Published in final edited form as:
Nat Comput Sci. 2021 November ; 1(11): 732–743. doi:10.1038/s43588-021-00155-3.

Physics-driven coarse-grained model for biomolecular phase separation with near-quantitative accuracy

Jerelle A. Joseph[#1,2,3,*], **Aleks Reinhardt**[#1,‡], **Anne Aguirre**[1], **Pin Yu Chew**[1], **Kieran O. Russell**[1], **Jorge R. Espinosa**[2], **Adiran Garaizar**[2], **Rosana Collepardo-Guevara**[1,2,3,§]

[1]Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge, CB2 1EW, UK

[2]Department of Physics, University of Cambridge, Cambridge, CB3 0HE, UK

[3]Department of Genetics, University of Cambridge, Cambridge, CB2 3EH, UK

[#] These authors contributed equally to this work.

Abstract

Various physics- and data-driven sequence-dependent protein coarse-grained models have been developed to study biomolecular phase separation and elucidate the dominant physicochemical driving forces. Here, we present Mpipi, a multiscale coarse-grained model that describes almost quantitatively the change in protein critical temperatures as a function of amino-acid sequence. The model is parameterised from both atomistic simulations and bioinformatics data and accounts for the dominant role of π–π and hybrid cation-π/π–π interactions and the much stronger attractive contacts established by arginines than lysines. We provide a 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; Mpipi predictions agree well with experiment on both fronts. Moreover, it can account for protein–RNA interactions, correctly predicts the multiphase behaviour of a charge-matched poly-arginine/poly-lysine/RNA system, and recapitulates experimental LLPS trends for sequence mutations on FUS, DDX4 and LAF-1 proteins.

Introduction

Under certain conditions, macromolecules within cells demix into membraneless organelles. These organelles, often termed biomolecular condensates, are sustained by the physicochemical process of liquid-liquid phase separation (LLPS) [1, 2] The ensuing condensates play important roles in cellular function as well as dysfunction [3]; therefore, delineating the mechanisms of intracellular LLPS is now an active area of research.

[*] jaj52@cam.ac.uk . [‡] ar732@cam.ac.uk . [§] rc597@cam.ac.uk .

Author Contributions Statement

J.A.J. and R.C.-G. conceived the project; J.A.J., A.R. and R.C.-G. designed the model and benchmarking framework; J.A.J. and A.R. implemented and optimised the model; J.A.J., A.R., A.A., P.Y.C., K.O.R., J.R.E. and A.G. validated the model and analysed the data; J.A.J. and A.R. wrote the manuscript with help from R.C.-G; all authors reviewed the manuscript; J.A.J., A.R. and R.C.-G. acquired funding; and J.A.J., A.R. and R.C.-G. supervised the research.

Competing Interests Statement

The authors declare no competing interests.

Europe PMC Funders Author Manuscripts

Intracellular LLPS is principally driven by biomolecular multivalency, i.e. the ability of multidomain proteins, intrinsically disordered proteins/regions (IDPs/IDRs), ribonucleic acids (RNA) and chromatin to engage multiple interaction partners simultaneously. This multivalency is, in turn, predominantly encoded in the chemical makeup of the macromolecules in question. It is well-established that both hydrophobic and electrostatic interactions are important drivers of biomolecular LLPS, including charge–charge, π–π, cation–π, dipole–dipole and non-polar interactions. Additionally, there is strong evidence that certain chemical building blocks have a bigger stake in biomolecular LLPS than others [4–7]. Biophysical models for studying LLPS must therefore be able to capture the correct balance between these myriad driving forces. In this paper, we aim to achieve precisely this.

Together, the stickers-and-spacers framework of Pappu and colleagues [8, 9] and the quantitative experimental phase diagrams of Mittag and colleagues position aromatic residues as being chief drivers of biomolecular phase separation [4, 10]. Moreover, it is evident that even within the subset of aromatic residues, tyrosine, for example, is a stronger contributor than phenylalanine to LLPS stability [7, 10–12], perhaps because it has more side-chain binding modes than phenylalanine: in addition to forming aromatic π–π contacts, tyrosine can form strong hydrogen bonds via its phenol group.

The dominant role of π–π interactions in LLPS was also suggested by Vernon et al. [13], who, via a comprehensive survey of the protein data bank (PDB), identified an abundance of planar π–π contacts involving not only aromatic but also non-aromatic residues in protein structures. Additionally, in recent work, π–π interactions emerged as a major driver of LLPS at both low and high salt concentration [7]. 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 [7].

The role of hydrophobic π–π contacts in LLPS is even more obvious when we consider differences in the strengths of cation–π interactions. Notably, cationic residues, namely arginine and lysine, have been shown to act as unequal contributors to LLPS [6, 7, 14–16]: arginine establishes appreciably stronger cation–π and charge–charge interactions due to the presence of the guanidinium group [6, 7, 13, 16–18]. Furthermore, Pappu and colleagues have recently demonstrated that the free energy of hydration of arginine is considerably less favourable than that of lysine; thus, although they both carry the same charge, arginine is significantly more hydrophobic than lysine [16]. Since π–based contacts play such a dominant role in biomolecular LLPS, achieving the correct balance of these interactions is essential for making quantitative predictions.

Complementing experimental and theoretical work, computer simulations have provided a unique lens for probing biomolecular LLPS. Because LLPS is a collective phenomenon, coarse-graining is essential to reduce the system dimensionality while retaining essential physicochemical information and allowing sufficient sampling of phase space in computationally tractable time scales. There are numerous possible approaches for parameterising biomolecular coarse-grained models [21], from ‘bottom-up’ strategies that rely on higher-resolution models [9, 10, 22–24], to ‘knowledge-driven’ approaches that

aim to reproduce experimental properties using a data-based parameterisation [25–27], to 'top-down' strategies that account for emergent behaviour by approximating fundamental physical forces [28, 29], to combinations of these [30, 31]. Coarsegrained models can also be broadly classed as 'system-specific', bottom-up parameterisations focussing on finding an optimum representation for a particular system using fine-grained simulations as a reference, often derived in a systematic way using for instance iterative Boltzmann inversion [32, 33] or force matching [34, 35], and 'transferable', either bottom-up or top-down parameterisations, aiming to achieve a generally applicable potential.

Developing a coarse-grained model involves invoking multiple approximations and making design decisions (e.g. type of model, resolution, bead characteristics, types of interaction) that are more or less appropriate depending on the question being investigated. As discussed by Choi et al. [9], there is no unequivocal reason that makes one scheme intrinsically superior to others: each approach has its advantages and drawbacks. For instance, systematic multiscale coarse-graining from higher-resolution models [23, 36] by construction results in an excellent description of the system under investigation and allows us to work out precisely what underlying building blocks have been coarse-grained. However, system-specific coarse-graining does not generally result in a unique solution [37] and requires sufficiently long simulations of the entire system of interest to be run with an expensive high-resolution potential. Indeed, bulk phase behaviour can be significantly different between a machine-learned potential and the underlying quantum-mechanical potential-energy surface even for systems much simpler than biomolecules [38], illustrating the significant challenge of this approach. Similarly, transferable data-driven or machine-learning-based approaches can give excellent agreement with the data they were parameterised from, but, since in such high-dimensional problems, many solutions are similarly good, careful curation is required to obtain statistically meaningful results for a specific system, and even these may still not be transferable to similar molecules [39]. On the other hand, a transferable 'physics-based' approach to interaction parameters provides us with a simple way of rationalising complex behaviour based on relatively simple interactions. However, it risks introducing our biases of which interactions are important into the predictions of the model, and the predictions and rationalisations of observed behaviours with such models can therefore be to some extent self-fulfilling.

Various coarse-graining strategies have now been applied to gain insights into the problem of biomolecular LLPS. For example, the mean-field stickers-and-spacers model can be parameterised to reach quantitative agreement with experimental phase diagrams of specific proteins, providing a tool to dissect the driving forces behind the observations [4, 8–10]. A different approach, pioneered by Mittal, Best and colleagues [28], combines residue-level coarse-grained models with direct-coexistence simulations [20], offering a transferable method to predict protein phase diagrams and augmenting our ability to link molecular sequences to their experimental phase behaviour.

Inspired by previous computational work and guided by the accumulated knowledge of the LLPS interaction landscape, we set out to design a chemically accurate coarse-grained model for predicting biomolecular LLPS. Specifically, the model aims to achieve optimal strengths of protein–protein and protein–RNA interactions. We demonstrate that our model

can accurately predict biomolecular phase separation while achieving quantitative agreement with experiment, recapitulating post-translational modification effects, and even capturing more complex features such as sequence-dependent multiphasic compartmentalisation. Simulations using our model are particularly simple to set up since all its components are already implemented in open-source software.

In what follows, we first describe the design of our multiscale π–π model (termed 'Mpipi') for probing phase separation of biomolecules. We outline the use of atomistic potential-of-mean-force (PMF) calculations coupled with bioinformatics data for yielding a chemically accurate interaction scale for coarse-grained simulations [Fig. 1]. We then assess the balance of key interactions in the Mpipi model alongside other commonly used residue-based coarse-grained models. Finally, we present benchmarks for several LLPS systems and directly compare our predictions with other models and against quantitative experimental phase diagrams and other experiments.

Results

Designing coarse-grained models for biomolecular LLPS

We have designed a residue-level coarse-grained model for predicting biomolecular phase behaviour (Fig. 1a–c) that captures the fundamental Van der Waals and electrostatic interactions of a 'top-down' approach and the interaction strengths obtained from 'bottom-up' atomistic simulations and bioinformatics data. In the Mpipi model, each amino (or nucleic) acid is mapped onto a unique bead (Fig. 1b) based on simulation and experimental data. Following Dignon et al. [28], the potential energy of molecules is computed as the sum of a harmonic bond energy, Debye–Hückel and short-ranged energy terms (Methods), which account for π–π, cation–π and other non-ionic interactions. 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 parameterisation of short-ranged interactions, and (3) the relative contribution of long-ranged electrostatics and short-ranged terms to the total energy. Specifically, for short-range interactions, we use the recently developed Wang–Frenkel [19] pair potential (Fig. 1b; see Methods), which accounts for key physical interactions, namely a short-ranged excluded-volume repulsion and a longer-ranged attraction which gradually decays to zero. The Wang–Frenkel potential has several advantages [19] over Lennard-Jones-like potentials that are commonly adopted in molecular simulations. We outline these and how our model is fitted within the Wang–Frenkel framework in the Methods section.

When deciding on the energy scale for short-ranged interactions, our main objective is to achieve the correct balance of π–π and non-π-based contacts. To this end, we combine bioinformatics data and atomistic short-ranged free energy estimates (Fig. 2a–d). In the Methods section, we explain our parameterisation of short-ranged pairwise contacts and long-ranged charge–charge interactions (Fig. 2e).

Cation–π, π–π and non-π interactions in residue-level models

To validate our model parameters, we first compare the Mpipi model with seven other residue-level coarse-grained models for LLPS, namely the KH (Kim–Hummer) [28], HPS-KR (hydrophobicity scale) [28], FB-HPS [26] and HPS-Urry [29] models, as well as the HPS model with augmented cation–π interactions [schemes (i) and (ii)] [40]. We also include analyses for TSCL-M2, the M2 parameter set of Tesei et al. [27] proposed while our work was under review. Das et al. [40] recently provided a thorough comparison of the KH, HPS-KR, HPS+cation–π(i) and HPS+cation–π(ii) models. Below, we briefly discuss the key features of all models and then evaluate them in terms of the balance of π–π, cation–π and non-π-based interactions.

A key difference between Mpipi and other residue-level models is the parameterisation of short-ranged interactions. In the KH model [28, 41], short-ranged interactions (ε_{ij}) are based on residue contact statistics of folded proteins. The energy scale of the KH model was tuned to reproduce approximately the radii of gyration of selected unfolded proteins/IDPs [28]; here, we utilise parameter set D [28] for interactions involving disordered proteins. The KH model has been successfully used to predict LLPS propensities for variants of the N-terminal domain (NTD) of the DEAD-Box Helicase 4 (DDX4) protein [40] and to describe qualitatively the phase behaviour of the proteins Fused in Sarcoma (FUS) and Lethal-And-Feminizing-1 (LAF-1) [28].

The next model, HPS-KR [28] is perhaps the most widely used sequence-based continuum model for studying biomolecular LLPS. In this model, short-ranged interactions are based on the hydrophobicity scale of Kapcha and Rossky [42], and each amino acid is assigned a λ_i value which accounts for its 'hydrophobicity', and residue–residue contacts (λ_{ij}) are determined by the arithmetic mean of the λ_i values of each residue [43]. Additionally, the absolute energy scale of the model was optimised to reproduce experimental radii of gyration (R_g) of an IDP subset. However, as previously noted [40], the HPS-KR model is inconsistent with experimental data when accounting for the balance between Arg and Lys interactions. An improved version of the HPS model, HPS-Urry [29], was recently parameterised, which employs instead the hydrophobicity scale of Urry et al. [44] to determine λ_{ij}. Moreover, two free parameters (μ and Δ) are introduced to scale and shift the λ_{ij} values; these are optimised to reproduce experimental R_g [29]. Recently, Dannenhoffer-Lafage and Best [26] also reparameterised the short-ranged interactions in the HPS-KR model by employing machine-learning techniques. Their model, FB-HPS, was optimised against experimental R_g of unfolded, phase-separating and intrinsically disordered proteins.

Prior to these studies, Das et al. [40] augmented the HPS-KR model so as better to account for cation–π interactions. They presented two schemes: scheme (i), where Arg/Lys–π interactions are scaled uniformly, and scheme (ii), where they 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 [40]. In another study, it was demonstrated that the HPS+cation–π(i) model can reasonably reproduce experimental trends of selected RNA binding proteins [45]. Here, we have considered these augmented models to achieve a more complete view of how cation–π interactions contribute to biomolecular LLPS.

Recently, Tesei and coworkers [27] used experimental data to reparameterise the hydrophobicity scale of HPS-KR via a Bayesian parameter-learning procedure. The M2 parameter set predicted well both single-molecule and collective behaviours of the tested IDPs; we have therefore included benchmarks for this parameter set in our work.

Fig. 3 summarises the relative contributions of selected π–π and non-π-based interactions for the residue-level coarse-grained models assessed in this work (see SI Fig. S5 and SI Fig. S6 for the HPS+cation–π(i) and TSCL-M2 models). The relative interaction strengths are obtained by computing the integral of the curves of the short-ranged potential (with consistent limits of σ–3σ). In the Mpipi, KH, FB-HPS and TSCL-M2 models, aromatic residue pairs (magenta bars in Fig. 3a,b,e and SI Fig. S6a) are generally considerably stronger than residue pairs not involving π contacts (dark yellow bars in Fig. 3a,b,e and SI Fig. S6a). Hence, consistently with the stickers-and-spacers framework, YY and FF are expected to act as stickers, while AA, SS and PP should behave as spacers in these models. Interestingly, in the FB-HPS model, glycine (see Figure 4 of Dannenhoffer-Lafage and Best [26]), which is normally classified as a spacer, has an interaction strength that is stronger than even the aromatic residues. While this result is attributed to glycine forming strong backbone π–π contacts [26], mutational studies have consistently found that replacing sticker-like residues with Gly significantly suppresses biomolecular LLPS [10, 11]. The stronger contacts for Gly arising from the machine-learning algorithm optimisation [26] may be a result of how commonly occurring glycine is in many proteins, particularly those for which experimental radii of gyration are available.

With regards to Tyr versus Phe, a survey of the PDB [13], our atomistic PMF calculations (Fig. 2d) and experiments [10–12] all suggest that Phe–Phe contacts are weaker than Tyr–Tyr ones. By contrast, the KH, HPS-KR, FB-HPS, HPS+cation–π(i) and HPS+cation–π(ii) models all predict stronger Phe–Phe contacts than Tyr–Tyr interactions (Fig. 3). The trend for the KH model is particularly striking, with the weighted interaction energy of FF predicted to be about twice that of YY (Fig. 3b). Taken together, we do not expect these models to reproduce LLPS propensities faithfully as far as Tyr vs Phe mutations are concerned.

We also examine the relative strengths of cation–π interactions in the coarse-grained models, again focussing on the contributions of cation–π contacts versus aromatic π–π contacts and Arg–π contacts compared to Lys–π ones. The HPS+cation–π(ii) (Fig. 3d) and the TSCL-M2 (SI Fig. S6a) models are most similar to the Mpipi model in terms of the relative contributions of Arg–π and Lys–π contacts. While in the KH model, Arg–π interactions are also stronger than Lys–π ones, the overall strength of these is low, which makes Arg–π interactions closer in strength to spacer-type interactions (Fig. 3b); thus, the dominant role of these interactions may not be properly accounted for in the KH model. The HPS-Urry model also captures the overall trend for Arg –vs Lys–π contacts (Fig. 3f); in addition, the weights of these interactions are more similar than those encoded in Mpipi, HPS+cation–π(ii) and TSCL-M2. The opposite trend is found in both the HPS-KR and the FB-HPS models, where Lys–π interactions are now stronger than Arg–π interactions. Moreover, in the HPS-KR model, non–π-based interactions are comparable to (or even stronger than) cation–π interactions (Fig. 3c). We therefore speculate that the HPS-KR

model might represent an upper bound in terms of predicting LLPS propensities of proteins. Strikingly, in the HPS+cation–π(i) model, cation–π interactions convincingly dominate all other types of interaction (SI Fig. S5). LLPS systems driven by cation–π interactions are thus likely to be over-stabilised with this model [40].

Estimating single-molecule radii of gyration

Certain single-molecule properties of proteins, such as R_g in the context of coil-to-globule transitions, are often governed by similar driving forces as bulk LLPS [46–48], and coiling transitions are sometimes used as a proxy for the upper critical solution temperature (T_c) [24]. Importantly, the strong correlation between single-molecule dimensions and T_c has been used as a target for optimising coarse-grained LLPS models. It is often assumed that models that correctly reproduce experimental R_g of single proteins (at infinite dilution) should accurately predict homotypic LLPS propensities.

Accordingly, we tested the ability of Mpipi to recapitulate experimental R_g of IDPs (Fig. 4a; see also SI Sec. S2.1 and SI Table III). The set of IDPs has a good distribution of net charge: from –44e for ProTα to +16e for Ashl, where e is the elementary charge. These proteins therefore provide an indirect measure of how well electrostatic and short-ranged pairwise interactions are balanced in the coarse-grained models. Most of the proteins in our test set largely comprise 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 that lack π electrons, since these residues form weaker contacts and so the resulting proteins are less prone to aggregation and self-assembly. Notwithstanding the dominance of this class, the test set of proteins does exhibit appreciable variation in protein composition.

Fig. 4b–g compares simulated R_g with experiment R_g for each coarse-grained model we have considered. Each protein is coloured according to its dominant residue class, using the same colouring code as Fig. 4a but ignoring the neutral class. The Mpipi, FB-HPS (Fig. 4f), HPS-Urry (Fig. 4g) and TSCL-M2 (SI Fig. S6b) models achieve the closest match with experiment. This result is not unexpected for the last three models, since they were all optimised to reproduce experimental R_g values, and several proteins in the current study were used either to train or to validate the respective model parameters. Importantly, compared to HPS-KR, HPS-Urry and TSCL-M2 both perform better at predicting single-molecule radii of gyration. Thus, in this regard, these models fulfil their goal of offering an improvement over their common predecessor.

Notably, Mpipi (Fig. 4b), whose parameters were not optimised on R_g data but rather on physicochemical information, is able to predict the R_g values to within a root mean squared deviation of 0.3 nm for the tested IDPs. Fits to the bioinformatics data and atomistic PMFs therefore appear to be physically sound, at least with respect to capturing sequence-dependent single-molecule chain dimensions.

While the HPS+cation–π(ii) (Fig. 4e), HPS-KR (Fig. 4d) and HPS+cation–π(i) (SI Fig. S5) models yield reasonable agreement with experiment, all generally predict more compact proteins than experiments. Interestingly, the KH model (Fig. 4c) gives the poorest agreement with experiment, perhaps because short-ranged pairwise interactions in the KH model were

obtained from residue-residue contacts in folded proteins and may thus overestimate the relative strengths of such interactions.

Recapitulating the phase behaviour of A1-LCD variants

To ascertain the extent to which the Mpipi potential is able to capture the bulk properties of protein solutions, we compute the critical solution temperatures for a series of variants of the LCD of the heterogeneous nuclear ribonucleoprotein A1 (hnRNPA1), referred to here as A1-LCD, whose experimental phase diagrams were recently determined [10].

We estimate the experimental critical temperatures [Fig. 5a] of a range of A1-LCD variants, using the fitting procedure described in SI Sec. S2.2. The variants' sequences are given in SI Sec. S2.2, following the nomenclature of Bremer and co-workers [10]. For each model, we show the computed phase diagrams in Fig. 5b–g, and the correlation between simulation and experimental values in Fig. 5h–m. The corresponding data for the HPS+cation-π(i) and TSCL-M2 models are provided in SI Fig. S5d,e and SI Fig. S6d,e, respectively.

Although the fitted linear regression has a positive gradient for the eight models considered, indicating that, broadly speaking, all the models capture some of the underlying physics, the Pearson correlation coefficient varies significantly across the models. Mpipi achieves a Pearson correlation coefficient (r) of 0.97 (Fig. 5h), indicating that the parameterisation works well not only for single-molecule properties (Fig. 4), but also accounts for bulk behaviour. The HPS-Urry (Fig. 5m; $r = 0.91$), TSCL-M2 (SI Fig. S6e; $r = 0.80$) and HPS+cation–π(ii) (Fig. 5k; $r = 0.79$) models also achieve high correlation with experiment. By contrast, the FB-HPS model, which was parameterised principally on R_g data, performs quite well when predicting the radius of gyration, but the improvement of its predictions of the phase behaviour of the A1-LCD variants relative to the underlying HPS-KR potential ((Fig. 5j; $r = 0.35$) is only marginal ((Fig. 5l; $r = 0.37$). Although R_g properties do correlate with phase behaviour in experiment, there are evidently several parameterisations of coarse-grained models which are able to capture one property but not the other.

Coarse-graining necessarily entails integrating out some degrees of freedom and so interaction 'energies' are therefore approximate free energies. It is thus not likely that such models could capture faithfully the behaviour of protein systems far from the temperature range in which they were parameterised. Nevertheless, given that all the models were parameterised to reproduce protein behaviour close to room temperature, we may also consider the agreement between the experimental and simulation temperature scales. To this end, we can compare the deviation of the experimental and simulation critical temperatures. For the A1-LCD wild type, this can be visualised by comparing the solid horizontal black lines in Fig. 5b–g, representing the experimental critical solution temperature of the A1-LCD wild type, to the maximum temperature of the binodals in black, the simulation results for the same system. In addition, a quantification of the deviation between the experimental and simulation results is given by the difference in slope between the black linear fits shown in Fig. 5h–m and the $y = x$ lines shown in red, as well as by the root mean squared deviation between the experimental and simulation critical temperatures (D).

These comparisons demonstrate that, at least within the range of experimental data available, Mpipi (D=9 K), HPS+cation-π(ii) (D=18 K) and TSCL-M2 (D=19 K) give good quantitative predictions for A1-LCD (i.e. high Pearson coefficients and small D values). Lastly, although HPS-Urry achieves good agreement with experiment when considering the Pearson coefficients, its range of predicted critical temperatures is smaller than in experiment, which results in a relatively large root mean squared deviation.

LLPS of other proteins and multiphasic compartmentalisation

To probe the model's transferability, we test its performance for other well-studied IDRs of RNA-binding proteins. Specifically, we compute phase diagrams for the prion-like domain (PLD) of FUS protein, three variants of the arginine/glycinerich (RGG) domain of LAF-1 (Fig. 6a), and four variants of the DDX4 NTD (Fig. 6b). We also compute phase diagrams for five variants of the full-length FUS protein (SI Sec. S2; Fig. 6c). For all these systems, Mpipi achieves good qualitative agreement with experiment and in some cases achieves quantitative accuracy as well. We provide a full account of these results in SI Sec. S7.

Finally, beyond predicting critical temperatures, achieving the correct balance of interactions is essential to recapitulate more complex condensate behaviours. Inside cells, condensates are multicomponent systems and can have complex molecular architectures that are meaningful to their functions (e.g. the nucleolus), and the variance in the chemical makeup of biological phase-separating systems can give rise to multilayered architectures [49]. For example, Fisher and Elbaum-Garfinkle [6] recently demonstrated that charge-matched mixtures of polyarginine (polyR), poly-lysine (polyK) and polynucleotides formed multiphasic droplets in which arginine is positioned towards the centre of the condensates and lysine is concentrated at the interface.

To investigate this behaviour in simulations, we extend Mpipi to include parameters for RNA (Methods and SI Fig. S4) and study the phase-separation behaviour of a mixture of polyK, polyR and polyU RNA. Consistent with experiments, our simulations recapitulate multiphase droplet architectures (Fig. 6d). Interestingly, we find that the density of the Argrich region at the droplet core is significantly higher than the Lys-rich phase density towards the interface; these results also agree well with experiment [6]. For comparison, we also simulate this mixture using the extended HPS-KR model which includes parameters for RNA [50]; however, using this model, multiphase droplets are not stabilised (SI Fig. S8a). This result is not especially surprising since, as noted above, the balance between Arg and Lys interactions in HPS-KR is incorrect [29, 40].

More broadly, we speculate that it is generally difficult to account for multiphasic behaviour if the standard arithmetic mixing rules for interaction strengths are utilised. Collectively, our work suggests that it is not sufficient to obtain the correct trends for short-ranged pairwise terms: one must also achieve the right balance of these interaction parameters to yield quantitative accuracy.

Discussion

While the balance of interactions in our current model is in agreement with several experimental and computational studies, there is conflicting evidence on the exact ordering of certain key interactions. For example, Bremer and colleagues [10] report weaker Arg–π contacts than the corresponding π–π ones, while our work and the work of Wang et al. [11] 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 these interactions [7], 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 π–π contacts. Hence, we postulate that 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 good agreement with experiment in some cases. Consequently, one set of measurements, be it experiments or simulations, will be unlikely to yield the complete picture.

A key assumption in our work is that differences in LLPS propensities of biomolecules can be captured via pairwise amino-acid interactions. This approximation allows us to construct a transferable coarse-grained model that can capture several qualitative and quantitative trends for phase-separating systems, especially for those characterised in vitro. However, in crowded intracellular environments, three- and higher-body energy terms may become important; accordingly, co-operative interactions can reshape the phase boundaries of LLPS systems [5, 51]. It is therefore important to consider carefully the contribution of such co-operative interactions in intracellular LLPS systems.

As we discussed above, interaction energies in coarse-grained potentials are in fact effective free energies, and they should in principle depend on temperature. 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 hydrogen bonding, could result in significantly different phase behaviour as the temperature is lowered. This can even result in complete mixing at low temperatures, leading to a lower critical solution temperature or re-entrant phase behaviour, especially in multi-component systems [52–54].

Capturing such effects within computational models can further extend our ability to elucidate the driving forces for intracellular LLPS and to probe the ensuing material properties. The current parameterisation of the Mpipi potential is not able to account for such phase behaviour; as an extension of the current work, an approach similar to that of Dignon and co-workers [55] for the HPS-KR model, involving an explicit temperature dependence of the interaction strengths, could be undertaken to enable successful simulations of a broader range of proteins to be performed.

This work highlights the promise that multiscale coarsegrained models can prove robust in delineating the link between chemical changes in biomolecules and their emergent collective behaviour. In particular, the ability of Mpipi to predict quantitatively both single-molecule radii of gyration (which are computationally inexpensive to determine) and the collective behaviour of proteins and RNA in solution (which is computationally more expensive) makes it a prime candidate for efficiently assisting the design of experiments and for gaining physical insight into LLPS at the microscopic scale. Our approach therefore augments the set of rigorous tools that are narrowing down the gap towards achieving 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 has the potential to help discover the molecular mechanisms underpinning phase separation and to provide biophysical understanding of how biomolecular condensates are formed, sustained and regulated.

Methods

Atomistic PMF calculations

To quantify the relative contributions of different types of interactions at physiological salt, we perform atomistic potential-of-mean-force (PMF) computations for a subset of residue pairs, namely WW, YY, FF, RY, RF, KY, KF, AA, SS, PP, RE, RD, KE, KD. All residue pairs from our previous work [7] were recomputed at 150 mM salt concentration, and we have included additional pairs. We also perform PMF calculations for four RNA dimer pairs (SI Sec. S4).

Preparation of structures—Amino acids and nucleic acids are modelled using the AMBER ff03ws force field [56]. This force field is well-suited for probing protein-protein interactions. For modelling the solvent (water) and ions, we use the JC-SPC/E-ion/TIP4P/2005 force field [57], as in our previous work [7]. The N- and C-terminal ends of each amino acid are capped with acetyl and N-methyl capping groups, respectively. Pairs of amino acids are orientated with their side-chains facing each other, based on the most common arrangements observed in protein structures. In cases where the interaction preference is uncertain, multiple arrangements are tested to determine the strongest interaction mode.

Each dimer is then immersed in a cubic box containing TIP4P/2005 water molecules (ca. 960–11,020 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 ~ 150 mM, as well as to produce charge-neutral systems. The resulting systems are then minimised (force tolerance = 500 J mol^{-1} pm^{-1}), with positional restraints of 200 J mol^{-1} pm^{-2} applied in each dimension to all heavy atoms.

Umbrella sampling—The interaction between each dimer is probed with umbrella sampling. For production runs, positional restraints of 1 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 force constant 6 J mol^{-1} pm^{-2}). All bonds with hydrogens are constrained using the

LINCS algorithm, permitting an integration time step of 2 fs. Periodic boundary conditions were used during molecular dynamics (MD) simulations. Electrostatics are computed using particle-mesh Ewald summations with a Coulomb cutoff of 0.9 nm. For each umbrella sampling run, approximately 40 windows, spaced at 50 pm from 0.1 nm to 2nm, are used per pair. Each window is 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). Umbrella sampling data is analysed using the weighted histogram analysis method (WHAM)The first 1 ns of simulations is used for equilibration and is not included in the WHAM analysis. Error analysis is performed using the Bayesian bootstrap method. All atomistic simulations and analyses are carried out using the GROMACS simulation package.

Although we focus on COM distances for fixed molecular orientations in PMF calculations, we ultimately map these to Cα–Cα distances of the coarse-grained potential. The effective free energy as expressed with different order parameters may not be the same and depends on the jacobian determinant of the transformation. However, our choice of order parameter cannot affect observable properties of the system, and the two distances are related by a simple linear relationship for a fixed molecular orientation. Provided we use the PMFs in a self-consistent manner, the resulting ratios of interaction strengths should not depend on this choice of order parameter.

Cation–π charge refitting—Cation–π interactions involve significant polarisation of π electron clouds of aromatic side-chains in the proximity of cationic side-chains (i.e. arginine and lysine), especially at physiological salt conditions. There have been many efforts to capture correctly cation–π interactions in atomistic force fields, both with fixed-charge and polarisable force fields (see discussion by Liu and co-workers [58]). Recently, Paloni *et al.* demonstrated that the fixed-charge AMBER 99SB-disp force field was able to account for Arg/Lys–π interactions for the DDX4 NTD [59]. In another study, Liu and colleagues used quantum-mechanical calculations to reparameterise the Lennard-Jones parameters in the CHARMM36 force field to model cation–π pairs [58]. Their modified parameters led to improved descriptions of the selected folded proteins [58], achieving a closer match to experimental crystal structures.

In this work, to model cation–π interactions atomistically, we follow our previous approach [7] and first refit the charges on tyrosine and phenylalanine side chains. Specifically, the dimers (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 is used to refit the side-chain charges of Tyr and Phe to the ESPs from the quantum-mechanical calculations; 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.

Mpipi model

In the Mpipi model, each amino acid or nucleic acid is represented by a single bead, with corresponding mass, molecular diameter (σ), charge (q), and an energy scale reflecting the relative planar π–π contact frequency (ε). We broadly follow the approach of Dignon et al. [28] to compute the potential energy of a given protein or RNA molecule as

$$E_{\text{Mpipi}} = E_{\text{bond}} + E_{\text{elec}} + E_{\text{pair}}. \tag{1}$$

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

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

where the spring constant k is set to 8.03 J mol^{-1} pm^{-2} and r_{i} is the bond length: reference bond lengths, $r_{i,\text{ ref}}$, of 381 pm and 500 pm are used when bond i connects two protein and two RNA beads, respectively. The electrostatic contribution to the potential energy is computed using a Coulomb term with Debye–Hückel electrostatic screening,

$$E_{\text{elec}} = \sum_{i,j} \frac{q_i q_j}{4\pi\varepsilon_{\text{r}}\varepsilon_0 r_{ij}} \exp\left(-\kappa r_{ij}\right), \tag{3}$$

where $\varepsilon_{\text{r}} = 80$ is the relative dielectric constant of water, ε_0 is the electric constant and $K^{-1} = 795$ pm is the Debye screening length, corresponding to a monovalent salt concentration of 0.15 M to be consistent with the PMF calculations. We use a Coulomb cutoff of 3.5 nm. The dielectric constant and the Debye length control the range of ionic interactions and determine the relative importance of charges relative to all other interactions. A more careful treatment of electrostatics, perhaps in the spirit of Wessen and co-workers [60], would be an important next step to consider in the development of more accurate potentials.

Finally, the non-bonded interactions between protein/RNA beads are modelled via the Wang–Frenkel (WF) potential [19]. The WF potential between two beads of types i and j a distance r apart is given by

$$\phi_{ij}(r) = \varepsilon_{ij}\alpha_{ij}\left[\left(\frac{\sigma_{ij}}{r}\right)^{2\mu_{ij}} - 1\right]\left[\left(\frac{R_{ij}}{r}\right)^{2\mu_{ij}} - 1\right]^{2\upsilon_{ij}}, \tag{4}$$

where

$$\alpha_{ij} = 2\upsilon_{ij}\left(\frac{R_{ij}}{\sigma_{ij}}\right)^{2\mu_{ij}}\left[\frac{2\upsilon_{ij}+1}{2\upsilon_{ij}\left(\left\{\frac{R_{ij}}{\sigma_{ij}}\right\}^{2\mu_{ij}} - 1\right)}\right]^{2\upsilon_{ij}+1}, \tag{5}$$

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 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 and removes any ambiguities or inconsistencies that may arise from one implementation to the next. For example, Lennard-Jones-based potentials can exhibit significant undesirable finite-size effects as a function of the cutoff distance and subsequent tail corrections [61]. The computational performance of the Wang–Frenkel potential is comparable to the Lennard-Jones potential for the same cutoff; although we have not done this here, if one wished to simulate particularly large systems, the Wang–Frenkel potential's more flexible functional form affords an opportunity 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 etal. [13], determine Ashbaugh–Hatch-styleLennard-Jones interactions following Dignon et al. [28], and from these obtain the initial WF parameters. The steepness of the repulsive region of the potential and the width of the attractions can easily be modulated in this framework by allowing the μ parameter to take values larger than unity. We next adjust the values of ε_{ij} by a suitable multiplicative factor so that the integrals of the well depths of the PMF curves of residue pairs i and j approximate their WF analogues, including any screened charge–charge interaction (SI Fig. S3) if relevant to ensure that the overall interaction energy is correctly taken into account [62]. We provide a full parameter listing in SI Table XI, and a LAMMPS implementation in the supporting data. Although it has been suggested [26, 43] 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 model greater flexibility. There is no a priori reason to assume that coarse-grained interactions between unlike species will be well described by an arithmetic mean of homotypic interactions, and, in particular, we find that the heterotypic interactions of arginine and lysine can be significantly different from the mixing-rule prediction. Parameters for the nucleic acids are determined directly by fitting the respective PMF well depths and widths to the WF framework (see below). Both disordered proteins/regions and RNA are modelled as fully flexible polymers.

Validation simulations use various previously reported models. Mostly, these are based on the functional form introduced in the work of Dignon et al. [28]. The bonded and electrostatic contributions to the potential are given by the same functional form in each case, although with slightly different constants (SI Table X).

Background on the parameterisation of Mpipi

To parameterise the non-bonded short-ranged terms described via the Wang–Frenkel potential, we first determine the relative π–π contact frequencies for amino acids from the work of Vernon et al. [13] (SI Table I), who predict the planar π–π contact frequencies from a survey of approximately 6000 high-resolution structures in the PDB. We utilise these contact frequencies as an initial energy scale for short-ranged interactions in our model.

We then refine this initial energy scale using atomistic PMF calculations, focussing on aromatic π–π (Fig. 2a), cation–π (Fig. 2b) and a subset of non-π-based (Fig. 2c) interactions. Pappu, Mittag and colleagues position aromatic ‘stickers’ as the chief drivers of biomolecular LLPS [4, 8–10]; our recent findings are also consistent with the stickers-and-spacers model, where we predict that aromatic π–π interactions constitute dominant forces in LLPS even at extremely high salt concentrations [7].

In this work, we compute the PMF between YY, FF (Fig. 2a) and WW (SI Fig. S2) at physiological salt concentration (see Methods) and find that, in agreement with the bioinformatics data [13] and experiments [10–12], the relative strength of aromatic π–π interactions increases in the order FF<YY<WW (magenta bars in Fig. 2d and SI Fig. S2). Importantly, we find that aromatic π–π interactions are at least twice as strong as non-π-based interactions (dark yellow bars in Fig. 2d). The latter interactions include non-polar, polar and special residues (e.g. Pro) and are commonly categorised as spacers [4, 8, 11]. Interestingly, Bremer et al. predict that the disparity in spacer–spacer and sticker–sticker residue interaction strengths can be as high as 1:8 [10]. Our fitted spacer-type interactions represent a compromise between the predictions of Bremer and co-workers [10] and those suggested by our PMF calculations [Fig. 2e].

We next concentrate on interactions between basic residues (Arg and Lys in particular) and aromatics. These ‘cation–π’ interactions also make significant contributions to LLPS of biomolecules. In an early bioinformatics survey, Gallivan and Dougherty [63] revealed that Trp was most likely to form cation–π interactions, followed by Tyr and then Phe. Song et al. [64] subsequently used experiments and simulations to demonstrate significantly higher binding strengths for RW interactions compared to RY/F ones, with RY slightly stronger than RF.

Furthermore, recent work by Wang et al. [11] suggests that Arg–Tyr interactions may be stronger drivers of LLPS than Tyr–Tyr contacts; our PMF calculations agree that Arg–Tyr interactions are stronger than Tyr–Tyr. However, whether cation–π interactions are indeed stronger contributors to protein LLPS than π––π interactions remains contested: the work of Bremer et al. [10] and single-residue solubility measurements [65] suggest instead that Tyr–Tyr interactions are stronger than Arg–Tyr contacts. A potential source of error in the relative ordering of interactions in our work might come from the approximate nature of atomistic PMFs simulations and the use of a pairwise energy to describe an interaction that is likely affected by co-operative effects. Reassuringly, despite the differences, in all cases, both cation–π and π–π interactions are significant.

A further important consideration is the balance of Arg–π and Lys–π interactions. The differences between Arg–π and Lys–π contacts were highlighted by Gallivan and Dougherty [63], who reported a higher percentage of Arg–π contacts in protein structures. We recently proposed that Arg–π interactions are best described as hybrid cation-π/π–π, whereas Lys–π contacts represent 'purer' cation–π interactions [7]. This distinction arises largely from the presence of π electrons in the Arg side-chain [6, 7, 13, 16–18], which enable Arg residues to interact much more strongly with n-binding partners than Lys can [6, 7, 15, 16]. The dominance of Arg over Lys in these interactions is also consistent with the less favourable hydration free energy recently reported for Arg versus Lys [16]. An earlier study by Kumar et al. revealed that, whereas Lys–π interactions are more favourable in the gas phase, in solution Arg establishes stronger interactions with aromatic rings than Lys, the latter being dominated by electrostatics and therefore weakened by the surrounding dielectric medium [66]. Collectively, our PMF calculations and previous studies all suggest a preference for Arg–π interactions over Lys–π contacts in biomolecular systems. Accordingly, we reparameterise cation–π interactions so that the relative weights in our model more closely match those suggested by the atomistic simulations (Fig. 2d). A summary of the relative interaction strengths between amino-acid pairs is provided in Fig. 2e. These interaction strengths correspond to the average interaction energy in the high-temperature limit relative to a fixed (albeit arbitrary) energy of zero obtained by numerically integrating Eq. (8). The integration over the energy well in this high-temperature limit enables the relative interaction strength to account, at least approximately, for both enthalpic and entropic contributions.

Direct-coexistence simulations

Proteins/RNA are represented via the Mpipi model (or other residue-level coarse-grained model) and direct-coexistence [20] simulations are used to compute their phase diagrams. In such simulations, the high- and low-density fluid phases are simulated in the same simulation box delimited by an interface.

The target number of copies of the protein are placed in an elongated box, which is initially simulated at high temperature and then cooled down to the desired temperature. Canonical-ensemble simulations are then run at temperatures below the estimated critical temperatures for each system. A relaxation time of 5 ps is typically used for the Langevin thermostat and an integration time step of 10 fs is used for all coarse-grained simulations. Calculations are carried out using LAMMPS. We discuss the effect of finite-size effects [9, 67, 68] below.

Estimation of critical points on phase diagrams—Critical temperatures are estimated using the law of coexistence densities,

$$\left(\rho_{\text{high}}(T) - \rho_{\text{low}}(T)\right)^{3.06} = d(1 - T/T_c), \tag{6}$$

and critical densities are computed by assuming that the law of rectilinear diameters holds,

$$\rho_{\text{high}}(T) + \rho_{\text{low}}(T) = 2\rho_c + 2A(T - T_c), \tag{7}$$

where $\rho_{\text{high}}(T)$, $\rho_{\text{low}}(T)$ and ρ_{c} are the densities of the high-density and low-density phases and the critical density, respectively; T_{c} is the critical temperature and d and A are fitting parameters.

Data analysis

When comparing relative interaction strengths, we compute the average energy in the high-temperature limit, i.e. assuming that the well depth is much smaller than $k_{\text{B}}T$ for each individual interaction. For each pair of amino-acid residues, we compute

$$\varepsilon_{\text{avg}} = \int_{\sigma}^{3\sigma} \phi(r)\mathrm{d}r + \int_{\sigma}^{3.5\text{nm}} E_{\text{elec}}(r)\mathrm{d}r, \tag{8}$$

where $\phi(r)$ is the Wang–Frenkel potential [Eq. (4)] and $E_{\text{elec}}(r)$ is the Coulomb energy [Eq. (3)], if relevant. We then normalise the result by the interaction strength of the RY (Arg–Tyr) pair. We compute the integral in Eq. (8) in one-dimensional space, i.e. not including the $4\pi r^2$ volume element, since we wish to compare these strengths to PMF calculations, where a constrained approach was used. However, including a spherical-polar volume element does not significantly affect the appearance of Fig. 2e.

To assess the agreement between simulation results and experiment for a given observable X (where X is either the critical temperature or the radius of gyration), we compute both a Pearson correlation coefficient (r), which is a measure of deviation from a linear fit to the data and is a good measure of the quality of the ordering of the predictions, and a root mean squared deviation D, whose square we define as

$$D^2 = \frac{1}{n}\sum_{i=1}^{n}\left[X_i(\text{experiment}) - X_i(\text{simulation})\right]^2, \tag{9}$$

where n is the number of data points. D is a measure of the absolute deviation from experimental results.

Radii of gyration computation

We compute single-molecule radii of gyration (R_{g}) for the protein sequences presented in SI Sec. S2.1. Each protein was simulated in a large cubic box (ca. 60 nm × 60 nm × 60 nm). Canonical-ensemble simulations were then performed at 300 K for 5 μs, with a time step of 10 fs. A Langevin thermostat was used, with a relaxation time of 50 ps. R_{g} measurements were made every 100,000 time steps (i.e. 1 ns). The first 1000 fs part of simulation was not used in the estimation of R_g values. Simulations were run using LAMMPS.

Estimation of the coil–globule transition temperature

We have estimated the coil–globule transition temperature T_θ with ABSINTH [69], a continuum solvation all-atom model of proteins. To determine this temperature, we simulated a single protein in a spherical cell with explicit ions, and determined the temperature at which there is a sudden change in the radius of gyration as a function of temperature [24].

Finite-size scaling

Finite-size effects can play a significant role in direct-coexistence simulations [67]. To ascertain that phase-diagram calculations with direct-coexistence simulations yield robust results, we first confirmed that reducing the system size by approximately 30 % yields the same phase diagrams, within error bars, for the hnRNPA1 variants and FUS-LCD as the results reported above. Since there is no difference in the predicted critical temperatures, we hypothesised that finite-size effects are not dominant in our simulations. To test this hypothesis more carefully, we investigated the finite-size scaling behaviour of the FUS-LCD system systematically. In SI Fig. S7, we show two sets of results for this system. We first tested the effect of the size of the cross-sectional area of the interface, starting from a particularly small area of 4 nm × 4 nm, i.e. with box dimensions only just larger than the largest cutoff distance in the interparticle potential of 3.5 nm. SI Fig. S7a shows a considerable spread in individual values, but with the possible exception of the smallest system size, there is no significant difference in the mean density computed across the entire high-density portion of the density profile, which is needed for the phase diagram. In other words, a cross-sectional area of approximately 10 nm × 10 nm used in the majority of phase-diagram calculations appears to be more than sufficient to avoid significant finite-size effects.

Next, we tested the finite-size scaling of the bulk of the system, by keeping a constant cross-sectional area and increasing the length of the long box dimension at a constant density, i.e. by increasing the number of chains in the system. We show these results in SI Fig. S7b. For the very smallest system size considered here, with a z-axis dimension of 11 nm, the density profile is close, but not exactly consistent with that of the larger systems. This is not especially surprising, since the ‘long’ box dimension is only marginally longer than the remaining two, and the interface is considerably more fluxional as a result. However, from the 22 nm simulation onwards, the high-density profile has a flat region that changes in width, but not the mean density, suggesting that finite-size effects are negligible beyond this point. The system sizes used in our phase-diagram calculations are shown in SI Table IX. All sizes are well beyond the point where finite-size effects dominate the system’s behaviour.

One caveat here is that the results for FUS-LCD we show in SI Fig. S7 correspond to a temperature just above 80 % of the critical temperature. The interface naturally becomes less well defined as the critical point is approached, and data points very close to the critical temperature are not usually very robust. However, such data points are not necessary to obtain to estimate the critical temperature from a fit to Eqs (6) and (7).

Implementation of parameters for RNA

We have parameterised an initial set of RNA nucleotide parameters that is compatible with the Mpipi model for proteins. Here, nucleotide–nucleotide interaction strengths are derived by first performing atomistic PMF calculations for homo-dimer pairs (SI Fig. S4a,b). We use dimers instead of monomers since it is more straightforward to study homodimers than single nucleic acid monomers in standard protein/RNA force fields; from these simulations we extract the relative weights of RNA nucleotide–nucleotide interactions. Specifically, we

compute the base–base binding free energies, which encode the short-range pairwise terms in the Mpipi model.

To map the PMFs to our Mpipi model parameters, we first fitted the weighted atomistic interaction energies (i.e. the integral of the PMF curves at 298.15 K; Eq. (8)) to the corresponding Wang–Frenkel weighted interaction energies at the same temperature. This procedure involves performing a linear fit between known WF interaction energies in our model and their atomistic counterparts (i.e. for the set in SI Fig. S2a). The fit parameters were then used to determine the corresponding weights for the RNA nucleotides. Next, using an iterative procedure, we determined the WF parameters (i.e. ε and μ) for each RNA bead that yield the target binding strengths.

Each RNA bead is then described by a unique set of WF parameters and a charge of –0.75 e. Finally, we reduced the ε in the WF part of the potential for the RNA beads until self-assembly for PolyA/PolyG RNA (50 beads; 64 chains) [i.e. nucleotides with stronger base-stacking propensities] was sufficiently destabilised at 200 K. In subsequent work, we aim to refine our RNA parameters (including short-ranged binding strengths, bond constants and angular constants).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Prof. Daan Frenkel for useful comments on the manuscript, Prof. Jeetain Mittal and Dr Gregory L. Dignon for helping us implement the HPS-KR potential in LAMMPS, and Dr Giulio Tesei and Prof. Kresten Landorff-Larsen for helping us debug our implementation of their potential. This project has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme [grant 803326; RC-G]. JAJ is a Junior Research Fellow at King's College. RC-G is an Advanced Fellow of the Winton Programme for the Physics of Sustainability. JRE acknowledges funding from the Oppenheimer Fellowship of the University of Cambridge and the Roger Ekins Fellowship from Emmanuel College. AG is funded by the EPSRC [Doctoral Training Partnership, Grant EP/N509620/1] and the Winton Programme for the Physics of Sustainability. PYC is funded by the University of Cambridge Ernest Oppenheimer Fund and the Winton Programme for the Physics of Sustainability. KOR is funded by the EPSRC [Doctoral Training Partnership, Grant EP/N509620/1]. This work was performed using resources provided by the Cambridge Tier-2 system operated by the University of Cambridge Research Computing Service funded by EPSRC Tier-2 capital grant EP/P020259/1 (RCG, JAJ, AR). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Data Availability

All relevant supporting data are available in the Figshare data repository at doi:10.6084/m9.figshare.16772812 [70]. Source data for Figures 2–6 are available with this manuscript.

The data for this study were generated with the simulation codes aand algorithms outlined in SI Table XIV, using the supporting code [70], alongside standard command-line tools.

Code Availability

LAMMPS input scripts and parameter files are available in the Figshare data repository at doi:10.6084/m9.figshare.16772812 [70].

Europe PMC Funders Author Manuscripts

References

[1]. Hyman AA, Simons K. Beyond oil and water-phase transitions in cells. Science. 2012; 337: 1047–1049. DOI: 10.1126/science.1223728 [PubMed: 22936764]

[2]. Li P, et al. Phase transitions in the assembly of multivalent signalling proteins. Nature. 2012; 483: 336–340. DOI: 10.1038/nature10879 [PubMed: 22398450]

[3]. Alberti S, Dormann D. Liquid-liquid phase separation in disease. Annu Rev Genet. 2019; 53: 171–194. DOI: 10.1146/annurev-genet-112618-043527 [PubMed: 31430179]

[4]. Martin EW, et al. Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. Science. 2020; 367: 694–699. DOI: 10.1126/science.aaw8653 [PubMed: 32029630]

[5]. Choi J-M, Holehouse AS, Pappu RV. Physical principles underlying the complex biology of intracellular phase transitions. Annu Rev Biophys. 2020; 49: 107–133. DOI: 10.1146/annurev-biophys-121219-081629 [PubMed: 32004090]

[6]. Fisher RS, Elbaum-Garfinkle S. Tunable multiphase dynamics of arginine and lysine liquid condensates. Nat Commun. 2020; 11: 4628. doi: 10.1038/s41467-020-18224-y [PubMed: 32934220]

[7]. Krainer G, et al. Reentrant liquid condensate phase of proteins is stabilized by hydrophobic and non-ionic interactions. Nat Commun. 2021; 12: 1085. doi: 10.1038/s41467-021-21181-9 [PubMed: 33597515]

[8]. Harmon TS, Holehouse AS, Rosen MK, Pappu RV. Intrinsically disordered linkers determine the interplay between phase separation and gelation in multivalent proteins. eLife. 2017; 6 e30294 doi: 10.7554/elife.30294 [PubMed: 29091028]

[9]. Choi JM, Dar F, Pappu RV. LASSI: A lattice model for simulating phase transitions of multivalent proteins. PLoS Comput Biol. 2019; 15 e1007028 doi: 10.1371/journal.pcbi.1007028 [PubMed: 31634364]

[10]. Bremer A, et al. Deciphering how naturally occurring sequence features impact the phase behaviors of disordered prion-like domains. bioRxiv. 2021; doi: 10.1101/2021.01.01.425046

[11]. Wang J, et al. A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins. Cell. 2018; 174: 688–699. e16 doi: 10.1016/j.cell.2018.06.006 [PubMed: 29961577]

[12]. Qamar S, et al. FUS phase separation is modulated by a molecular chaperone and methylation of arginine cation-π interactions. Cell. 2018; 173: 720–734. e15 doi: 10.1016/j.cell.2018.03.056 [PubMed: 29677515]

[13]. Vernon RM, et al. Pi–pi contacts are an overlooked protein feature relevant to phase separation. eLife. 2018; 7 e31486 doi: 10.7554/elife.31486 [PubMed: 29424691]

[14]. Brady JP, et al. Structural and hydrodynamic properties of an intrinsically disordered region of a germ cell-specific protein on phase separation. Proc Natl Acad Sci U S A. 2017; 114: E8194–E8203. DOI: 10.1073/pnas.1706197114 [PubMed: 28894006]

[15]. Dubreuil B, Matalon O, Levy ED. Protein abundance biases the amino acid composition of disordered regions to minimize non-functional interactions. J Mol Biol. 2019; 431: 4978–4992. DOI: 10.1016/j.jmb.2019.08.008 [PubMed: 31442477]

[16]. Fossat MJ, Zeng X, Pappu RV. Uncovering differences in hydration free energies and structures for model compound mimics of charged side chains of amino acids. J Phys Chem B. 2021; 125: 4148–4161. DOI: 10.1021/acs.jpcb.1c01073 [PubMed: 33877835]

[17]. Dyson HJ, Wright PE, Scheraga HA. The role of hydrophobic interactions in initiation and propagation of protein folding. Proc Natl Acad Sci U S A. 2006; 103: 13057–13061. DOI: 10.1073/pnas.0605504103 [PubMed: 16916929]

[18]. Andrew CD, et al. Stabilizing interactions between aromatic and basic side chains in α-helical peptides and proteins. Tyrosine effects on helix circular dichroism. J Am Chem Soc. 2002; 124: 12706–12714. DOI: 10.1021/ja027629h [PubMed: 12392418]

[19]. Wang X, Ramírez-Hinestrosa S, Dobnikar J, Frenkel D. The Lennard-Jones potential: when (not) to use it. Phys Chem Chem Phys. 2020; 22: 10624–10633. DOI: 10.1039/c9cp05445f [PubMed: 31681941]

[20]. Opitz A. Molecular dynamics investigation of a free surface of liquid argon. Phys Lett A. 1974; 47: 439–440. DOI: 10.1016/0375-9601(74)90566-0

[21]. Noid WG. Perspective: Coarse-grained models for bio-molecular systems. J Chem Phys. 2013; 139 090901 doi: 10.1063/1.4818908 [PubMed: 24028092]

[22]. Hills RD, Lu L, Voth GA. Multiscale coarse-graining of the protein energy landscape. PLOS Comput Biol. 2010; 6 e1000827 doi: 10.1371/journal.pcbi.1000827 [PubMed: 20585614]

[23]. Ruff KM, Harmon TS, Pappu RV. CAMELOT: A machine learning approach for coarse-grained simulations of aggregation of block-copolymeric protein sequences. J Chem Phys. 2015; 143 243123 doi: 10.1063/1.4935066 [PubMed: 26723608]

[24]. Zeng X, Holehouse AS, Chilkoti A, Mittag T, Pappu RV. Connecting coil-to-globule transitions to full phase diagrams for intrinsically disordered proteins. Biophys J. 2020; 119: 402–418. DOI: 10.1016/j.bpj.2020.06.014 [PubMed: 32619404]

[25]. Latham AP, Zhang B. Consistent force field captures homologue-resolved HP1 phase separation. J Chem Theory Comput. 2021; 17: 3134–3144. DOI: 10.1021/acs.jctc.0c01220 [PubMed: 33826337]

[26]. Dannenhoffer-Lafage T, Best RB. A data-driven hydrophobicity scale for predicting liquid-liquid phase separation of proteins. J Phys Chem B. 2021; 125: 4046–4056. DOI: 10.1021/acs.jpcb.0c11479 [PubMed: 33876938]

[27]. Tesei G, Schulze TK, Crehuet R, Lindorff-Larsen K. Accurate model of liquid–liquid phase behaviour of intrinsically-disordered proteins from optimization of single-chain properties. bioRxiv. 2021; doi: 10.1101/2021.06.23.449550

[28]. Dignon GL, Zheng WW, Kim YC, Best RB, Mittal J. Sequence determinants of protein phase behavior from a coarse-grained model. PLOS Comput Biol. 2018; 14 e1005941 doi: 10.1371/journal.pcbi.1005941 [PubMed: 29364893]

[29]. Regy RM, Thompson J, Kim YC, Mittal J. Improved coarse-grained model for studying sequence dependent phase separation of disordered proteins. Protein Sci. 2021; doi: 10.1002/pro.4094

[30]. Souza PCT, et al. Martini 3: a general purpose force field for coarse-grained molecular dynamics. Nat Methods. 2021; 18: 382–388. DOI: 10.1038/s41592-021-01098-3 [PubMed: 33782607]

[31]. Benayad Z, von Bulow S, Stelzl LS, Hummer G. Simulation of FUS protein condensates with an adapted coarse-grained model. J Chem Theory Comput. 2021; 17: 525–537. DOI: 10.1021/acs.jctc.0c01064 [PubMed: 33307683]

[32]. Reith D, Putz M, Muller-Plathe F. Deriving effective mesoscale potentials from atomistic simulations. J Comput Chem. 2003; 24: 1624–1636. DOI: 10.1002/jcc.10307 [PubMed: 12926006]

[33]. van Hoof B, Markvoort AJ, van Santen RA, Hilbers PA. A novel method for coarse graining of atomistic simulations using Boltzmann inversion. Biophys J. 2011; 100: 309a. doi: 10.1016/j.bpj.2010.12.1888

[34]. Ercolessi F, Adams JB. Interatomic potentials from first-principles calculations: the force-matching method. Europhys Lett. 1994; 26: 583–588. DOI: 10.1209/0295-5075/26/8/005

[35]. Lu L, Dama JF, Voth GA. Fitting coarse-grained distribution functions through an iterative force-matching method. J Chem Phys. 2013; 139 121906 doi: 10.1063/1.4811667 [PubMed: 24089718]

[36]. Izvekov S, Voth GA. A multiscale coarse-graining method for biomolecular systems. J Phys Chem B. 2005; 109: 2469–2473. DOI: 10.1021/jp044629q [PubMed: 16851243]

[37]. Johnson ME, Head-Gordon T, Louis AA. Representability problems for coarse-grained water potentials. J Chem Phys. 2007; 126 144509 doi: 10.1063/1.2715953 [PubMed: 17444725]

[38]. Reinhardt A, Cheng B. Quantum-mechanical exploration of the phase diagram of water. Nat Commun. 2021; 12: 588. doi: 10.1038/s41467-020-20821-w [PubMed: 33500405]

[39]. Wang J, et al. Machine learning of coarse-grained molecular dynamics force fields. ACS Cent Sci. 2019; doi: 10.1021/acscentsci.8b00913

[40]. Das S, Lin Y-H, Vernon RM, Forman-Kay JD, Chan HS. Comparative roles of charge, π and hydrophobic interactions in sequence-dependent phase separation of intrinsically disordered proteins. Proc Natl Acad Sci U S A. 2020; 117: 28795–28805. DOI: 10.1073/pnas.2008122117 [PubMed: 33139563]

[41]. Kim YC, Hummer G. Coarse-grained models for simulations of multiprotein complexes: Application to ubiquitin binding. J Mol Biol. 2008; 375: 1416–1433. DOI: 10.1016/j.jmb.2007.11.063 [PubMed: 18083189]

[42]. Kapcha LH, Rossky PJ. A simple atomic-level hydrophobicity scale reveals protein interfacial structure. J Mol Biol. 2014; 426: 484–498. DOI: 10.1016/j.jmb.2013.09.039 [PubMed: 24120937]

[43]. Li H, Tang C, Wingreen NS. Nature of driving force for protein folding: A result from analyzing the statistical potential. Phys Rev Lett. 1997; 79: 765–768. DOI: 10.1103/physrevlett.79.765

[44]. Urry DW, et al. Hydrophobicity scale for proteins based on inverse temperature transitions. Biopolymers. 1992; 32: 1243–1250. DOI: 10.1002/bip.360320913 [PubMed: 1420991]

[45]. Tejedor AR, Garaizar A, Ramírez J, Espinosa JR. Dual RNA modulation of protein mobility and stability within phase-separated condensates. bioRxiv. 2021; doi: 10.1101/2021.03.05.434111

[46]. Lin Y-H, Chan HS. Phase separation and single-chain compactness of charged disordered proteins are strongly correlated. Biophys J. 2017; 112: 2043–2046. DOI: 10.1016/j.bpj.2017.04.021 [PubMed: 28483149]

[47]. Riback JA, et al. Stress-triggered phase separation is an adaptive, evolutionarily tuned response. Cell. 2017; 168: 1028–1040. e19 doi: 10.1016/j.cell.2017.02.027 [PubMed: 28283059]

[48]. Dignon GL, Zheng W, Best RB, Kim YC, Mittal J. Relation between single-molecule properties and phase behavior of intrinsically disordered proteins. Proc Natl Acad Sci U S A. 2018; 115: 9929–9934. DOI: 10.1073/pnas.1804177115 [PubMed: 30217894]

[49]. Fare CM, Villani A, Drake LE, Shorter J. Higher-order organization of biomolecular condensates. Open Biol. 2021; 11 210137 doi: 10.1098/rsob.210137 [PubMed: 34129784]

[50]. Regy RM, Dignon GL, Zheng W, Kim YC, Mittal J. Sequence dependent phase separation of protein-polynucleotide mixtures elucidated using molecular simulations. Nucleic Acids Res. 2020; 48: 12593–12603. DOI: 10.1093/nar/gkaa1099 [PubMed: 33264400]

[51]. Choi J-M, Hyman AA, Pappu RV. Generalized models for bond percolation transitions of associative polymers. Phys Rev E. 2020; 102 doi: 10.1103/physreve.102.042403

[52]. Zeng X, et al. Design of intrinsically disordered proteins that undergo phase transitions with lower critical solution temperatures. APL Mater. 2021; 9 021119 doi: 10.1063/5.0037438

[53]. Banerjee PR, Milin AN, Moosa MM, Onuchic PL, Deniz AA. Reentrant phase transition drives dynamic substructure formation in ribonucleoprotein droplets. Angew Chem Int Ed. 2017; 56: 11354–11359. DOI: 10.1002/anie.201703191

[54]. Alshareedah I, et al. Interplay between short-range attraction and long-range repulsion controls reentrant liquid condensation of ribonucleoprotein–RNA complexes. J Am Chem Soc. 2019; 141: 14593–14602. DOI: 10.1021/jacs.9b03689 [PubMed: 31437398]

[55]. Dignon GL, Zheng W, Kim YC, Mittal J. Temperature-controlled liquid–liquid phase separation of disordered proteins. ACS Cent Sci. 2019; 5: 821–830. DOI: 10.1021/acscentsci.9b00102 [PubMed: 31139718]

[56]. Best RB, Zheng W, Mittal J. Balanced protein–water interactions improve properties of disordered proteins and non-specific protein association. J Chem Theory Comput. 2014; 10: 5113–5124. DOI: 10.1021/ct500569b [PubMed: 25400522]

[57]. Benavides AL, Aragones JL, Vega C. Consensus on the solubility of NaCl in water from computer simulations using the chemical potential route. J Chem Phys. 2016; 144 124504 doi: 10.1063/1.4943780 [PubMed: 27036458]

[58]. Liu H, Fu H, Shao X, Cai W, Chipot C. Accurate description of cation-π interactions in proteins with a nonpolarizable force field at no additional cost. J Chem Theory Comput. 2020; 16: 6397–6407. DOI: 10.1021/acs.jctc.0c00637 [PubMed: 32852943]

[59]. Paloni M, Bailly R, Ciandrini L, Barducci A. Unraveling molecular interactions in liquid–liquid phase separation of disordered proteins by atomistic simulations. J Phys Chem B. 2020; 124: 9009–9016. DOI: 10.1021/acs.jpcb.0c06288 [PubMed: 32936641]

[60]. Wessén J, Pal T, Das S, Lin Y-H, Chan HS. A simple explicit-solvent model of polyampholyte phase behaviors and its ramifications for dielectric effects in biomolecular condensates. J Phys Chem B. 2021; 125: 4337–4358. DOI: 10.1021/acs.jpcb.1c00954 [PubMed: 33890467]

Europe PMC Funders Author Manuscripts

[61]. Holcomb CD, Clancy P, Zollweg JA. A critical study of the simulation of the liquid-vapour interface of a Lennard-Jones fluid. Mol Phys. 1993; 78: 437–459. DOI: 10.1080/00268979300100321

[62]. Reinhardt A. Phase behavior of empirical potentials of titanium dioxide. J Chem Phys. 2019; 151 064505 doi: 10.1063/1.5115161

[63]. Gallivan JP, Dougherty DA. Cation-π interactions in structural biology. Proc Natl Acad Sci U S A. 1999; 96: 9459–9464. DOI: 10.1073/pnas.96.17.9459 [PubMed: 10449714]

[64]. Song J, Ng SC, Tompa P, Lee KAW, Chan HS. Polycation-π interactions are a driving force for molecular recognition by an intrinsically disordered oncoprotein family. PLOS Comput Biol. 2013; 9 e1003239 doi: 10.1371/journal.pcbi.1003239 [PubMed: 24086122]

[65]. Auton M, Bolen DW. Application of the transfer model to understand how naturally occurring osmolytes affect protein stability. Methods Enzymol. 2007; 428: 397–418. DOI: 10.1016/s0076-6879(07)28023-1 [PubMed: 17875431]

[66]. Kumar K, et al. Cation-π interactions in protein–ligand binding: theory and data-mining reveal different roles for lysine and arginine. Chem Sci. 2018; 9: 2655–2665. DOI: 10.1039/c7sc04905f [PubMed: 29719674]

[67]. Chapela GA, Saville G, Thompson SM, Rowlinson JS. Computer simulation of a gas-liquid surface Part 1. J Chem Soc Faraday Trans. 1977; 273: 1133–1144. DOI: 10.1039/F29777301133

[68]. Nilsson D, Irbäck A. Finite-size scaling analysis of protein droplet formation. Phys Rev E. 2020; 101 022413 doi: 10.1103/PhysRevE.101.022413 [PubMed: 32168715]

[69]. Vitalis A, Pappu RV. ABSINTH: A new continuum solvation model for simulations of polypeptides in aqueous solutions. J Comput Chem. 2009; 30: 673–699. DOI: 10.1002/jcc.21005 [PubMed: 18506808]

[70]. Joseph JA, et al. Code and data for 'Physics-driven coarse-grained model for biomolecular phase separation with near-quantitative accuracy'. 2021; doi: 10.6084/m9.figshare.16772812

[71]. Debye P, Huckel E. Zur Theorie der Elektrolyte. I. Gefri-erpunktserniedrigung und verwandte Erscheinungen. Phys Z. 1923; 24: 185–206.

[72]. Araki K, et al. A small-angle X-ray scattering study of alpha-synuclein from human red blood cells. Sci Rep. 2016; 6 30473 doi: 10.1038/srep30473 [PubMed: 27469540]

[73]. Kjaergaard M, et al. Temperature-dependent structural changes in intrinsically disordered proteins: Formation of α-helices or loss of polyproline II? Protein Sci. 2010; 19: 1555–1564. DOI: 10.1002/pro.435 [PubMed: 20556825]

[74]. Martin EW, et al. Sequence determinants of the conformational properties of an intrinsically disordered protein prior to and upon multisite phosphorylation. J Am Chem Soc. 2016; 138: 15323–15335. DOI: 10.1021/jacs.6b10272 [PubMed: 27807972]

[75]. Fuertes G, et al. Decoupling of size and shape fluctuations in heteropolymeric sequences reconciles discrepancies in SAXS vs. FRET measurements. Proc Natl Acad Sci U S A. 2017; 114: E6342–E6351. DOI: 10.1073/pnas.1704692114 [PubMed: 28716919]

[76]. Mylonas E, et al. Domain conformation of tau protein studied by solution small-angle X-ray scattering. Biochemistry. 2008; 47: 10345–10353. DOI: 10.1021/bi800900d [PubMed: 18771286]

[77]. Wells M, et al. Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. Proc Natl Acad Sci U S A. 2008; 105: 5762–5767. DOI: 10.1073/pnas.0801353105 [PubMed: 18391200]

[78]. Uversky VN, Gillespie JR, Fink AL. Why are “natively unfolded” proteins unstructured under physiologic conditions? Proteins. 2000; 41: 415–427. DOI: 10.1002/1097-0134(20001115)41:3<415::aid-prot130>3.0.co;2-7 [PubMed: 11025552]

[79]. Baul U, Chakraborty D, Mugnai ML, Straub JE, Thirumalai D. Sequence effects on size, shape, and structural heterogeneity in intrinsically disordered proteins. J Phys Chem B. 2019; 123: 3462–3474. DOI: 10.1021/acs.jpcb.9b02575 [PubMed: 30913885]

[80]. Arbesú M, et al. The unique domain forms a fuzzy intramolecular complex in Src family kinases. Structure. 2017; 25: 630–640. e4 doi: 10.1016/j.str.2017.02.011 [PubMed: 28319009]

[81]. Gomes G-NW, et al. Conformational ensembles of an intrinsically disordered protein consistent with NMR, SAXS, and single-molecule FRET. J Am Chem Soc. 2020; 142: 15697–15710. DOI: 10.1021/jacs.0c02088 [PubMed: 32840111]

[82]. Lichtinger SM, Garaizar A, Collepardo-Guevara R, Reinhardt A. Targeted modulation of protein liquid–liquid phase separation by evolution of amino-acid sequence. PLOS Comput Biol. 2021; 17 e1009328 doi: 10.1371/journal.pcbi.1009328 [PubMed: 34428231]

[83]. Rowlinson, JS, Widom, B. Molecular theory of capillarity. Dover; 2013.

[84]. Schuster BS, et al. Identifying sequence perturbations to an intrinsically disordered protein that determine its phase-separation behavior. Proc Natl Acad Sci U S A. 2020; 117: 11421–11431. DOI: 10.1073/pnas.2000223117 [PubMed: 32393642]

[85]. Hub JS, de Groot BL, van der Spoel D. g_wham—A free weighted histogram analysis implementation including robust error and autocorrelation estimates. J Chem Theory Comput. 2010; 6: 3713–3720. DOI: 10.1021/ct100494z

[86]. Portz B, Lee BL, Shorter J. FUS and TDP-43 phases in health and disease. Trends Biochem Sci. 2021; 46: 550–563. DOI: 10.1016/j.tibs.2020.12.005 [PubMed: 33446423]

[87]. Akerlof GC, Oshry HI. The dielectric constant of water at high temperatures and in equilibrium with its vapor. J Am Chem Soc. 1950; 72: 2844–2847. DOI: 10.1021/ja01163a006

[88]. Ashbaugh HS, Hatch HW. Natively unfolded protein stability as a coil-to-globule transition in charge/hydropathy space. J Am Chem Soc. 2008; 130: 9536–9542. DOI: 10.1021/ja802124e [PubMed: 18576630]

[89]. Torrie G, Valleau J. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. J Comput Phys. 1977; 23: 187–199. DOI: 10.1016/0021-9991(77)90121-8

[90]. Kästner J. Umbrella sampling. WIREs Comput Mol Sci. 2011; 1: 932–942. DOI: 10.1002/wcms.66

[91]. Kumar S, Rosenberg JM, Bouzida D, Swendsen RH, Kollman PA. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. J Comput Chem. 1992; 13: 1011–1021. DOI: 10.1002/jcc.540130812

[92]. Bayly CI, Cieplak P, Cornell W, Kollman PA. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. J Phys Chem. 1993; 97: 10269–10280. DOI: 10.1021/j100142a004

[93]. Frisch, MJ; , et al. Gaussian 09. Revision D.01. 2013.

[94]. Vitalis A, Pappu RV. Methods for Monte Carlo simulations of biomacromolecules. Annu Rep Comput Chem. 2009; 5: 49–76. DOI: 10.1016/s1574-1400(09)00503-9 [PubMed: 20428473]

[95]. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM. LINCS: A linear constraint solver for molecular simulations. J Comput Chem. 1997; 18: 1463–1472. DOI: 10.1002/(sici)1096-987x(199709)18:12<1463::aid-jcc4>3.0.co;2-h

[96]. Essmann U, et al. A smooth particle mesh Ewald method. J Chem Phys. 1995; 103: 8577–8593. DOI: 10.1063/1.470117

[97]. Schrödinger. PyMol Molecular Graphics System, version 2.4.2.

[98]. Ladd A, Woodcock L. Triple-point coexistence properties of the Lennard-Jones system. Chem Phys Lett. 1977; 51: 155–159. DOI: 10.1016/0009-2614(77)85375-x

[99]. Fernández RG, Abascal JLF, Vega C. The melting point of ice Ih for common water models calculated from direct coexistence of the solid-liquid interface. J Chem Phys. 2006; 124 144506 doi: 10.1063/1.2183308 [PubMed: 16626213]

[100]. Espinosa JR, Sanz E, Valeriani C, Vega C. On fluid-solid direct coexistence simulations: The pseudo-hard sphere model. J Chem Phys. 2013; 139 144502 doi: 10.1063/1.4823499 [PubMed: 24116630]

[101]. Plimpton S. Fast parallel algorithms for short-range molecular dynamics. J Comput Phys. 1995; 117: 1–19. DOI: 10.1006/jcph.1995.1039

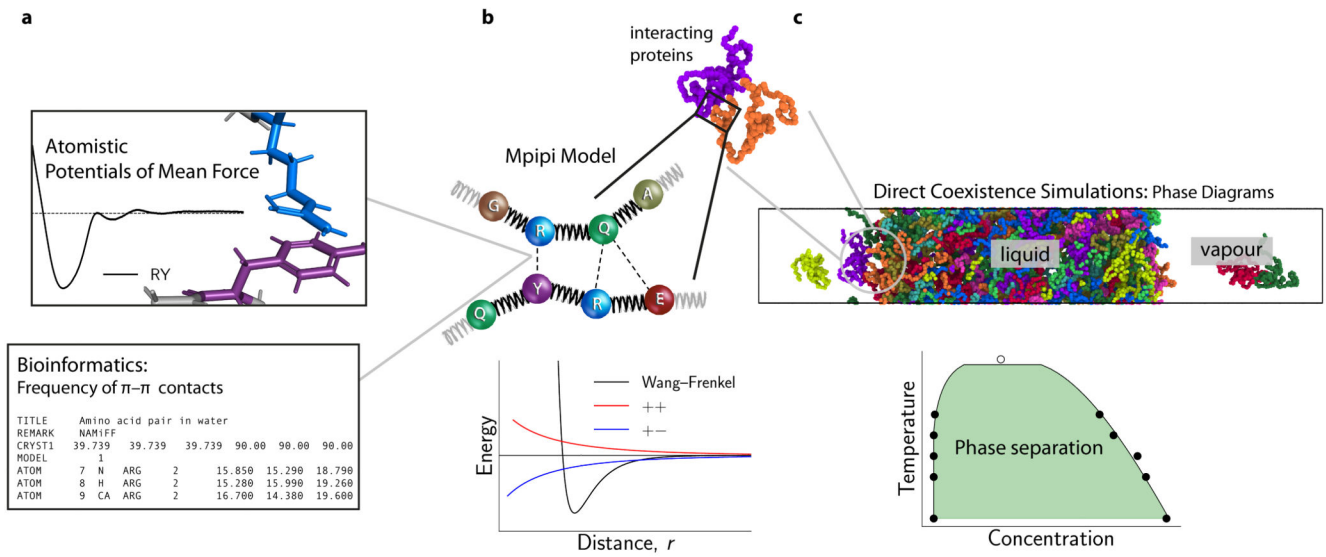

Figure 1. Designing a coarse-grained model for LLPS from potential-of-mean-force calculations and bioinformatics data.

a (Top) Potential of mean force (PMF) of selected amino-acid (or nucleic-acid) pairs are computed in all-atom simulations with explicit solvent and ions. The computed curves provide a free energy of interaction for the pair in question. (Bottom) The frequencies of π–π contacts for amino acids are obtained from bioinformatics work [13]. Together, these data are used to parameterise the pairwise interactions in the Mpipi model. **b** (Top) In the Mpipi model, each amino acid (or nucleic acid) is represented by a unique bead. The potential energy is computed as a sum of short-ranged pairwise terms, electrostatic interactions and bonded interactions modelled as harmonic springs. (Bottom) Short-ranged pairwise and electrostatic interactions are computed via a Coulomb term with Debye–Hückel screening (red and blue curves) and the Wang–Frenkel potential [19] (black curve), respectively. **c** To study biomolecular phase behaviour, we use direct-coexistence molecular-dynamics simulations [20] and compute phase diagrams in the temperature–concentration (or density) space.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

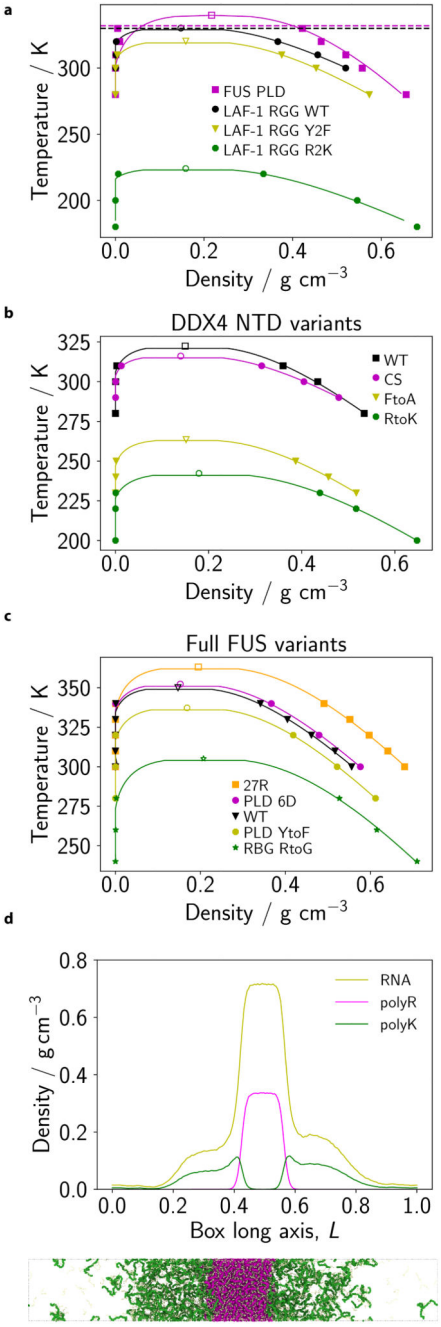

Figure 2. Obtaining the correct balance of π–π and non-π-based interactions in the Mpipi model.

a–c PMF calculations at 150 mM NaCl salt concentration for π–π, cation–π and non-π-based interactions, respectively, as a function of the centre-of-mass (COM) distance. Statistical errors (mean±s.d.) are given as error bands, and are only just larger than the line width. They were computed via Bayesian bootstrapping of 3 independent simulations. Each pair is labelled using one-letter amino-acid codes (SI Table I). **d** Comparison of relative interaction strengths of selected residue pairs (SI Table I) from the PMF calculations with those implemented in the Mpipi model, relative to the Arg–Tyr (RY) interaction.

Values are computed by taking the integral of the curves in **a–c** and the integral of the Wang-Frenkel potential only (between σ and 3σ) for the PMF and Mpipi sets, respectively; for the PMF data only the leftmost well is considered. These correspond to mean energies in the high-temperature limit. **e** Summary of relative interaction strengths in the Mpipi model. These relative interaction strengths include electrostatic interactions and are computed by numerically integrating Eq. (8) and normalising the result by the RY interaction strength.

Europe PMC Funders Author Manuscripts

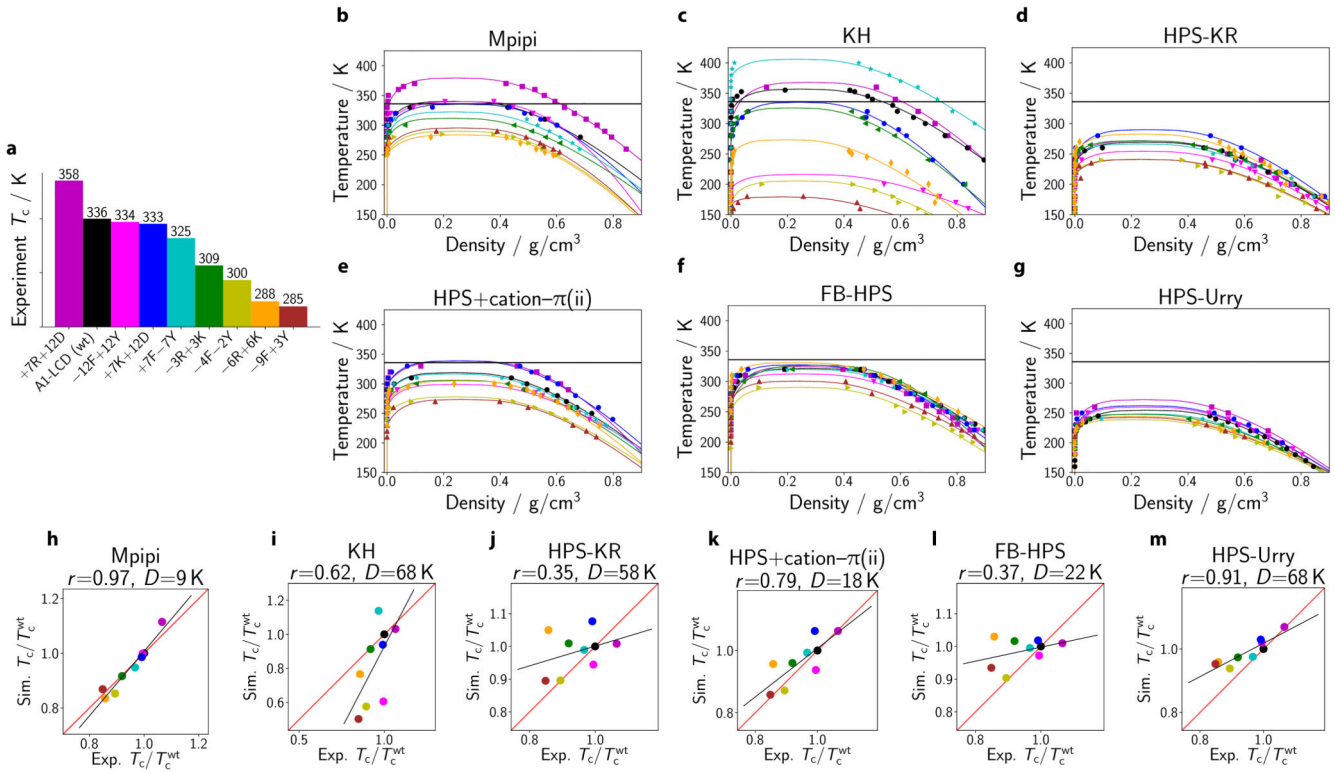

Figure 3. Relative contributions of π–π, cation–π and non-π-based interactions in different residue-level models.

a–f Relative interaction strengths [Eq. (8)] for selected residue pairs (see SI Table I for one-letter amino-acid codes) in Mpipi, KH, HPS-KR, FB-HPS, HPS+cation–π(i) and HPS+cation–π(ii) models. For each model, the data set is normalised 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 coloured in magenta, Arg–π in blue, Lys–π in cyan and non-π-based interactions in dark yellow.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

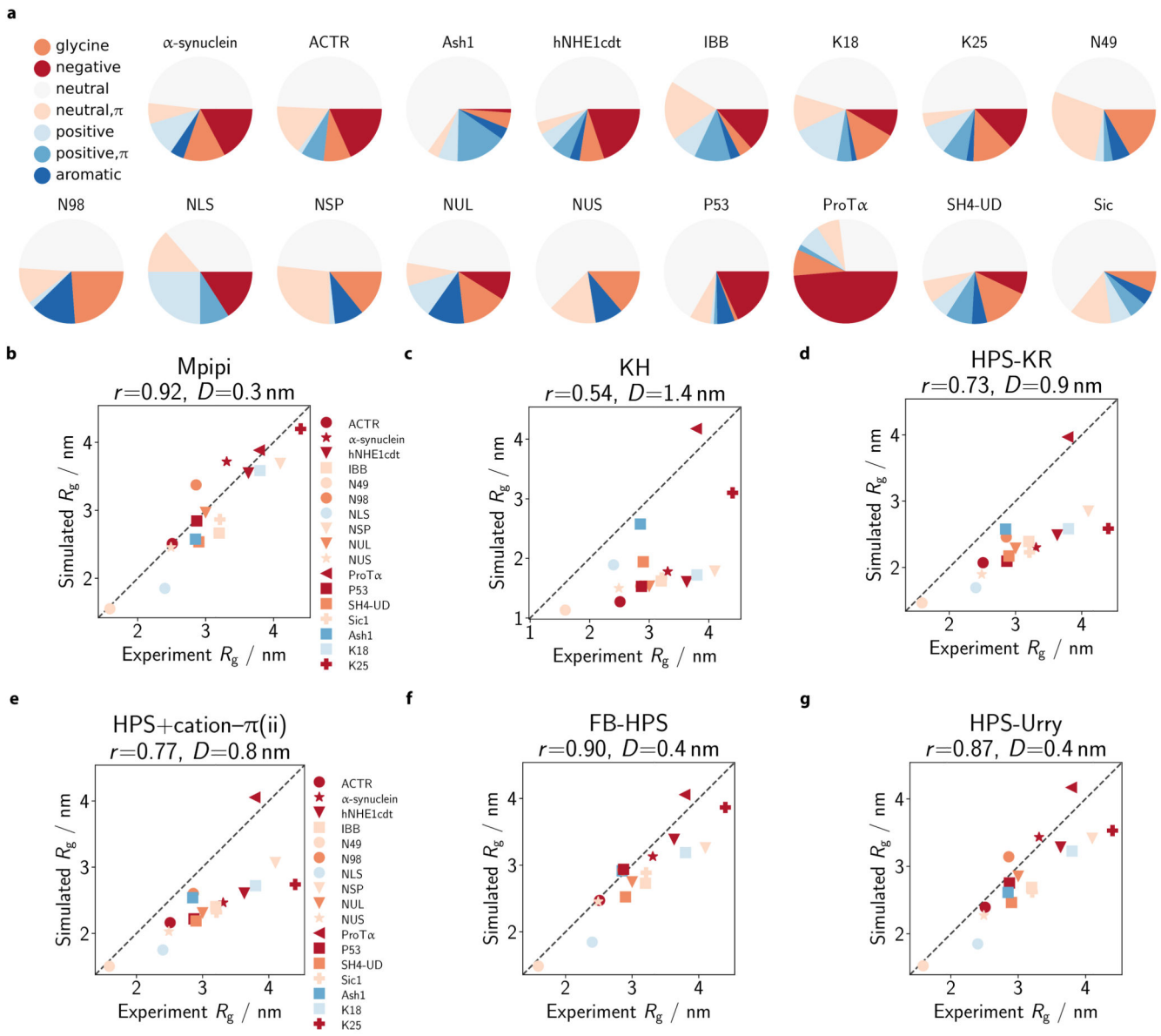

Figure 4. Comparison of single-molecule radii of gyration with experiment.

a Composition of simulated IDPs. We select 17 IDPs for which experimental radii of gyration (R_g) are available (see SI Sec. S2.1 and SI Table III) and assess the composition of the IDPs in terms of the proportion 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), neutral 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) and aromatic (magenta: F, W, Y) residues. **b–g** Comparison of simulated and experiment R_g. R_g values are computed at 300 K in each model. Each protein is coloured based on its dominant residue class (as categorised in **a** and excluding the 'neutral' class). The broken line represents the 'perfect fit' line. For each model, the Pearson correlation coefficient r and the root mean squared deviation D are reported in the respective figure title.

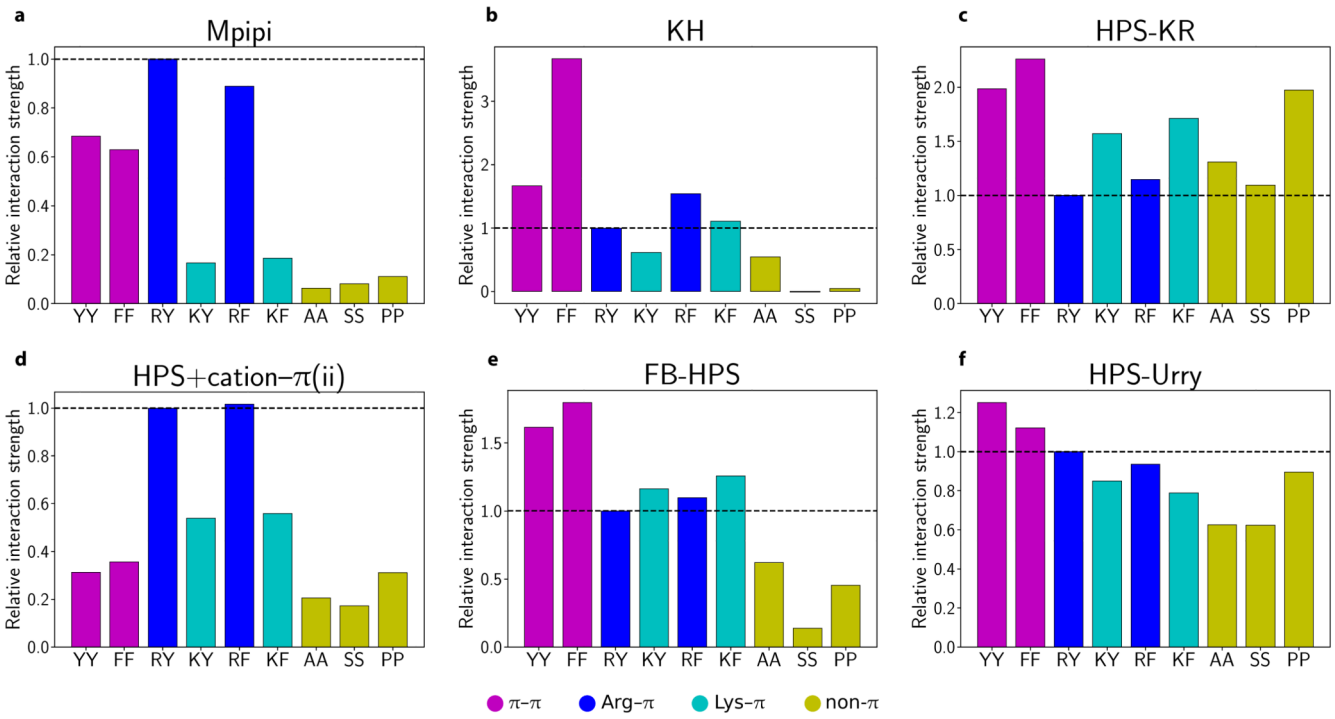

Figure 5. Recapitulating the phase behaviour of A1-LCD variants.

a Nine variants of the A1-LCD (including the wild-type) are studied in this work. Variants are prepared following Bremer et al. [10] Experimental critical temperatures are estimated as described in SI Sec. S2.2. The colour of each variant used in panel **a** is also used in all remaining panels. **b–g** Phase diagrams for A1-LCD variants obtained via direct-coexistence simulations using the Mpipi, KH, HPS-KR, HPS+cation–π(ii), FB-HPS and HPS-Urry models, respectively. Estimation of critical points of simulated phase diagrams is described in the Methods section. Curves are derived from empirical fits of the data to Eqs (6) and (7); typical errors are discussed in SI Sec. S8.4. **h–m** Simulated critical temperature T_c relative to the critical temperature of the wild type (T_c^{wt}) shown against the experimental analogue. The Pearson correlation coefficient r and the root mean squared deviation D are provided above each graph. The red lines correspond to a perfect fit to the experimental data, while the black lines represent the linear regression fit.

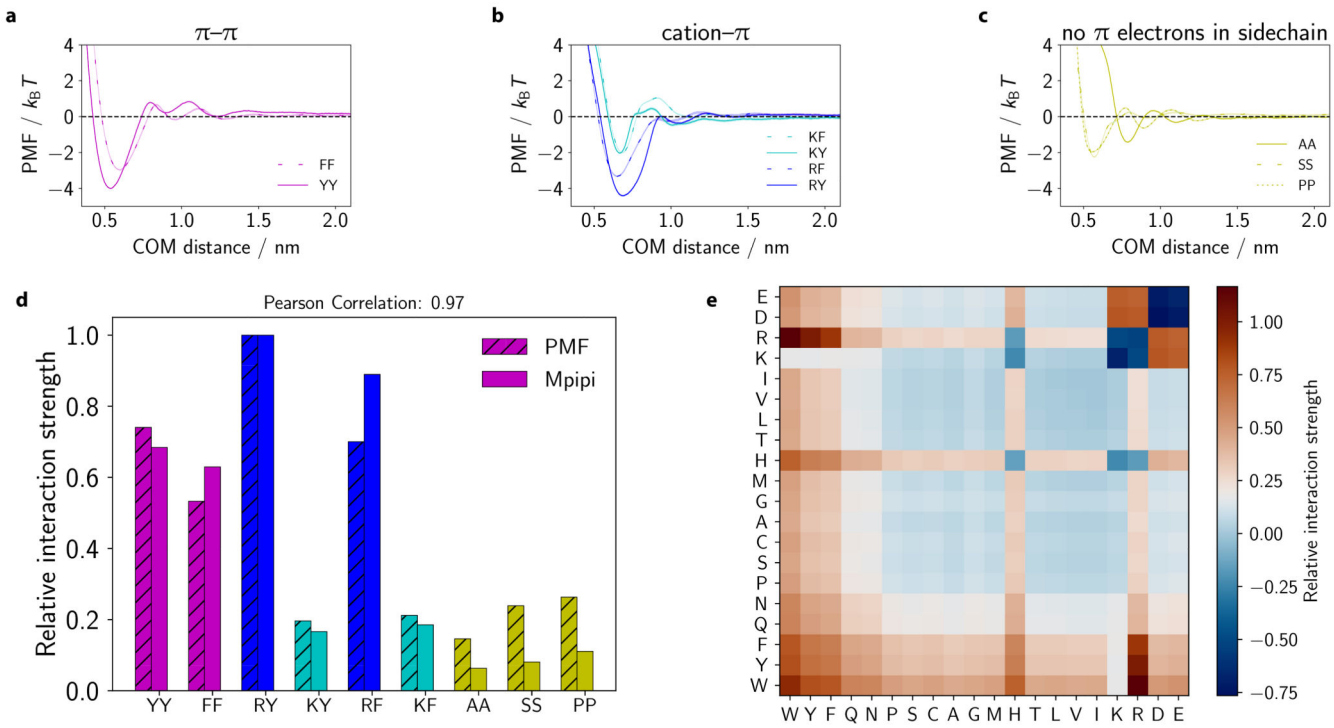

Figure 6. Predicting LLPS propensities of other proteins and multiphasic compartmentalisation.

a Temperature–density phase diagrams for FUS PLD, LAF-1 RGG (WT) and two other variants of LAF-1 RGG for the Mpipi model. Filled symbols represent simulation data, while empty symbols depict estimated simulation critical points (see Methods). The horizontal dashed lines represent estimated T_θ (temperature of the coil-to-globule transition) for FUS PLD (magenta) and LAF-1 RGG (WT) (black) obtained with the Absinth potential. **b, c** Same as in **a,** but for four DDX4 variants and full FUS variants, respectively. **d** We simulate a mixture of PolyK (50 residues; 128 chains), PolyR (50 residues; 128 chains) and RNA (10 residues; 1280 chains) with an extended Mpipi model (see Methods and SI Fig. S4). The density profile along the simulation box's long axis (L; normalised) is given for each mixture component. A simulation snapshot is provided below the density plot. The colour code in the snapshot is consistent with that used in the density plot. The mixture is simulated at $T/T_c \approx 0.8$, where T_c is the critical temperature for liquid-vapour phase separation.