P., Mathieu, C., Kolaitis, R.-M., Zhang, P., Messing, J., Yurtsever, U., Yang, Z., Wu, J., Li, Y., Pan, Q., and et al. (2020). G3bp1 is a tunable switch that triggers phase separation to assemble stress granules. *Cell*, 181(2).

[183] Yao, R.-W., Xu, G., Wang, Y., Shan, L., Luan, P.-F., Wang, Y., Wu, M., Yang, L.-Z., Xing, Y.-H., Yang, L., and et al. (2019). Nascent pre-rrna sorting via phase separation drives the assembly of dense fibrillar components in the human nucleolus. *Molecular Cell*, 76(5).

[184] You, W., Tang, Z., and Chang, C.-e. A. (2019). Potential mean force from umbrella sampling simulations: What can we learn and what is missed? *Journal of Chemical Theory and Computation*, 15(4):2433–2443.

[185] Yukawa, H. (1955). On the Interaction of Elementary Particles. I. *Progress of Theoretical Physics Supplement*, 1:1–10.

[186] Zbinden, A., Pérez-Berlanga, M., De Rossi, P., and Polymenidou, M. (2020). Phase separation and neurodegenerative diseases: A disturbance in the force. *Developmental Cell*, 55(1):45–68.

[187] Zeng, M., Chen, X., Guan, D., Xu, J., Wu, H., Tong, P., and Zhang, M. (2018). Reconstituted postsynaptic density as a molecular platform for understanding synapse formation and plasticity. *Cell*, 174(5).

[188] Zeng, M., Shang, Y., Araki, Y., Guo, T., Huganir, R. L., and Zhang, M. (2016). Phase transition in postsynaptic densities underlies formation of synaptic complexes and synaptic plasticity. *Cell*, 166(5).

[189] Zeng, X., Holehouse, A. S., Chilkoti, A., Mittag, T., and Pappu, R. V. (2020). Connecting coil-to-globule transitions to full phase diagrams for intrinsically disordered proteins. *Biophysical Journal*, 119(2):402–418.

[190] Zhang, J., Li, X., and Li, J.-D. (2019). The roles of post-translational modifications on -synuclein in the pathogenesis of parkinson's diseases. *Frontiers in Neuroscience*, 13.

[191] Zwanzig, R. W. (1954). High-temperature equation of state by a perturbation method. i. nonpolar gases. *The Journal of Chemical Physics*, 22(8):1420–1426.

Appendix A

A.1 Sequences

A.1.1 FUS variants

A.1.1.1 WT

```
  1 MASNDYTQQA TQSYGAYPTQ PGQGYSQQSS QPYGQQSYSG YSQSTDTSGY GQSSYSSYGQ
 61 SQNTGYGTQS TPQGYGSTGG YGSSQSSQSS YGQQSSYPGY GQQPAPSSTS GSYGSSSQSS
121 SYGQPQSGSY SQQPSYGGQQ QSYGQQQSYN PPQGYGQQNQ YNSSSGGGGG GGGGGNYGQD
181 QSSMSSGGGS GGGYGNQDQS GGGGSGGYGQ QDRGGRGRGG SGGGGGGGGG GYNRSSGGYE
241 PRGRGGGRGG RGGMGGSDRG GFNKFGGPRD QGSRHDSEQD NSDNRADFNR GGGNGRGGRG
301 RGGPMGRGGY GGGGSGGGGR GGFPSGGGGG GGQQGPGGGP GGSHMGGNYG DDRRGGRGGY
361 DRGGYRGRGG DRGGFRGGRG GGDRGGFGPG KMDSRGEHRQ DRRERPY
```

A.1.1.2 PLD 6D

```
  1 MASNDYTQQA TQSYDAYPTQ PGQGYDQQSS QPYDQQSYDG YDQSTDTSGY DQSSYSSYGQ
 61 SQNTGYGTQS TPQGYGSTGG YGSSQSSQSS YGQQSSYPGY GQQPAPSSTS GSYGSSSQSS
121 SYGQPQSGSY SQQPSYGGQQ QSYGQQQSYN PPQGYGQQNQ YNSSSGGGGG GGGGGNYGQD
181 QSSMSSGGGS GGGYGNQDQS GGGGSGGYGQ QDRGGRGRGG SGGGGGGGGG GYNRSSGGYE
241 PRGRGGGRGG RGGMGGSDRG GFNKFGGPRD QGSRHDSEQD NSDNRADFNR GGGNGRGGRG
301 RGGPMGRGGY GGGGSGGGGR GGFPSGGGGG GGQQGPGGGP GGSHMGGNYG DDRRGGRGGY
361 DRGGYRGRGG DRGGFRGGRG GGDRGGFGPG KMDSRGEHRQ DRRERPY
```

A.1.1.3 27R

```
  1 MASNDYTQQA RQSYGAYPTQ PRQGYSQQRS QPYGQQSYSG YSQRTDRSGY GQSSYSSYGQ
 61 RQNTGYGTQR TPQGYGSRGG YGSRQSRQSS YGQQSSYPGY GQQPAPRSRS GSYGSSRQSS
121 SYGQPQRGSY SQQPSYGGRQ QSYGQRQSYN PPQGYGQRNQ YNSSRGRGRG RGRGGNYGQD
181 QRSMSRGGGR GGGYGNQDQR GGGRSGGYGQ QASDRGGRGR GGSGGGGGGG GGGYNRSSGG
241 YEPRGRGGGR GGRGGMGGSD RGGFNKFGGP RDQGSRHDSE QDNSDNNTIF VQGLGENVTI
301 ESVADYFKQI GIIKTNKKTG QPMINLYTDR ETGKLKGEAT VSFDDPPSAK AAIDWFDGKE
361 FSGNPIKVSF ATRRADFNRG GGNGRGGRGR GGPMGRGGYG GGGSGGGGRG GFPSGGGGGG
421 GQQRAGDWKC PNPTCENMNF SWRNECNQCK APKPDGPGGG PGGSHMGGNY GDDRRGGRGG
481 YDRGGYRGRG GDRGGFRGGR GGGDRGGFGP GKMDSRGEHR QDRRERPY
```

A.1.1.4 PLD Y $\rightarrow$ F

```
  1 MASNDFTQQA TQSFGAFPTQ PGQGFSQQSS QPFGQQSFSG FSQSTDTSGY GQSSFSSFGQ
 61 SQNTGFGTQS TPQGFGSTGG FGSSQSSQSS FGQQSSFPGF GQQPAPSSTS GSFGSSSQSS
121 SFGQPQSGSF SQQPSFGGQQ QSFGQQQSFN PPQGFGQQNQ FNSSSGGGGG GGGGGNYGQD
181 QSSMSSGGGS GGGYGNQDQS GGGGSGGYGQ QDRGGRGRGG SGGGGGGGGG GYNRSSGGYE
241 PRGRGGGRGG RGGMGGSDRG GFNKFGGPRD QGSRHDSEQD NSDNRADFNR GGGNGRGGRG
301 RGGPMGRGGY GGGGSGGGGR GGFPSGGGGG GGQQGPGGGP GGSHMGGNYG DDRRGGRGGY
361 DRGGYRGRGG DRGGFRGGRG GGDRGGFGPG KMDSRGEHRQ DRRERPY
```

A.1.1.5 RBD R $\rightarrow$ G

```
  1 MASNDYTQQA TQSYGAYPTQ PGQGYSQQSS QPYGQQSYSG YSQSTDTSGY GQSSYSSYGQ
 61 SQNTGYGTQS TPQGYGSTGG YGSSQSSQSS YGQQSSYPGY GQQPAPSSTS GSYGSSSQSS
121 SYGQPQSGSY SQQPSYGGQQ QSYGQQQSYN PPQGYGQQNQ YNSSSGGGGG GGGGGNYGQD
181 QSSMSSGGGS GGGYGNQDQS GGGGSGGYGQ QDGGGGGGGG SGGGGGGGGG GYNGSSGGYE
241 PGGGGGGGGG GGGMGGSDGG GFNKFGGPGD QGSGHDSEQD NSDNGADFNG GGGNGGGGGG
301 GGGPMGGGGY GGGGSGGGGG GGFPSGGGGG GGQQGPGGGP GGSHMGGNYG DDGGGGGGGY
361 DGGGYGGGGG DGGGFGGGGG GGDGGGFGPG KMDSGGEHGQ DGGEGPY
```

A.1.2 LAF1 RGG variants

A.1.2.1 WT

```
  1 MESNOSNNGG SGNAALNRGG RYVPPHLRGG DGGAAAAASA GGDDRRGGAG GGGYRRGGGN
 61 SGGGGGGGYD RGYNDNRDDR DNRGGSGGYG RDRNYEDRGY NGGGGGGGNR GYNNNRGGGG
121 GGYNRQDRGD GGSSNFSRGG YNNRDEGSDN RGSGRSYNND RRDNGGDG
```

A.1.3 DdX4 variants

A.1.3.1 WT

```
  1 MGDEDWEAEI NPHMSSYVPI FEKDRYSGEN GDNFNRTPAS SSEMDDGPSR RDHFMKSGFA
 61 SGRNFGNRDA GECNKRDNTS TMGGFGVGKS FGNRGFSNSR FEDGDSSGFW RESSNDCEDN
121 PTRNRGFSKR GGYRDGNNSE ASGPYRRGGR GSFRGCRGGF GLGSPNNDLD PDECMQRTGG
181 LFGSRRPVLS GTGNGDTSQS RSGSGSERGG YKGLNEEVIT GSGKNSWKSE AEGGES
```

A.1.3.2 CS

```
  1 MGDRDWRAEI NPHMSSYVPI FEKDRYSGEN GRNFNDTPAS SSEMRDGPSE RDHFMKSGFA
 61 SGDNFGNRDA GKCNERDNTS TMGGFGVGKS FGNEGFSNSR FERGDSSGFW RESSNDCRDN
121 PTRNDGFSDR GGYEKGNNSE ASGPYERGGR GSFDGCRGGF GLGSPNNRLD PRECMQRTGG
181 LFGSDRPVLS GTGNGDTSQS RSGSGSERGG YKGLNEKVIT GSGENSWKSE ARGGES
```

A.1.3.3 R $\rightarrow$ K

```
  1 MGDEDWEAEI NPHMSSYVPI FEKDKYSGEN GDNFNKTPAS SSEMDDGPSK KDHFMKSGFA
 61 SGKNFGNKDA GECNKKDNTS TMGGFGVGKS FGNKGFSNSK FEDGDSSGFW KESSNDCEDN
121 PTKNKGFSKK GGYKDGNNSE ASGPYKKGGK GSFKGCKGGF GLGSPNNDLD PDECMQKTGG
181 LFGSKKPVLS GTGNGDTSQS KSGSGSEKGG YKGLNEEVIT GSGKNSWKSE AEGGES
```

A.1.3.4 F → A

```
  1 MGDEDWEAEI NPHMSSYVPI AEKDRYSGEN GDNANRTPAS SSEMDDGPSR RDHAMKSGAA
 61 SGRNAGNRDA GECNKRDNTS TMGGAGVGKS AGNRGASNSR AEDGDSSGAW RESSNDCEDN
121 PTRNRGASKR GGYRDGNNSE ASGPYRRGGR GSARGCRGGA GLGSPNNDLD PDECMQRTGG
181 LAGSRRPVLS GTGNGDTSQS RSGSGSERGG YKGLNEEVIT GSGKNSWKSE AEGGES
```

A.1.4 G3BP1

```
  1 FGGFVTEPQE ESEEEVEEPE ERQQTPEVVP DDSGTFYDQA VVSNDMEEHL EEPVAEPEPD
 61 PEPEPEQEPV SEIQEEKPEP VLEETAPEDA QKSSSPAPAD IAQTVQEDLR TFSWASVTSK
121 NLPPSGAVPV TGIPPHVVKV PASQPRPESK PESQIPPQRP QRDQRVREQR INIPPQRGPR
181 PIREAGEQGD IEPRRMVRHP DSHEEKKTRA AREGDRRDNR LRGPGGPRGG LGGGMRGPPR
241 GGMVQKPGFG VGRGLAPRQF GGFVTEPQEE SEEEVEEPEE RQQTPEVVPD DSGTFYDQAV
301 VSNDMEEHLE EPVAEPEPDP EPEPEQEPVS EIQEEKPEPV LEETAPEDAQ KSSSPAPADI
361 AQTVQEDLRT FSWASVTSKN LPPSGAVPVT GIPPHVVKVP ASQPRPESKP ESQIPPQRPQ
421 RDQRVREQRI NIPPQRGPRP IREAGEQGDI EPRRMVRHPD SHEEKKTRAA REGDRRDNRL
481 RGPGGPRGGL GGGMRGPPRG GMVQKPGFGV GRGLAPRQ
```

A.1.5 MED1-IDR

```
  1 HHSGSQGPLL TTGDLGKEKT QKRVKEGNGT SNSTLSGPGL DSKPGKRSRT PSNDGKSKDK
 61 PPKRKKADTE GKSPSHSSSN RPFTPPTSTG GSKSPGSAGR SQTPPGVATP PIPKITIQIP
121 KGTVMVGKPS SHSQYTSSGS VSSSGSKSHH SHSSSSSSSA STSGKMKSSK SEGSSSSKLS
181 SSMYSSQGSS GSSQSKNSSQ SGGKPGSSPI TKHGLSSGSS STKMKPQGKP SSLMNPSLSK
241 PNISPSHSRP PGGSDKLASP MKPVPGTPPS SKAKSPISSG SGGSHMSGTS SSSGMKSSSG
301 LGSSGSLSQK TPPSSNSCTA SSSSFSSSGS SMSSSQNQHG SSKGKSPSRN KKPSLTAVID
361 KLKHGVVTSG PGGEDPLDGQ MGVSTNSSSH PMSSKHNMSG GEFQGKREKS DKDKSKVSTS
421 GSSVDSSKKT SESKNVGSTG VAKIIISKHD GGSPSIKAKV TLQKPGESSG EGLRPQMASS
481 KNYGSPLISG STPKHERGSP SHSKSPAYTP QNLDSESESG SSIAEKSYQN SPSSDDGIRP
```

```
541 LPEYSTEKHK KHKKEKKKVK DKDRDRDRDK DRDKKKSHSI KPESWSKSPI SSDQSLSMTS
601 NTILSADRPS RLSPDFMIGE EDDDLM
```

A.1.6 BRD4-IDR

```
  1 CLRKKRKPQA EKVDVIAGSS KMKGFSSSES ESSSESSSSD SEDSETEMAP KSKKKGHPGR
 61 EQKKHHHHHH QQMQQAPAPV PQQPPPPPQQ PPPPPPPQQQ QQPPPPPPPP SMPQQAAPAM
121 KSSPPPFIAT QVPVLEPQLP GSVFDPIGHF TQPILHLPQP ELPPHLPQPP EHSTPPHLNQ
181 HAVVSPPALH NALPQQPSRP SNRAAALPPK PARPPAVSPA LTQTPLLPQP PMAQPPQVLL
241 EDEEPPAPPL TSMQMQLYLQ QLQKVQPPTP LLPSVKVQSQ PPPPLPPPPH PSVQQQLQQQ
301 PPPPPPPQPQ PPPQQQHQPP PRPVHLQPMQ FSTHIQQPPP PQGQQPPHPP PGQQPPPPQP
361 AKPQQVIQHH HSPRHHKSDP YSTGHLREAP SPLMIHSPQM SQFQSLTHQS PPQQNVQPKK
421 QELRAASVVQ PQPLVVVKEE KIHSPIIRSE PFSPSLRPEP PKHPESIKAP VHLPQRPEMK
481 PVDVGRPVIR PPEQNAPPPG APDKDKQKQE PKTPVAPKKD LKIKNMGSWA SLVQKHPTTP
541 SSTAKSSSDS FEQFRRAARE KEEREKALKA QAEHAEKEKE RLRQERMRSR EDEDALEQAR
601 RAHEEARRRQ EQQQQQRQEQ QQQQQQQAAA VAAAATPQAQ SSQPQSMLDQ QRELARKREQ
661 ERRRREAMAA TIDMNFQS
```

A.1.7 Nanog CTD

```
  1 NSNGVTQKAS APTWPSLWSS WHQGCLVNPT GNLPMWSNQT WNNSTWSNQT QNIQSWSNHS
 61 WNTQTWCTQS WNNQAWNSPF WNCGEESLQS CMQFQPNSPA SDLEAALEAA GEGLNVIQQT
121 TRWFSTPQTM DLFLNWSMNM QPEDV
```

A.1.8 Ddx4 and Ddx3 orthologues

A.1.8.1 Ddx4N

```
  1 GAMGSMGDED WEAEINPHMS SYVPIFEKDR YSGENGDNFN RTPASSSEMD DGPSRRDHFM
 61 KSGFASGRNF GNRDAGECNK RDNTSTMGGF GVGKSFGNRG FSNSRFEDGD SSGFWRESSN
121 DCEDNPTRNR GFSKRGGYRD GNNSEASGPY RRGGRGSFRG CRGGFGLGSP NNDLDPDECM
```

```
181 QRTGGLFGSR RPVLSGTGNG DTSQSRSGSG SERGGYKGLN EEVITGSGKN SWKSEAEGGE
241 SSD
```

A.1.8.2 VasaN

```
  1 GAMGSMSDDW DDEPIVDTRG ARGGDWSDDE DTAKSFSGEA EGDGVGGSGG EGGGYQGGNR
 61 DVFGRIGGGR GGGAGGYRGG NRDGGGFHGG RREGERDFRG GEGGFRGGQG GSRGGQGGSR
121 GGQGGFRGGE GGFRGRLYEN EDGDERRGRL DREERGGERR GRLDREERGG ERGERGDGGF
181 ARRRRNEDDI NNNNNIV
```

A.1.8.3 BelN

```
  1 GAMGSMSNAI NQNGTGLEQQ VAGLDLNGGS ADYSGPITSK TSTNSVTGGV YVPPHLRGGG
 61 GNNNAADAES QGQGQGQGQG FDSRSGNPRQ ETRDPQQSRG GGGEYRRGGG GGGRGFNRQS
121 GDYGYGSGGG GRRGGGGRFE DNYNGGEFDS RRGGDWNRSG GGGGGGRGFG RGPSYRGGGG
181 GSGSNLNEQT AEDGQAQQQQ QPRNDRWQEP ERPAGFDGSE GGQSAGGNRS YNNRGERGGG
241 GYNSRWKEGG GSNVDYT
```

A.1.8.4 Ddx3yN

```
  1 GAMGSMSHVV VKNDPELDQQ LANLDLNSEK QSGGASTASK GRYIPPHLRN REASKGFHDK
 61 DSSGWSCSKD KDAYSSFGSR DSRGKPGYFS ERGSGSRGRF DDRGRSDYDG IGNRERPGFG
121 RFERSGHSRW CDKSVEDDWS
```

A.1.8.5 Ddx3xN-Flag

```
  1 GAMGSMSHVA VENALGLDQQ FAGLDLNSSD NQSGGSTASK GRYIPPHLRN REATKGFYDK
 61 DSSGWSSSKD KDAYSSFGSR SDSRGKSSFF SDRGSGSRGR FDDRGRSDYD GIGSRGDRSG
121 FGKFERGGNS RWCDKSDEDD WSDYKDDDDK
```

A.2 Supporting information for Chapter 5

A.2.1 hnRNPA1 variants

The wild-type LCD sequence is shown below.

A.2.1.1 hnRNPA1 WT

residues 186-320 of UniProt sequence P09651-2

```
  1 MASASSSQRG RSGSGNFGGG RGGGFGGNDN FGRGGNFSGR GGFGGSRGGG GYGGSGDGYN
 61 GFGNDGSNFG GGGSYNDFGN YNNQSSNFGP MKGGNFGGRS SGPYGGGGQY FAKPRNQGGY
121 GGSSSSSSYG SGRRF
```

The sequences of the variants of hnRNPA1 we have considered are shown below, using the nomenclature of Bremer and co-workers [29]. The amino-acid residues different from the wild type are highlighted in red.

A.2.1.2 −3R+3K

```
  1 MASASSSQRG KSGSGNFGGG RGGGFGGNDN FGRGGNFSGR GGFGGSKGGG GYGGSGDGYN
 61 GFGNDGSNFG GGGSYNDFGN YNNQSSNFGP MKGGNFGGRS SGGSGGGGQY FAKPRNQGGY
121 GGSSSSSSYG SGRKF
```

A.2.1.3 −4F−2Y

```
  1 MASASSSQRG RSGSGNSGGG RGGGFGGNDN FGRGGNSSGR GGFGGSRGGG GYGGSGDGYN
 61 GFGNDGSNSG GGGSSNDFGN YNNQSSNFGP MKGGNFGGRS SGGSGGGGQY SAKPRNQGGY
121 GGSSSSSSSG SGRRF
```

A.2.1.4 −6R+6K

```
  1 MASASSSQKG KSGSGNFGGG RGGGFGGNDN FGKGGNFSGR GGFGGSKGGG GYGGSGDGYN
 61 GFGNDGSNFG GGGSYNDFGN YNNQSSNFGP MKGGNFGGKS SGGSGGGGQY FAKPRNQGGY
121 GGSSSSSSYG SGRKF
```

A.2.1.5 +7F−7Y

```
  1 MASASSSQRG RSGSGNFGGG RGGGFGGNDN FGRGGNFSGR GGFGGSRGGG GFGGSGDGFN
 61 GFGNDGSNFG GGGSFNDFGN FNNQSSNFGP MKGGNFGGRS SGGSGGGGQF FAKPRNQGGF
121 GGSSSSSSFG SGRRF
```

A.2.1.6 +7K+12D

```
  1 MASADSSQRD RDDKGNFGDG RGGGFGGNDN FGRGGNFSDR GGFGGSRGDG KYGGDGDKYN
 61 GFGNDGKNFG GGGSYNDFGN YNNQSSNFDP MKGGNFKDRS SGPYDKGGQY FAKPRNQGGY
121 GGSSSSKSYG SDRRF
```

A.2.1.7 +7R+12D

```
  1 MASADSSQRD RDDRGNFGDG RGGGFGGNDN FGRGGNFSDR GGFGGSRGDG RYGGDGDRYN
 61 GFGNDGRNFG GGGSYNDFGN YNNQSSNFDP MKGGNFRDRS SGPYDRGGQY FAKPRNQGGY
121 GGSSSSRSYG SDRRF
```

A.2.1.8 −9F+3Y

```
  1 MASASSSQRG RSGSGNFGGG RGGGYGGNDN GGRGGNYSGR GGFGGSRGGG GYGGSGDGYN
 61 GGGNDGSNYG GGGSYNDSGN GNNQSSNFGP MKGGNYGGRS SGGSGGGGQY GAKPRNQGGY
121 GGSSSSSSYG SGRRS
```

A.2.1.9 −12F+12Y

```
  1 MASASSSQRG RSGSGNYGGG RGGGYGGNDN YGRGGNYSGR GGYGGSRGGG GYGGSGDGYN
 61 GYGNDGSNYG GGGSYNDYGN YNNQSSNYGP MKGGNYGGRS SGGSGGGGQY YAKPRNQGGY
121 GGSSSSSSYG SGRRY
```

A.2.2 Sequences of proteins used in radius of gyration calculations

A.2.2.1 α-synuclein [10]

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

A.2.2.2 ACTR [91]

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

A.2.2.3 Ash1 [118]

```
  1 GASASSSPSP STPTKSGKMR SRSSSPVRPK AYTPSPRSPN YHRFALDSPP QSPRRSSNSS
 61 ITKKGSRRSS GSSPTRHTTR VCV
```

A.2.2.4 hNHE1cdt [91]

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

A.2.2.5 IBB [61]

```
  1 GCTNENANTP AARLHRFKNK GKDSTEMRRR RIEVNVELRK AKKDDQMLKR RNVSSFPDDA
 61 TSPLQENRNN QGTVNWSVDD IVKGINSSNV ENQLQAT
```

A.2.2.6 K18 [131]

```
  1 MQTAPVPMPD LKNVKSKIGS TENLKHQPGG GKVQIINKKL DLSNVQSKCG SKDNIKHVPG
 61 GGSVQIVYKP VDLSKVTSKC GSLGNIHHKP GGGQVEVKSE KLDFKDRVQS KIGSLDNITH
121 VPGGGNKKIE
```

A.2.2.7 K25 [131]

```
  1 MAEPRQEFEV MEDHAGTYGL GDRKDQGGYT MHQDQEGDTD AGLKAEEAGI GDTPSLEDEA
 61 AGHVTQARMV SKSKDGTGSD DKKAKGADGK TKIATPRGAA PPGQKGQANA TRIPAKTPPA
121 PKTPPSSGEP PKSGDRSGYS SPGSPGTPGS RSRTPSLPTP PTREPKKVAV VRTPPKSPSS
181 AKSRL
```

A.2.2.8 N49 [61]

```
  1 GCQTSRGLFG NNNTNNININS SSGMNNASAG LFGSKP
```

A.2.2.9 N98 [61]

```
  1 GCFNKSFGTP FGGGTGGFGT TSTFGQNTGF GTTSGGAFGT SAFGSSNNTG GLFGNSQTKP
 61 GGLFGTSSFS QPATSTSTGF GFGTSTGTAN TLFGTASTGT SLFSSQNNAF AQNKPTGFGN
121 FGTSTSSGGL FGTTNTTSNP FGSTSGSLFG P
```

A.2.2.10 NLS [61]

```
  1 ACETNKRKRE QISTDNEAKM QIQEEKSPKK KRKKRSSKAN KPPE
```

A.2.2.11 NSP [61]

```
  1 GCNFNTPQQN KTPFSFGTAN NNSNTTNQNS STGAGAFGTG QSTFGFNNSA PNNTNNANSS
 61 ITPAFGSNNT GNTAFGNSNP TSNVFGSNNS TTNTFGSNSA GTSLFGSSSA QQTKSNGTAG
121 GNTFGSSSLF NNSTNSNTTK PAFGGLNFGG GNNTTPSSTG NANTSNNLFG ATANAN
```

A.2.2.12 NUL [61]

```
  1 GCGFKGFDTS SSSSNSAASS SFKFGVSSSS SGPSQTLTST GNFKFGDQGG FKIGVSSDSG
 61 SINPMSEGFK FSKPIGDFKF GVSSESKPEE VKKDSKNDNF KFGLSSGLSN PV
```

A.2.2.13 NUS [61]

```
  1 GCPSASPAFG ANQTPTFGQS QGASQPNPPG FGSISSSTAL FPTGSQPAPP TFGTVSSSSQ
 61 PPVFGQQPSQ SAFGSGTTPN
```

A.2.2.14 P53 [177]

```
  1 MEEPQSDPSV EPPLSQETFS DLWKLLPENN VLSPLPSQAM DDLMLSPDDI EQWFTEDPGP
 61 DEAPRMPEAA PPVAPAPAAP TPAAPAPAPS WPL
```

A.2.2.15 ProTα [168, 17]

```
  1 MSDAAVDTSS EITTKDLKEK KEVVEEAENG RDAPANGNAE NEENGEQEAD NEVDEEEEEG
 61 GEEEEEEEEG DGEEEDGDED EEAESATGKR AAEDDEDDDV DTKKQKTDED D
```

A.2.2.16 SH4-UD [11]

```
  1 MGSNKSKPKD ASQRRRSLEP AENVHGAGGG AFPASQTPSK PASADGHRGP SAAFAPAAAE
 61 PKLFGGFNSS DTVTSPQRAG PLAGG
```

A.2.2.17 Sic [123]

```
  1 GSMTPSTPPR SRGTRYLAQP SGNTSSSALM QGQKTPQKPS QNLVPVTPST TKSFKNAPLL
 61 APPNSNMGMT SPFNGLTSPQ RSPFPKSSVK RT
```

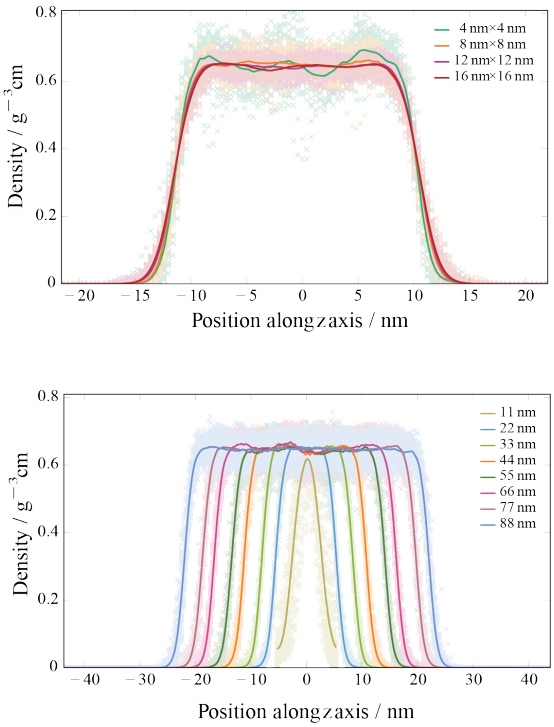


Fig. A.1 **Finite size effect analysis of Mpipi model**.

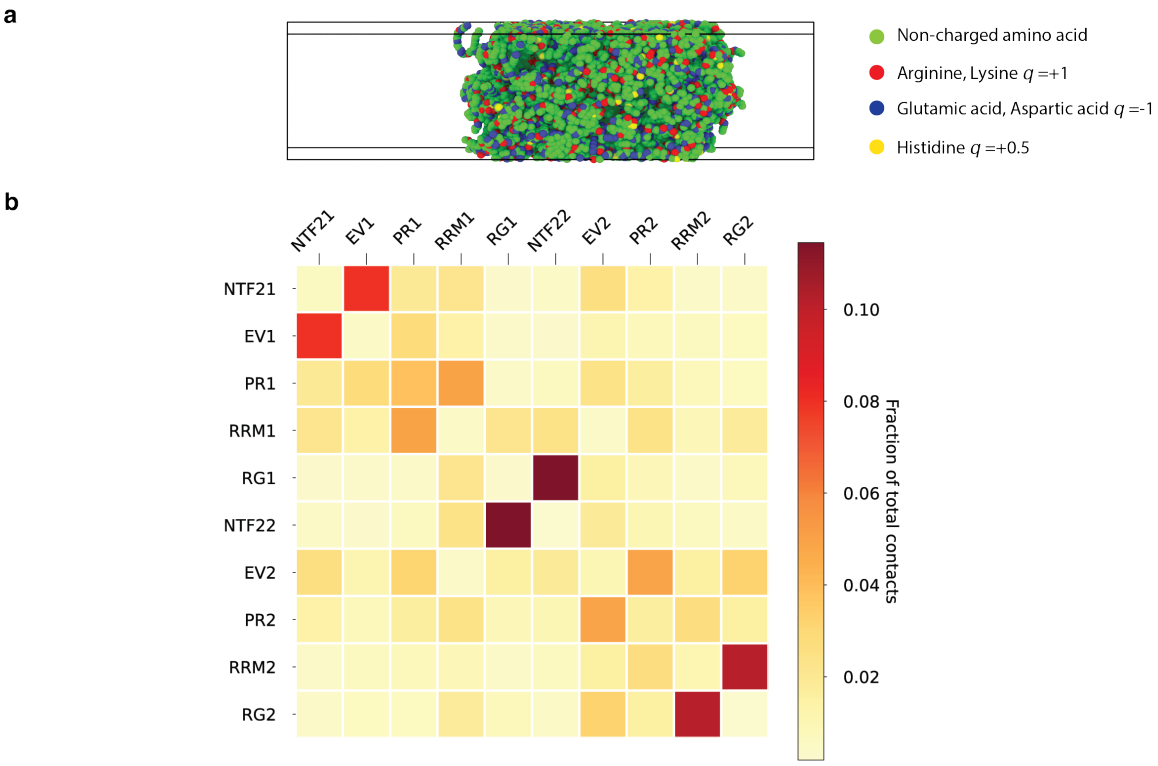


Fig. A.2 **Simulations of G3BP1 dimer with Mpipi model.** (a) Snapshot of direct coexistence simulation of G3BP1 dimer at 300 K. (b) Domain-wise normalised contact map of G3BP1 dimer from direct coexistence simulation at 300 K.

Table A.1 **Experimental radii of gyration for proteins, alongside the experimental salt concentration and the corresponding Debye screening constant** (computed using the equation immediately following Eq. (12) of Ref. 44 expressed in SI instead of gaussian units).

Protein	R_g / nm	[salt] / mM	κ / nm^{-1}
α-synuclein [10]	3.31	185	1.40
ACTR [91]	2.51	199	1.45
Ash1 [118]	2.85	150	1.26
hNHE1cdt [91]	3.63	199	1.45
IBB [61]	3.20	162	1.31
K18 [131]	3.80	163	1.31
K25 [131]	4.40	163	1.31
N49 [61]	1.59	162	1.31
N98 [61]	2.86	162	1.31
NLS [61]	2.40	162	1.31
NSP [61]	4.10	162	1.31
NUL [61]	3.00	162	1.31
NUS [61]	2.49	162	1.31
P53 [177]	2.87	208	1.49
ProTα [168, 17]	3.79	155	1.28
SH4-UD [11]	2.90	217	1.52
Sic1 [68]	3.21	162	1.31

A.3 Supporting information for Chapter 6

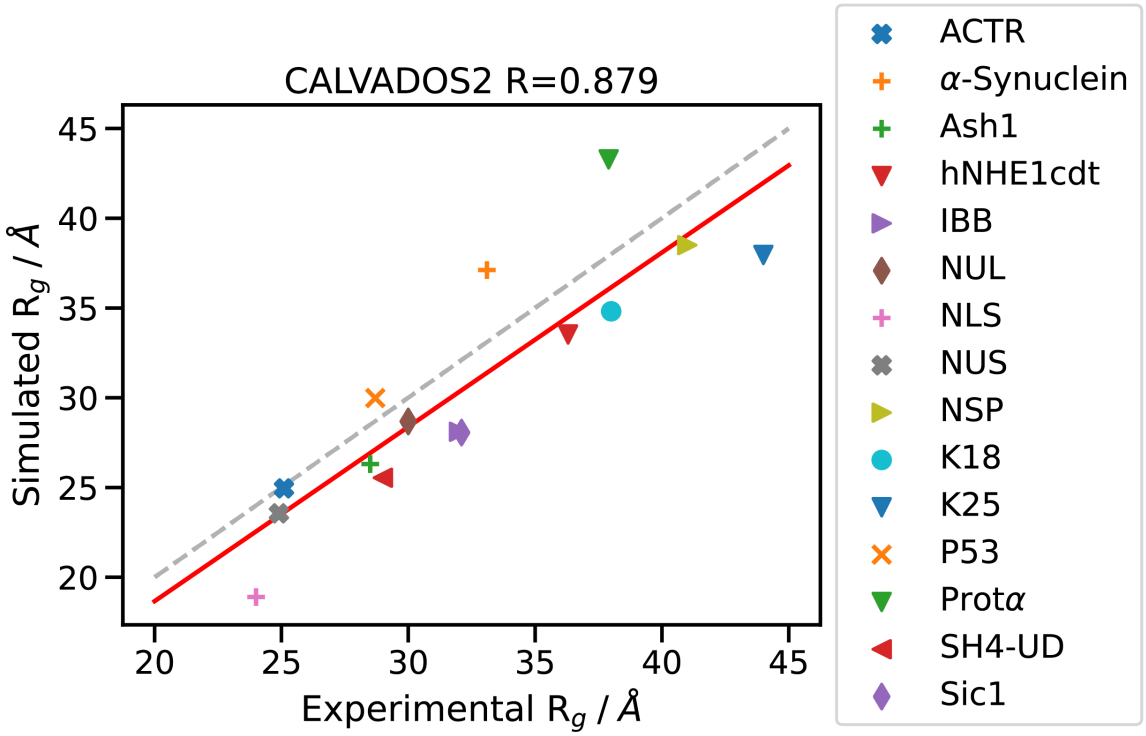


Fig. A.3 **Radii of gyration of IDPs tested on Calvados2 model developed by Tesei *et al*. [164].**

A.4 Supporting information for Chapter 7

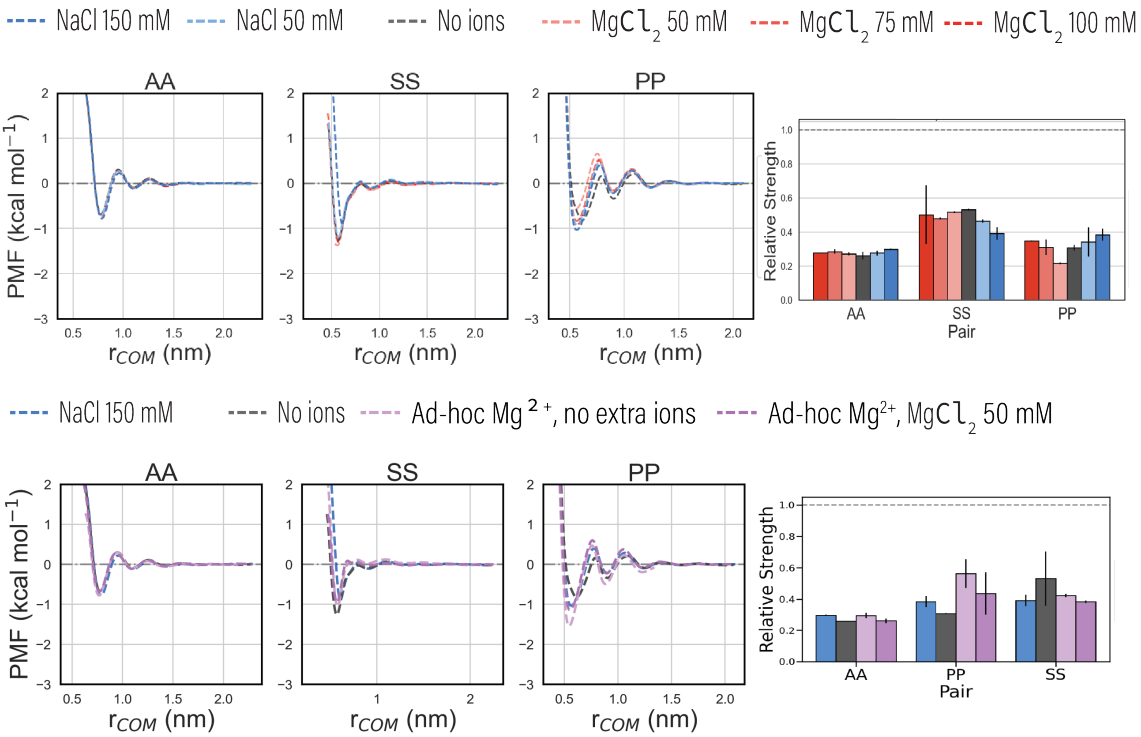


Fig. A.4 **Strengths of pairwise non-π, uncharged residue-residue interactions, namely AA (hydrophobic), SS (polar) and PP (non-polar)**. The top plots show the PMF curves (left) and relative interaction strengths (right) at different conditions of salt, including in the presence of NaCl, in the presence of increasing concentrations of $MgCl_2$ in which none of the Mg^{2+}ions binds to the residue pair through its first hydration layer, or in the absence of any salt. The bottom plots, on the other hand, show the PMF curves, on the left, and relative interaction strengths, on the right, including two cases in which Mg^{2+}ions are directly bound to the residue pair through their first hydration layer. The relative interaction strengths are computed via the depth of the well in the case of attractive interactions, and as the inflection point in repulsive interactions, then normalised by said value for RY interaction in no-salt condition.

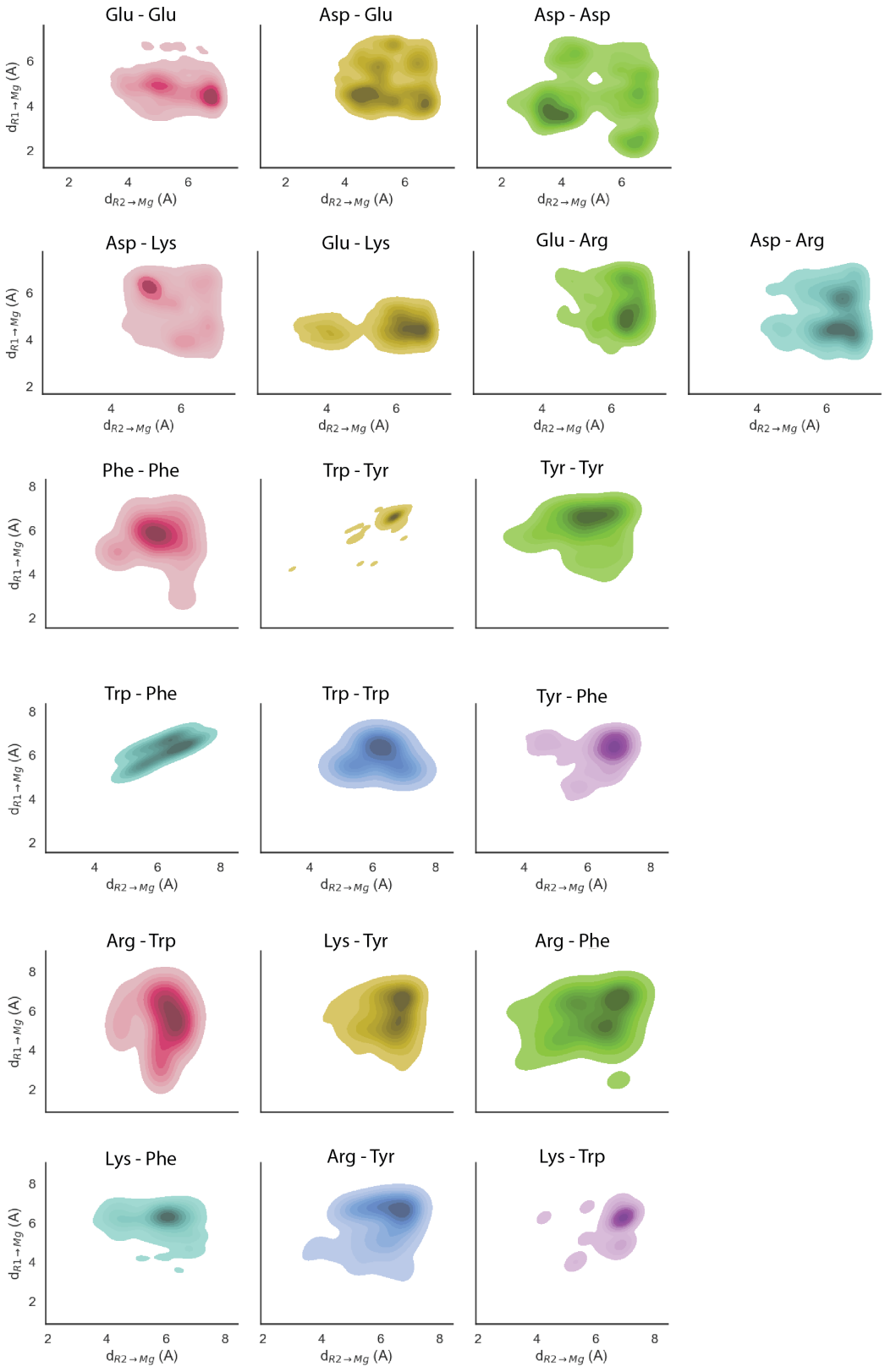


Fig. A.5 **Distribution of center-of-mass distances between specified residues mediated by Mg^{2+} ions.**

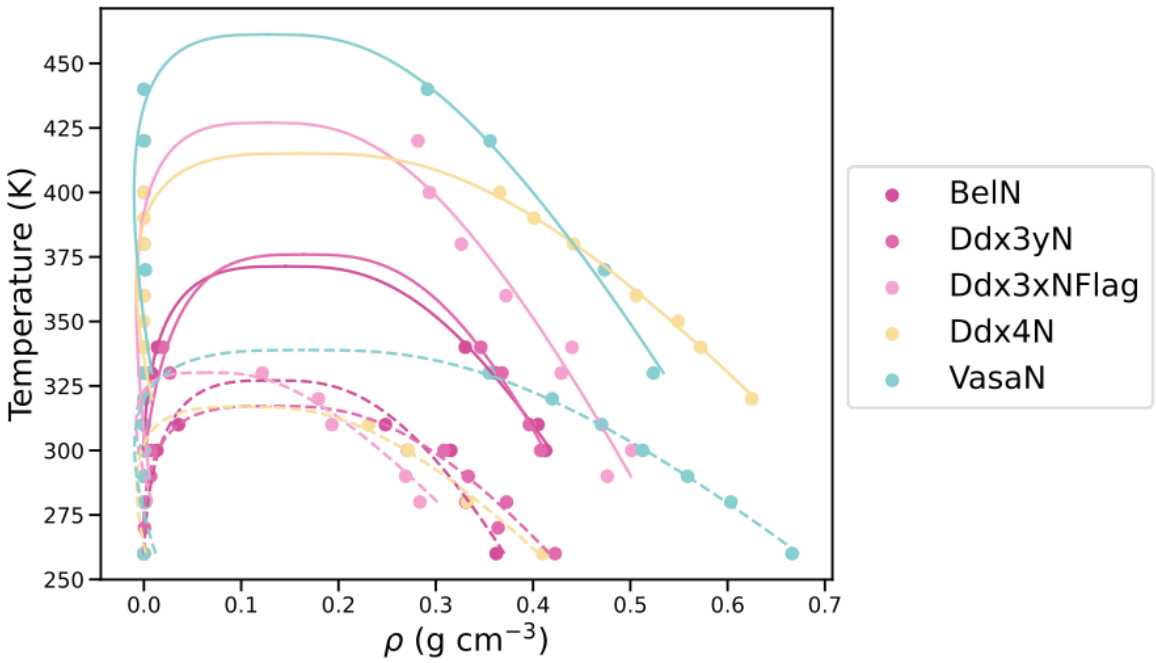


Fig. A.6 **LLPS propensities of Ddx4 and Ddx3 orthologues, as temperature-dependant phase diagrams with and without Mg^{2+}-mediated interactions.**

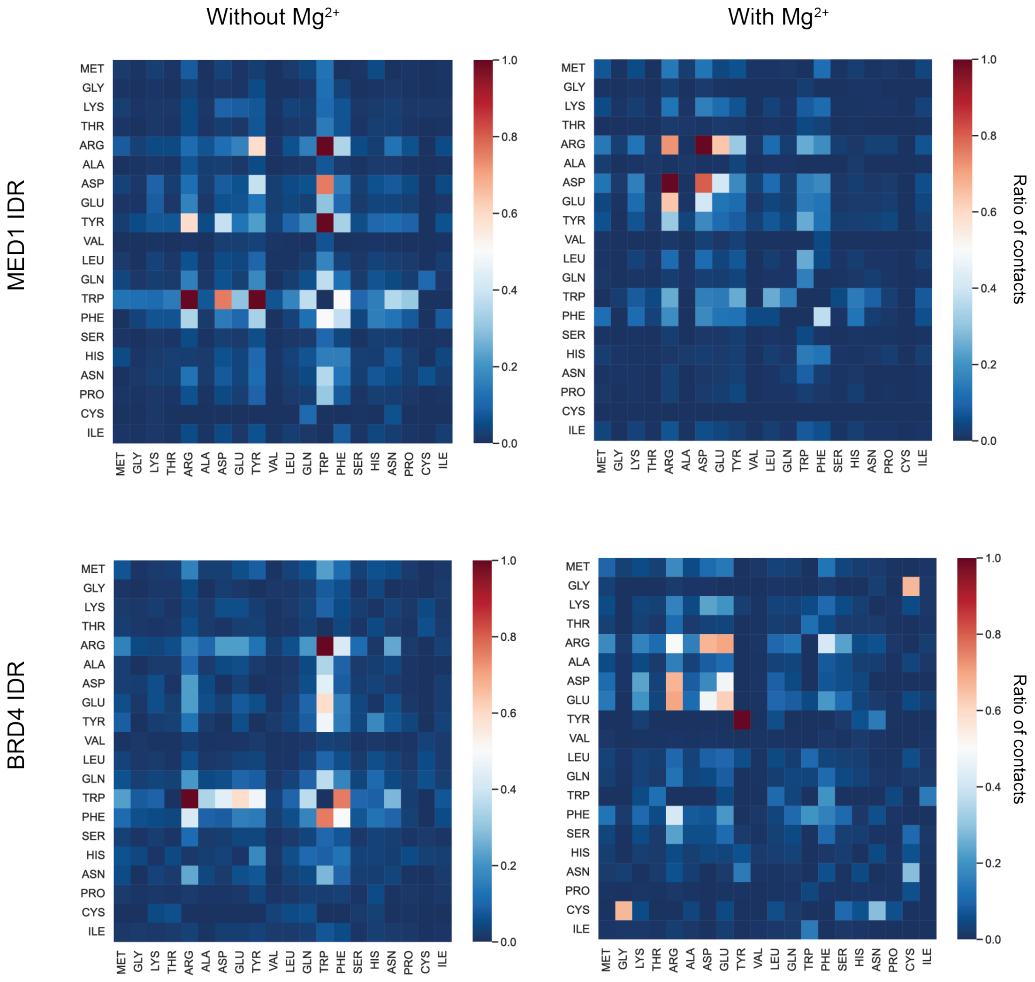


Fig. A.7 **Comparison of residue–residue contacts of MED1 (top) and BRD4 (bottom) without and with Mg^{2+}.**

Simulation Type	Simulation resolution	Force field	Integrator Timestep	Simulation Length
Umbrella Sampling	All-atom	Amber03ws [22]	2 fs	30 ns
Direct Coexistence	One bead per residue	Mpipi [83]	10 fs	0.5–1 μs
Direct Coexistence	One bead per residue	Mpipi Recharged	10 fs	1–2 μs
Direct Coexistence	One bead per residue	MagPi	10 fs	1–2 μs

Table A.2 **Simulation and force field resolution of MD models used in Chapters 4 to 7.**