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] and 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*

(Dated: 22 August 2021)

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. 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

* jaj52@cam.ac.uk
† These authors contributed equally to this work
‡ ar732@cam.ac.uk
§ rc597@cam.ac.uk

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]. Coarse-grained 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.

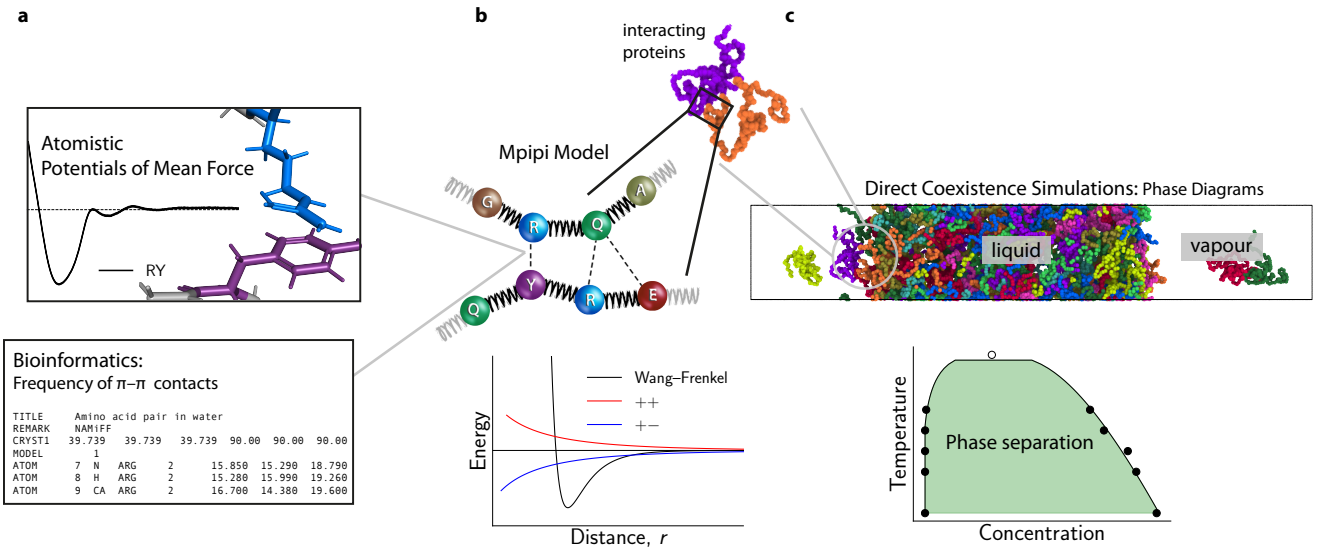

Figure 1. **Designing a coarse-grained model for LLPS from potential-of-mean-force calculations and bioinformatics data**. **a** (Top) Potential of mean force (PMF) of selected amino-acid (or nucleic-acid) pairs are computed in all-atom simulations with explicit solvent and ions. The computed curves 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.

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.

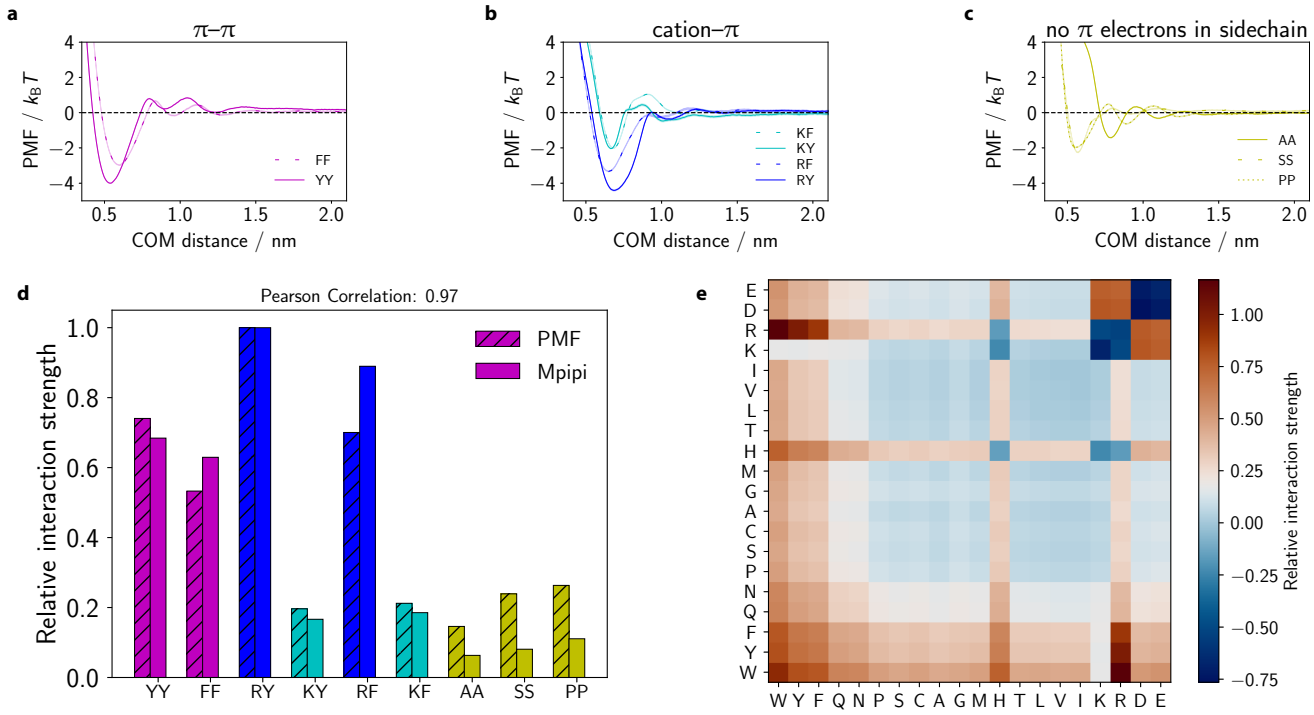

Figure 2. **Obtaining the correct balance of π–π and non-π-based interactions in the Mpipi model. a–c** PMF calculations at 150 mM NaCl salt concentration for π–π, cation–π and non-π-based interactions, respectively, as a function of the centre-of-mass (COM) distance. Statistical errors (mean±s.d.) are 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.

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

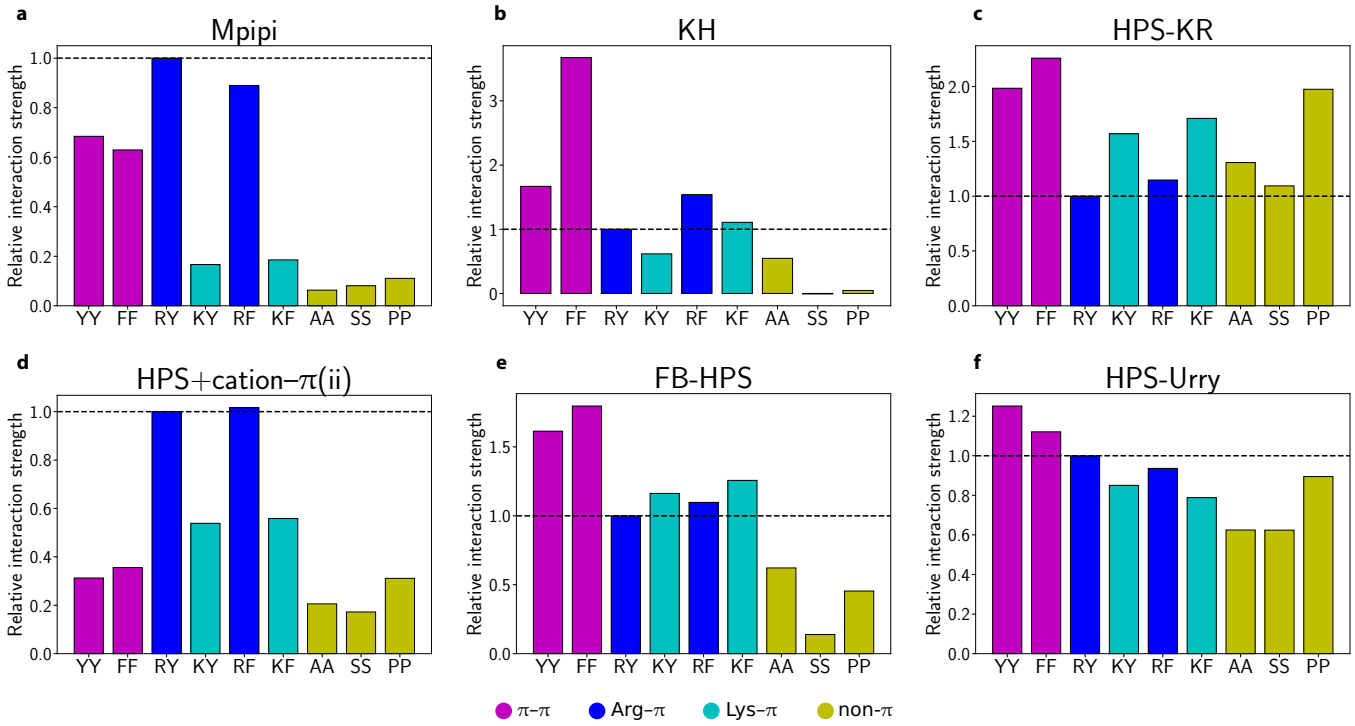

Figure 3. **Relative contributions of π–π, cation–π and non-π-based interactions in different residue-level models. a–f** Relative interaction strengths [Eq. (8)] for selected residue pairs (see SI Table I for one-letter amino-acid codes) in Mpipi, KH, HPS-KR, FB-HPS, HPS+cation–π(i) and HPS+cation–π(ii) 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.

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 Ash1, where e is the elementary charge. These proteins therefore provide an indirect measure of how well electrostatic and short-ranged pairwise

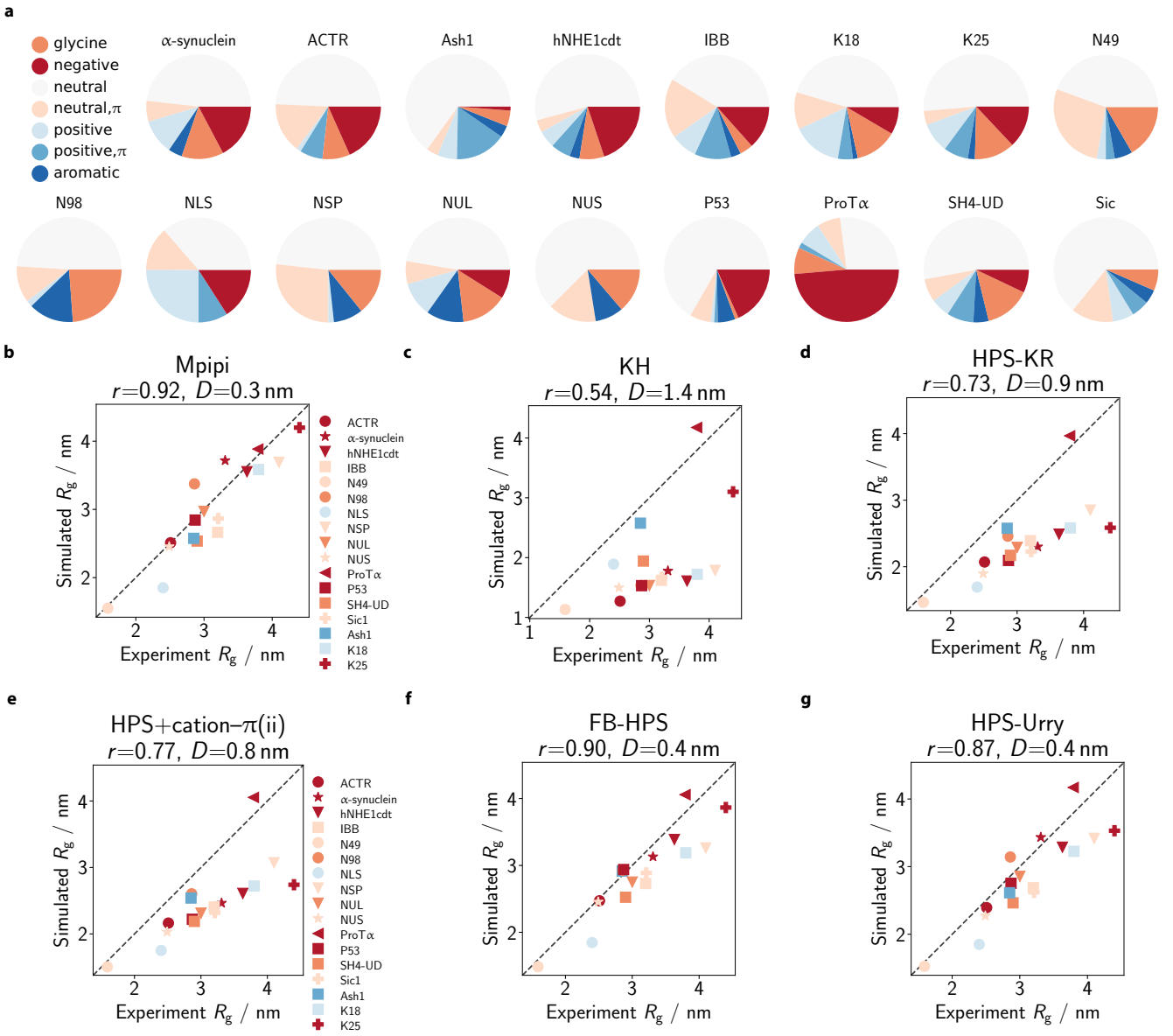

Figure 4. **Comparison of single-molecule radii of gyration with experiment. a** Composition of simulated IDPs. We select 17 IDPs for which experimental radii of gyration (R_g) are available (see SI Sec. S2.1 and SI Table III) and assess the composition of the IDPs in terms of the proportion of glycine (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.

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

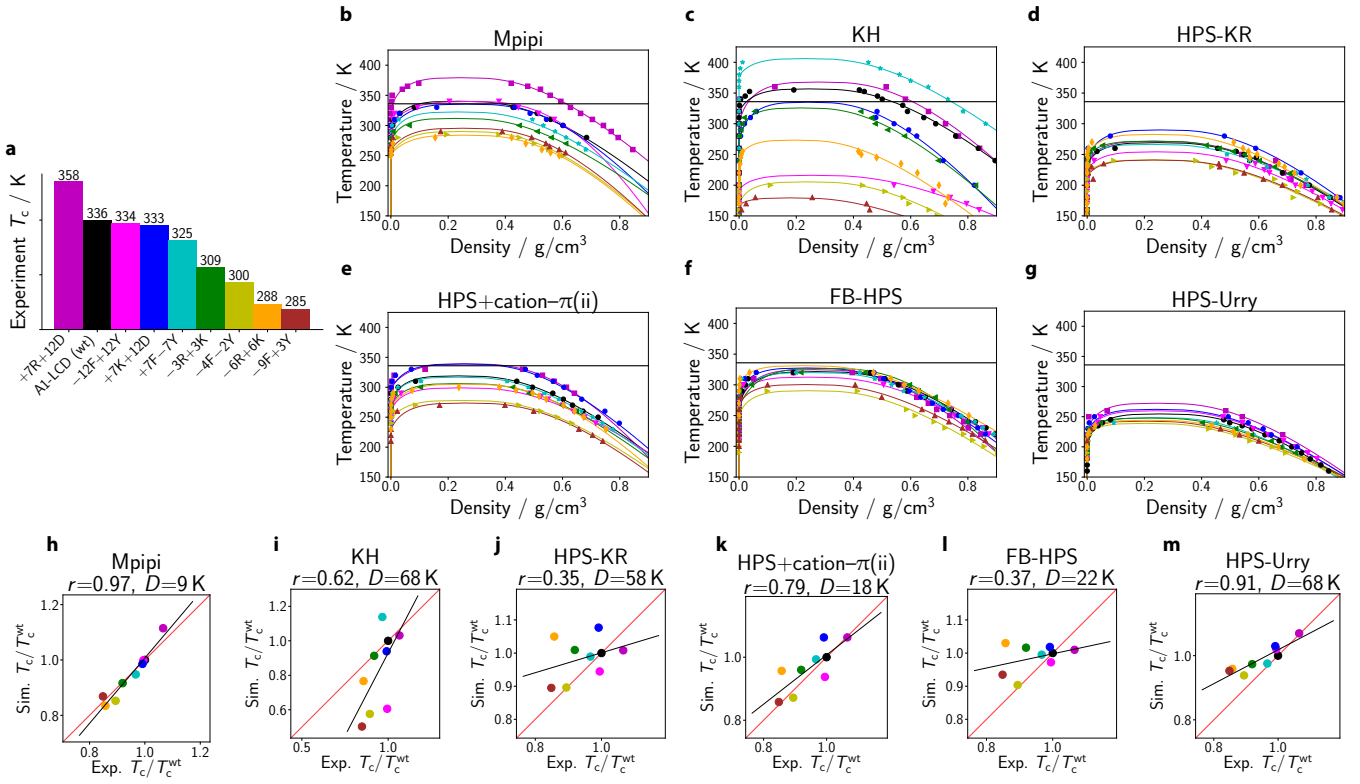

Figure 5. **Recapitulating the phase behaviour of A1-LCD variants. a** Nine variants of the A1-LCD (including the wild-type) are studied in this work. Variants are prepared following Bremer et al. [10] Experimental critical temperatures are estimated as described in SI Sec. S2.2. The colour of each variant 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.

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/glycine-rich (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

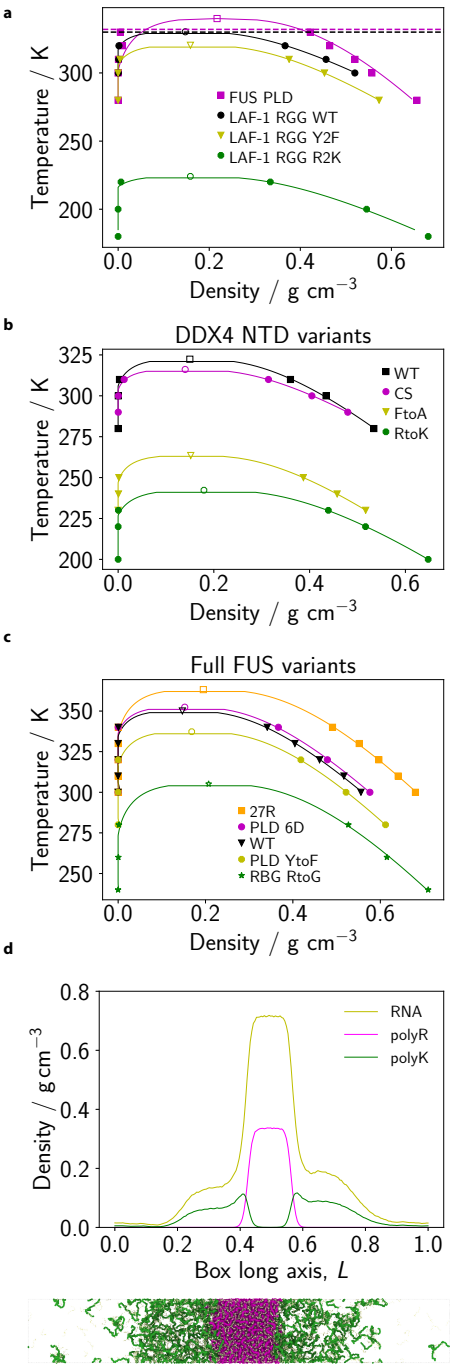

Figure 6. **Predicting LLPS propensities of other proteins and multiphasic compartmentalisation.** **a** Temperature–density phase diagrams for FUS PLD, LAF-1 RGG (WT) and two other variants of LAF-1 RGG for the Mpipi model. Filled symbols represent simulation data, while empty symbols depict estimated 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.

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 Arg-rich 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 multicomponent 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 coarse-grained 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 2 nm, 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

$$E_{\text{Mpipi}} = E_{\text{bond}} + E_{\text{elec}} + E_{\text{pair}}. \quad (1)$$

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

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

where the spring constant k is set to $8.03\,\text{J mol}^{-1}\,\text{pm}^{-2}$ 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,

$$E_{\text{elec}} = \sum_{i,j} \frac{q_i q_j}{4\pi\varepsilon_r\varepsilon_0 r_{ij}} \exp(-\kappa r_{ij}), \quad (3)$$

where $\varepsilon_r = 80$ is the relative dielectric constant of water, ε_0 is the electric constant and $\kappa^{-1} = 795$ 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

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

where

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

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 et al. [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 π-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,

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

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 $k_B T$ for each individual interaction. For each pair of amino-acid residues, we compute

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

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

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

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).

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].

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.

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.

[1] Hyman, A. A. & Simons, K. Beyond oil and water-phase transitions in cells. *Science* **337**, 1047–1049, DOI: 10.1126/science.1223728 (2012).

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

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

[4] Martin, E. W. *et al.* Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science* **367**, 694–699, DOI: 10.1126/science.aaw8653 (2020).

[5] Choi, J.-M., Holehouse, A. S. & Pappu, R. V. Physical principles underlying the complex biology of intracellular phase transitions. *Annu. Rev. Biophys.* **49**, 107–133, DOI: 10.1146/annurev-biophys-121219-081629 (2020).

[6] Fisher, R. S. & Elbaum-Garfinkle, S. Tunable multiphase dynamics of arginine and lysine liquid condensates. *Nat. Commun.* **11**, 4628, DOI: 10.1038/s41467-020-18224-y (2020).

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

[8] Harmon, T. S., Holehouse, A. S., Rosen, M. K. & Pappu, R. V. Intrinsically disordered linkers determine the interplay between phase separation and gelation in multivalent proteins. *eLife* **6**, e30294, DOI: 10.7554/elife.30294 (2017).

[9] Choi, J. M., Dar, F. & Pappu, R. V. LASSI: A lattice model for simulating phase transitions of multivalent proteins. *PLoS Comput. Biol.* **15**, e1007028, DOI: 10.1371/journal.pcbi.1007028 (2019).

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

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

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

[13] Vernon, R. M. *et al.* Pi–pi contacts are an overlooked protein feature relevant to phase separation. *eLife* **7**, e31486, DOI: 10.7554/elife.31486 (2018).

[14] Brady, J. P. *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.* **114**, E8194–E8203, DOI: 10.1073/pnas.1706197114 (2017).

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

[16] Fossat, M. J., Zeng, X. & Pappu, R. V. Uncovering differences in hydration free energies and structures for model compound mimics of charged side chains of amino acids. *J. Phys. Chem. B* **125**, 4148–4161, DOI: 10.1021/acs.jpcb.1c01073 (2021).

[17] Dyson, H. J., Wright, P. E. & Scheraga, H. A. The role of hydrophobic interactions in initiation and propagation of protein folding. *Proc. Natl Acad. Sci. U. S. A.* **103**, 13057–13061, DOI: 10.1073/pnas.0605504103 (2006).

[18] Andrew, C. D. *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.* **124**, 12706–12714, DOI: 10.1021/ja027629h (2002).

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

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

[21] Noid, W. G. Perspective: Coarse-grained models for biomolecular systems. *J. Chem. Phys.* **139**, 090901, DOI: 10.1063/1.4818908 (2013).

[22] Hills, R. D., Lu, L. & Voth, G. A. Multiscale coarse-graining of the protein energy landscape. *PLOS Comput. Biol.* **6**, e1000827, DOI: 10.1371/journal.pcbi.1000827 (2010).

[23] Ruff, K. M., Harmon, T. S. & Pappu, R. V. CAMELOT: A machine learning approach for coarse-grained simulations of aggregation of block-copolymeric protein sequences. *J. Chem. Phys.* **143**, 243123, DOI: 10.1063/1.4935066 (2015).

[24] Zeng, X., Holehouse, A. S., Chilkoti, A., Mittag, T. & Pappu, R. V. Connecting coil-to-globule transitions to full phase diagrams for intrinsically disordered proteins. *Biophys. J.* **119**, 402–418, DOI: 10.1016/j.bpj.2020.06.014 (2020).

[25] Latham, A. P. & Zhang, B. Consistent force field captures homologue-resolved HP1 phase separation. *J. Chem. Theory Comput.* **17**, 3134–3144, DOI: 10.1021/acs.jctc.0c01220 (2021).

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

[27] Tesei, G., Schulze, T. K., Crehuet, R. & Lindorff-Larsen, K. Accurate model of liquid–liquid phase behaviour of intrinsically-disordered proteins from optimization of single-chain properties. *bioRxiv* DOI: 10.1101/2021.06.23.449550 (2021).

[28] Dignon, G. L., Zheng, W. W., Kim, Y. C., Best, R. B. & Mittal, J. Sequence determinants of protein phase behavior from a coarse-grained model. *PLOS Comput. Biol.* **14**, e1005941, DOI: 10.1371/journal.pcbi.1005941 (2018).

[29] Regy, R. M., Thompson, J., Kim, Y. C. & Mittal, J. Improved coarse-grained model for studying sequence dependent phase separation of disordered proteins. *Protein Sci.* DOI: 10.1002/pro.4094 (2021).

[30] Souza, P. C. T. *et al.* Martini 3: a general purpose force field for coarse-grained molecular dynamics. *Nat. Methods* **18**, 382–388, DOI: 10.1038/s41592-021-01098-3 (2021).

[31] Benayad, Z., von Bülow, S., Stelzl, L. S. & Hummer, G. Simulation of FUS protein condensates with an adapted coarse-grained model. *J. Chem. Theory Comput.* **17**, 525–537, DOI: 10.1021/acs.jctc.0c01064 (2021).

[32] Reith, D., Pütz, M. & Müller-Plathe, F. Deriving effective mesoscale potentials from atomistic simulations. *J. Comput. Chem.* **24**, 1624–1636, DOI: 10.1002/jcc.10307 (2003).

[33] van Hoof, B., Markvoort, A. J., van Santen, R. A. & Hilbers, P. A. A novel method for coarse graining of atomistic simulations using Boltzmann inversion. *Biophys. J.* **100**, 309a, DOI: 10.1016/j.bpj.2010.12.1888 (2011).

[34] Ercolessi, F. & Adams, J. B. Interatomic potentials from first-principles calculations: the force-matching method. *Europhys.*

Lett. **26**, 583–588, DOI: 10.1209/0295-5075/26/8/005 (1994).

[35] Lu, L., Dama, J. F. & Voth, G. A. Fitting coarse-grained distribution functions through an iterative force-matching method. *J. Chem. Phys.* **139**, 121906, DOI: 10.1063/1.4811667 (2013).

[36] Izvekov, S. & Voth, G. A. A multiscale coarse-graining method for biomolecular systems. *J. Phys. Chem. B* **109**, 2469–2473, DOI: 10.1021/jp044629q (2005).

[37] Johnson, M. E., Head-Gordon, T. & Louis, A. A. Representability problems for coarse-grained water potentials. *J. Chem. Phys.* **126**, 144509, DOI: 10.1063/1.2715953 (2007).

[38] Reinhardt, A. & Cheng, B. Quantum-mechanical exploration of the phase diagram of water. *Nat. Commun.* **12**, 588, DOI: 10.1038/s41467-020-20821-w (2021).

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

[40] Das, S., Lin, Y.-H., Vernon, R. M., Forman-Kay, J. D. & Chan, H. S. Comparative roles of charge, π, and hydrophobic interactions in sequence-dependent phase separation of intrinsically disordered proteins. *Proc. Natl Acad. Sci. U. S. A.* **117**, 28795–28805, DOI: 10.1073/pnas.2008122117 (2020).

[41] Kim, Y. C. & Hummer, G. Coarse-grained models for simulations of multiprotein complexes: Application to ubiquitin binding. *J. Mol. Biol.* **375**, 1416–1433, DOI: 10.1016/j.jmb.2007.11.063 (2008).

[42] Kapcha, L. H. & Rossky, P. J. A simple atomic-level hydrophobicity scale reveals protein interfacial structure. *J. Mol. Biol.* **426**, 484–498, DOI: 10.1016/j.jmb.2013.09.039 (2014).

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

[44] Urry, D. W. *et al.* Hydrophobicity scale for proteins based on inverse temperature transitions. *Biopolymers* **32**, 1243–1250, DOI: 10.1002/bip.360320913 (1992).

[45] Tejedor, A. R., Garaizar, A., Ramírez, J. & Espinosa, J. R. Dual RNA modulation of protein mobility and stability within phase-separated condensates. *bioRxiv* DOI: 10.1101/2021.03.05.434111 (2021).

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

[47] Riback, J. A. *et al.* Stress-triggered phase separation is an adaptive, evolutionarily tuned response. *Cell* **168**, 1028–1040.e19, DOI: 10.1016/j.cell.2017.02.027 (2017).

[48] Dignon, G. L., Zheng, W., Best, R. B., Kim, Y. C. & Mittal, J. Relation between single-molecule properties and phase behavior of intrinsically disordered proteins. *Proc. Natl Acad. Sci. U. S. A.* **115**, 9929–9934, DOI: 10.1073/pnas.1804177115 (2018).

[49] Fare, C. M., Villani, A., Drake, L. E. & Shorter, J. Higher-order organization of biomolecular condensates. *Open Biol.* **11**, 210137, DOI: 10.1098/rsob.210137 (2021).

[50] Regy, R. M., Dignon, G. L., Zheng, W., Kim, Y. C. & Mittal, J. Sequence dependent phase separation of protein-polynucleotide mixtures elucidated using molecular simulations. *Nucleic Acids Res.* **48**, 12593–12603, DOI: 10.1093/nar/gkaa1099 (2020).

[51] Choi, J.-M., Hyman, A. A. & Pappu, R. V. Generalized models for bond percolation transitions of associative polymers. *Phys. Rev. E* **102**, DOI: 10.1103/physreve.102.042403 (2020).

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

[53] Banerjee, P. R., Milin, A. N., Moosa, M. M., Onuchic, P. L. & Deniz, A. A. Reentrant phase transition drives dynamic substructure formation in ribonucleoprotein droplets. *Angew. Chem., Int. Ed.* **56**, 11354–11359, DOI: 10.1002/anie.201703191 (2017).

[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.* **141**, 14593–14602, DOI: 10.1021/jacs.9b03689 (2019).

[55] Dignon, G. L., Zheng, W., Kim, Y. C. & Mittal, J. Temperature-controlled liquid–liquid phase separation of disordered proteins. *ACS Cent. Sci.* **5**, 821–830, DOI: 10.1021/acscentsci.9b00102 (2019).

[56] Best, R. B., Zheng, W. & Mittal, J. Balanced protein–water interactions improve properties of disordered proteins and non-specific protein association. *J. Chem. Theory Comput.* **10**, 5113–5124, DOI: 10.1021/ct500569b (2014).

[57] Benavides, A. L., Aragones, J. L. & Vega, C. Consensus on the solubility of NaCl in water from computer simulations using the chemical potential route. *J. Chem. Phys.* **144**, 124504, DOI: 10.1063/1.4943780 (2016).

[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.* **16**, 6397–6407, DOI: 10.1021/acs.jctc.0c00637 (2020).

[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* **124**, 9009–9016, DOI: 10.1021/acs.jpcb.0c06288 (2020).

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

[61] Holcomb, C. D., Clancy, P. & Zollweg, J. A. A critical study of the simulation of the liquid-vapour interface of a Lennard-Jones fluid. *Mol. Phys.* **78**, 437–459, DOI: 10.1080/00268979300100321 (1993).

[62] Reinhardt, A. Phase behavior of empirical potentials of titanium dioxide. *J. Chem. Phys.* **151**, 064505, DOI: 10.1063/1.5115161 (2019).

[63] Gallivan, J. P. & Dougherty, D. A. Cation-π interactions in structural biology. *Proc. Natl Acad. Sci. U. S. A.* **96**, 9459–9464, DOI: 10.1073/pnas.96.17.9459 (1999).

[64] Song, J., Ng, S. C., Tompa, P., Lee, K. A. W. & Chan, H. S. Polycation-π interactions are a driving force for molecular recognition by an intrinsically disordered oncoprotein family. *PLOS Comput. Biol.* **9**, e1003239, DOI: 10.1371/journal.pcbi.1003239 (2013).

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

[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.* **9**, 2655–2665, DOI: 10.1039/c7sc04905f (2018).

[67] Chapela, G. A., Saville, G., Thompson, S. M. & Rowlinson, J. S. Computer simulation of a gas–liquid surface. Part 1. *J. Chem. Soc., Faraday Trans. 2* **73**, 1133–1144, DOI: 10.1039/F29777301133 (1977).

[68] Nilsson, D. & Irbäck, A. Finite-size scaling analysis of protein droplet formation. *Phys. Rev. E* **101**, 022413, DOI: 10.1103/PhysRevE.101.022413 (2020).

[69] Vitalis, A. & Pappu, R. V. ABSINTH: A new continuum solvation model for simulations of polypeptides in aqueous solutions. *J. Comput. Chem.* **30**, 673–699, DOI: 10.1002/jcc.21005 (2009).

[70] Joseph, J. A. *et al.* Code and data for 'Physics-driven coarse-grained model for biomolecular phase separation with near-quantitative accuracy', DOI: 10.6084/m9.figshare.16772812 (2021).

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

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

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

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

[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.* **114**, E6342–E6351, DOI: 10.1073/pnas.1704692114 (2017).

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

[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.* **105**, 5762–5767, DOI: 10.1073/pnas.0801353105 (2008).

[78] Uversky, V. N., Gillespie, J. R. & Fink, A. L. Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins* **41**, 415–427, DOI: 10.1002/1097-0134(20001115)41:3<415::aid-prot130>3.0.co;2-7 (2000).

[79] Baul, U., Chakraborty, D., Mugnai, M. L., Straub, J. E. & Thirumalai, D. Sequence effects on size, shape, and structural heterogeneity in intrinsically disordered proteins. *J. Phys. Chem. B* **123**, 3462–3474, DOI: 10.1021/acs.jpcb.9b02575 (2019).

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

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

[82] Lichtinger, S. M., Garaizar, A., Collepardo-Guevara, R. & Reinhardt, A. Targeted modulation of protein liquid–liquid phase separation by evolution of amino-acid sequence. *PLOS Comput. Biol.* **17**, e1009328, DOI: 10.1371/journal.pcbi.1009328 (2021).

[83] Rowlinson, J. S. & Widom, B. *Molecular theory of capillarity* (Dover, 2013).

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

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

[86] Portz, B., Lee, B. L. & Shorter, J. FUS and TDP-43 phases in health and disease. *Trends Biochem. Sci.* **46**, 550–563, DOI: 10.1016/j.tibs.2020.12.005 (2021).

[87] Akerlof, G. C. & Oshry, H. I. The dielectric constant of water at high temperatures and in equilibrium with its vapor. *J. Am. Chem. Soc.* **72**, 2844–2847, DOI: 10.1021/ja01163a006 (1950).

[88] Ashbaugh, H. S. & Hatch, H. W. Natively unfolded protein stability as a coil-to-globule transition in charge/hydropathy space. *J. Am. Chem. Soc.* **130**, 9536–9542, DOI: 10.1021/ja802124e (2008).

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

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

[91] Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H. & Kollman, P. A. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* **13**, 1011–1021, DOI: 10.1002/jcc.540130812 (1992).

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

[93] Frisch, M. J. *et al.* Gaussian 09. Revision D.01 (2013).

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

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

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

[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.* **51**, 155–159, DOI: 10.1016/0009-2614(77)85375-x (1977).

[99] Fernández, R. G., Abascal, J. L. F. & Vega, C. The melting point of ice Ih for common water models calculated from direct coexistence of the solid–liquid interface. *J. Chem. Phys.* **124**, 144506, DOI: 10.1063/1.2183308 (2006).

[100] Espinosa, J. R., Sanz, E., Valeriani, C. & Vega, C. On fluid-solid direct coexistence simulations: The pseudo-hard sphere model. *J. Chem. Phys.* **139**, 144502, DOI: 10.1063/1.4823499 (2013).

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

SUPPLEMENTARY INFORMATION

S1 AMINO-ACID CODES AND BIOINFORMATICS DATA

Table I. The 20 naturally occurring amino acids with their one- and three-letter codes, alongside their charges. In simulations with most (but not all) models considered, the charge of histidine is set to half that of the other non-zero charges [see SI Table X below]. Amino acids marked with a '⋆' are aromatic. The last column corresponds to the planar π–π contact interaction frequencies for each residue, extracted from Fig. 1B of Ref. 13, and rescaled to a range between 0 and 1.

Full name	Code		Charge	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

S2 AMINO-ACID SEQUENCES OF PROTEINS STUDIED

We give below the amino-acid sequences of all the proteins considered, namely those proteins used in simulations of the radius of gyration; the hnRNPA1 intrinsically disordered region and its 8 variants used for validating phase behaviour; LAF-1 RGG and 2 of its variants; and the FUS IDR, using one-letter codes [Table I] for the amino acids.

S2.1 Sequences of proteins used in radius of gyration calculations

α-synuclein
MDVFM KGLSK AKEGV VAAAE KTKQG VAEAA GKTKE
GVLYV GSKTK EGVVH GVATV AEKTK EQVTN VGGAV
VTGVT AVAQK TVEGA GSIAA ATGFV KKDQL GKNEE
GAPQE GILED MPVDP DNEAY EMPSE EGYQD YEPEA

ACTR
GTQNR PLLRN SLDDL VGPPS NLEGQ SDERA LLDQL
HTLLS NTDAT GLEEI DRALG IPELV NQGQA LEPKQ
D

Ash1
GASAS SSPSP STPTK SGKMR SRSSS PVRPK AYTPS
PRSPN YHRFA LDSPP QSPRR SSNSS ITKKG SRRSS
GSSPT RHTTR VCV

hNHE1cdt
MVPAH KLDSP TMSRA RIGSD PLAYE PKEDL PVITI
DPASP QSPES VDLVN EELKG KVLGL SRDPA KVAEE
DEDDD GGIMM RSKET SSPGT DDVFT PAPSD SPSSQ
RIQRC LSDPG PHPEP GEGEP FFPKG Q

IBB
GCTNE NANTP AARLH RFKNK GKDST EMRRR RIEVN
VELRK AKKDD QMLKR RNVSS FPDDA TSPLQ ENRNN
QGTVN WSVDD IVKGI NSSNV ENQLQ AT

K18
MQTAP VPMPD LKNVK SKIGS TENLK HQPGG GKVQI
INKKL DLSNV QSKCG SKDNI KHVPG GGSVQ IVYKP
VDLSK VTSKC GSLGN IHHKP GGGQV EVKSE KLDFK
DRVQS KIGSL DNITH VPGGG NKKIE

K25
MAEPR QEFEV MEDHA GTYGL GDRKD QGGYT MHQDQ
EGDTD AGLKA EEAGI GDTPS LEDEA AGHVT QARMV
SKSKD GTGSD DKKAK GADGK TKIAT PRGAA PPGQK
GQANA TRIPA KTPPA PKTPP SSGEP PKSGD RSGYS
SPGSP GTPGS RSRTP SLPTP PTREP KKVAV VRTPP
KSPSS AKSRL

N49
GCQTS RGLFG NNNTN NINNS SSGMN NASAG LFGSK
P

N98
GCFNK SFGTP FGGGT GGFGT TSTFG QNTGF GTTSG
GAFGT SAFGS SNNTG GLFGN SQTKP GGLFG TSSFS
QPATS TSTGF GFGTS TGTAN TLFGT ASTGT SLFSS
QNNAF AQNKP TGFGN FGTST SSGGL FGTTN TTSNP
FGSTS GSLFG P

NLS
ACETN KRKRE QISTD NEAKM QIQEE KSPKK KRKKR
SSKAN KPPE

NSP
GCNFN TPQQN KTPFS FGTAN NNSNT TNQNS STGAG
AFGTG QSTFG FNNSA PNNTN NANSS ITPAF GSNNT
GNTAF GNSNP TSNVF GSNNS TTNTF GSNSA GTSLF
GSSSA QQTKS NGTAG GNTFG SSSLF NNSTN SNTTK
PAFGG LNFGG GNNTT PSSTG NANTS NNLFG ATANA
N

NUL
GCGFK GFDTS SSSSN SAASS SFKFG VSSSS SGPSQ
TLTST GNFKF GDQGG FKIGV SSDSG SINPM SEGFK
FSKPI GDFKF GVSSE SKPEE VKKDS KNDNF KFGLS
SGLSN PV

NUS
GCPSA SPAFG ANQTP TFGQS QGASQ PNPPG FGSIS
SSTAL FPTGS QPAPP TFGTV SSSSQ PPVFG QQPSQ
SAFGS GTTPN

P53
MEEPQ SDPSV EPPLS QETFS DLWKL LPENN VLSPL
PSQAM DDLML SPDDI EQWFT EDPGP DEAPR MPEAA
PPVAP APAAP TPAAP APAPS WPL

ProTα
MSDAA VDTSS EITTK DLKEK KEVVE EAENG RDAPA
NGNAE NEENG EQEAD NEVDE EEEEG GEEEE EEEEG
DGEEE DGDED EEAES ATGKR AAEDD EDDDV DTKKQ
KTDED D

SH4-UD
MGSNK SKPKD ASQRR RSLEP AENVH GAGGG AFPAS
QTPSK PASAD GHRGP SAAFA PAAAE PKLFG GFNSS
DTVTS PQRAG PLAGG

Sic1
GSMTP STPPR SRGTR YLAQP SGNTS SSALM QGQKT
PQKPS QNLVP VTPST TKSFK NAPLL APPNS NMGMT
SPFNG LTSPQ RSPFP KSSVK RT

Table III. 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. 71 expressed in SI instead of gaussian units).

Protein	R_g / nm	[salt] / mM	κ / nm^{-1}
α-synuclein [72]	3.31	185	1.40
ACTR [73]	2.51	199	1.45
Ash1 [74]	2.85	150	1.26
hNHE1cdt [73]	3.63	199	1.45
IBB [75]	3.20	162	1.31
K18 [76]	3.80	163	1.31
K25 [76]	4.40	163	1.31
N49 [75]	1.59	162	1.31
N98 [75]	2.86	162	1.31
NLS [75]	2.40	162	1.31
NSP [75]	4.10	162	1.31
NUL [75]	3.00	162	1.31
NUS [75]	2.49	162	1.31
P53 [77]	2.87	208	1.49
ProTα [78, 79]	3.79	155	1.28
SH4-UD [80]	2.90	217	1.52
Sic1 [81]	3.21	162	1.31

S2.2 hnRNPA1 variants

The wild-type hnRNPA1-LCD sequence is shown below.

[residues 186–320 of UniProt sequence P09651-2]

hnRNPA1
MASAS SSQRG RSGSG NFGGG RGGGF GGNDN FGRGG
NFSGR GGFGG SRGGG GYGGS GDGYN GFGND GSNFG
GGGSY NDFGN YNNQS SNFGP MKGGN FGGRS SGPYG
GGGQY FAKPR NQGGY GGSSS SSSYG SGRRF

The sequences of the variants of hnRNPA1 we have considered are shown below, using the nomenclature of Bremer and co-workers [10]. The amino-acid residues different from the wild type are highlighted in red. Estimates of their critical temperatures are given in SI Table IV. For hnRNPA1 variants, we recently reported a validation for the KH model similar to that described in the main text [82].

−3R+3K
MASAS SSQRG KSGSG NFGGG RGGGF GGNDN FGRGG
NFSGR GGFGG SKGGG GYGGS GDGYN GFGND GSNFG
GGGSY NDFGN YNNQS SNFGP MKGGN FGGRS SGGSG
GGGQY FAKPR NQGGY GGSSS SSSYG SGRKF

−4F−2Y
MASAS SSQRG RSGSG NSGGG RGGGF GGNDN FGRGG
NSSGR GGFGG SRGGG GYGGS GDGYN GFGND GSNSG
GGGSS NDFGN YNNQS SNFGP MKGGN FGGRS SGGSG
GGGQY SAKPR NQGGY GGSSS SSSSG SGRRF

−6R+6K
MASAS SSQKG KSGSG NFGGG RGGGF GGNDN FGKGG
NFSGR GGFGG SKGGG GYGGS GDGYN GFGND GSNFG
GGGSY NDFGN YNNQS SNFGP MKGGN FGGKS SGGSG
GGGQY FAKPR NQGGY GGSSS SSSYG SGRKF

+7F−7Y
MASAS SSQRG RSGSG NFGGG RGGGF GGNDN FGRGG
NFSGR GGFGG SRGGG GFGGS GDGFN GFGND GSNFG
GGGSF NDFGN FNNQS SNFGP MKGGN FGGRS SGGSG
GGGQF FAKPR NQGGF GGSSS SSSFG SGRRF

+7K+12D
MASAD SSQRD RDDKG NFGDG RGGGF GGNDN FGRGG
NFSDR GGFGG SRGDG KYGGD GDKYN GFGND GKNFG
GGGSY NDFGN YNNQS SNFDP MKGGN FKDRS SGPYD
KGGQY FAKPR NQGGY GGSSS SKSYG SDRRF

+7R+12D
MASAD SSQRD RDDRG NFGDG RGGGF GGNDN FGRGG
NFSDR GGFGG SRGDG RYGGD GDRYN GFGND GRNFG
GGGSY NDFGN YNNQS SNFDP MKGGN FRDRS SGPYD
RGGQY FAKPR NQGGY GGSSS SRSYG SDRRF

−9F+3Y
MASAS SSQRG RSGSG NFGGG RGGGY GGNDN GGRGG
NYSGR GGFGG SRGGG GYGGS GDGYN GGGND GSNYG
GGGSY NDSGN GNNQS SNFGP MKGGN YGGRS SGGSG
GGGQY GAKPR NQGGY GGSSS SSSYG SGRRS

−12F+12Y
MASAS SSQRG RSGSG NYGGG RGGGY GGNDN YGRGG
NYSGR GGYGG SRGGG GYGGS GDGYN GYGND GSNYG
GGGSY NDYGN YNNQS SNYGP MKGGN YGGRS SGGSG
GGGQY YAKPR NQGGY GGSSS SSSYG SGRRY

We have estimated the upper critical solution temperatures of these hnRNPA1 variants from the experimental phase diagrams given by Bremer and co-workers [10]. It is possible in the first instance to estimate critical temperatures by visual inspection in the light of the law of rectilinear diameter [83], which provides an initial crude estimate. To quantify the data more systematically, we fitted the experimental coexistence data points [10] to

$$T_{\mathrm{coex}} = \alpha \frac{\left(\frac{c_{\mathrm{coex}}}{\beta} - \gamma\right)^2 - \left(\frac{c_{\mathrm{coex}}}{\beta} - \gamma\right)}{(r-1)\left(\frac{c_{\mathrm{coex}}}{\beta} - \gamma\right) + 1}, \tag{S1}$$

where c_{coex} is the concentration at coexistence, and α, β, γ and r are fitting parameters. This is a slightly generalised form of the spinodal curve arising from Flory–Huggins–Staverman theory, chosen here solely because the resulting function has the desired shape. With this approach, we can obtain critical temperatures in a systematic way for all variants considered that have data points reported for both the vapour-like and the liquid-like branch. We show two examples of such fitting in

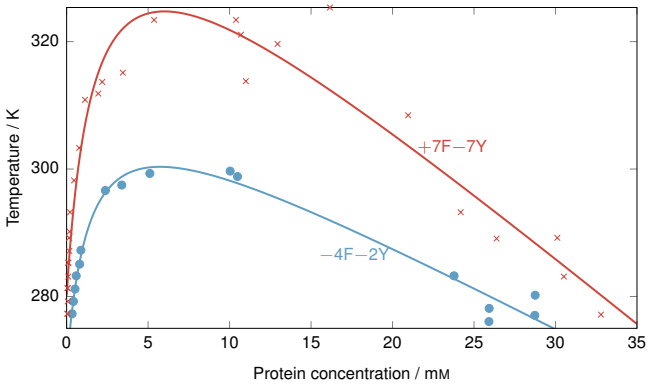

Figure S1. **Estimation of experimental critical temperatures.** We show data points reproduced from the work of Bremer and co-workers [10] alongside fits to Eq. (S1) for two variants of hnRNPA1. The maximum of the fit is taken to correspond to the critical temperature.

Table IV. Experimental upper critical solution temperatures of the hnRNPA1 variants studied, estimated from the phase diagrams reported by Bremer and co-workers [10]. Data points reported to only 3 significant figures are obtained by manually extrapolating the data points and assuming typical binodal behaviour.

Variant	WT	−3R+3K	−4F−2Y	−6R+6K	+7F−7Y	+7K+12D	+7R+12D	−9F+3Y	−12F+12Y
T_c / K	335.9	308.8	300.3	288.0	324.7	333	358	285	334.1

Supplementary Figure S1, and our estimates of the critical temperature in SI Table IV. However, these estimates should be interpreted with a pinch of salt: they are likely to give us the correct ordering, but the error associated with the numerical values is likely to be not insignificant. There were three variants of hnRNPA1 for which insufficient high-density data were reported [10] for the fit to Eq. (S1) to be possible and for which experimental T_c estimates are more approximate [see SI Table IV]; however, even if we remove these variants from further analysis, this does not significantly affect the Pearson coefficients.

S2.3 Additional protein sequences

The sequence of the FUS protein and the variants [11] we have considered is shown below. Changes in sequences are not highlighted for the first three variants since they entail not just single-point mutations, but also additions of residues within the chain that shift the remainder of the sequence.

[UniProt sequence P35637-1]

FUS
MASND YTQQA TQSYG AYPTQ PGQGY SQQSS QPYGQ
QSYSG YSQST DTSGY GQSSY SSYGQ SQNTG YGTQS
TPQGY GSTGG YGSSQ SSQSS YGQQS SYPGY GQQPA
PSSTS GSYGS SSQSS SYGQP QSGSY SQQPS YGGQQ
QSYGQ QQSYN PPQGY GQQNQ YNSSS GGGGG GGGGG
NYGQD QSSMS SGGGS GGGYG NQDQS GGGGS GGYGQ
QDRGG RGRGG SGGGG GGGGG GYNRS SGGYE PRGRG
GGRGG RGGMG GSDRG GFNKF GGPRD QGSRH DSEQD
NSDNN TIFVQ GLGEN VTIES VADYF KQIGI IKTNK
KTGQP MINLY TDRET GKLKG EATVS FDDPP SAKAA
IDWFD GKEFS GNPIK VSFAT RRADF NRGGG NGRGG
RGRGG PMGRG GYGGG GSGGG GRGGF PSGGG GGGGQ
QRAGD WKCPN PTCEN MNFSW RNECN QCKAP KPDGP
GGGPG GSHMG GNYGD DRRGG RGGYD RGGYR GRGGD
RGGFR GGRGG GDRGG FGPGK MDSRG EHRQD RRERP
Y

27R
MASND YTQQA RQSYG AYPTQ PRQGY SQQRS QPYGQ
QSYSG YSQRT DRSGY GQSSY SSYGQ RQNTG YGTQR
TPQGY GSRGG YGSRQ SRQSS YGQQS SYPGY GQQPA
PRSRS GSYGS SRQSS SYGQP QRGSY SQQPS YGGRQ
QSYGQ RQSYN PPQGY GQRNQ YNSSR GRGRG RGRGG
NYGQD QRSMS RGGGR GGGYG NQDQR GGGRS GGYGQ
QASDR GGRGR GGSGG GGGGG GGGYN RSSGG YEPRG
RGGGR GGRGG MGGSD RGGFN KFGGP RDQGS RHDSE
QDNSD NNTIF VQGLG ENVTI ESVAD YFKQI GIIKT
NKKTG QPMIN LYTDR ETGKL KGEAT VSFDD PPSAK
AAIDW FDGKE FSGNP IKVSF ATRRA DFNRG GGNGR
GGRGR GGPMG RGGYG GGGSG GGGRG GFPSG GGGGG
GQQRA GDWKC PNPTC ENMNF SWRNE CNQCK APKPD
GPGGG PGGSH MGGNY GDDRR GGRGG YDRGG YRGRG
GDRGG FRGGR GGGDR GGFGP GKMDS RGEHR QDRRE
R

PLD Y→F
MASND FTQQA TQSFG AFPTQ PGQGF SQQSS QPFGQ
QSFSG FSQST DTSGF GQSSF SSFGQ SQNTG FGTQS
TPQGF GSTGG FGSSQ SSQSS FGQQS SFPGF GQQPA
PSSTS GSFGS SSQSS SFGQP QSGSF SQQPS FGGQQ
QSFGQ QQSFN PPQGF GQQNQ FNSSS GGGGG GGGGG
NFGQD QSSMS SGGGS GGGFG NQDQS GGGGS GGFGQ
QASDR GGRGR GGSGG GGGGG GGGYN RSSGG YEPRG
RGGGR GGRGG MGGSD RGGFN KFGGP RDQGS RHDSE
QDNSD NNTIF VQGLG ENVTI ESVAD YFKQI GIIKT
NKKTG QPMIN LYTDR ETGKL KGEAT VSFDD PPSAK
AAIDW FDGKE FSGNP IKVSF ATRRA DFNRG GGNGR
GGRGR GGPMG RGGYG GGGSG GGGRG GFPSG GGGGG
GQQRA GDWKC PNPTC ENMNF SWRNE CNQCK APKPD
GPGGG PGGSH MGGNY GDDRR GGRGG YDRGG YRGRG
GDRGG FRGGR GGGDR GGFGP GKMDS RGEHR QDRRE
R

RBD R→G
MASND YTQQA TQSYG AYPTQ PGQGY SQQSS QPYGQ
QSYSG YSQST DTSGY GQSSY SSYGQ SQNTG YGTQS
TPQGY GSTGG YGSSQ SSQSS YGQQS SYPGY GQQPA
PSSTS GSYGS SSQSS SYGQP QSGSY SQQPS YGGQQ
QSYGQ QQSYN PPQGY GQQNQ YNSSS GGGGG GGGGG
NYGQD QSSMS SGGGS GGGYG NQDQS GGGGS GGYGQ
QASDG GGGGG GGSGG GGGGG GGGYN RSSGG YEPGG
GGGGG GGGGG MGGSD GGGFN KFGGP RDQGS RHDSE
QDNSD NNTIF VQGLG ENVTI ESVAD YFKQI GIIKT
NKKTG QPMIN LYTDR ETGKL KGEAT VSFDD PPSAK
AAIDW FDGKE FSGNP IKVSF ATRRA DFNGG GGNGG
GGGGG GGPMG GGGYG GGGSG GGGGG GFPSG GGGGG
GQQRA GDWKC PNPTC ENMNF SWRNE CNQCK APKPD
GPGGG PGGSH MGGNY GDDRG GGGGG YDGGG YGGGG
GDGGG FGGGG GGGDG GGFGP GKMDS GGEHR QDRRE
R

PLD 6D
MASND YTQQA TQSYD AYPTQ PGQGY DQQSS QPYDQ
QSYDG YDQST DTSGY DQSSY SSYGQ SQNTG YGTQS
TPQGY GSTGG YGSSQ SSQSS YGQQS SYPGY GQQPA
PSSTS GSYGS SSQSS SYGQP QSGSY SQQPS YGGQQ
QSYGQ QQSYN PPQGY GQQNQ YNSSS GGGGG GGGGG
NYGQD QSSMS SGGGS GGGYG NQDQS GGGGS GGYGQ
QDRGG RGRGG SGGGG GGGGG GYNRS SGGYE PRGRG
GGRGG RGGMG GSDRG GFNKF GGPRD QGSRH DSEQD
NSDNN TIFVQ GLGEN VTIES VADYF KQIGI IKTNK
KTGQP MINLY TDRET GKLKG EATVS FDDPP SAKAA
IDWFD GKEFS GNPIK VSFAT RRADF NRGGG NGRGG
RGRGG PMGRG GYGGG GSGGG GRGGF PSGGG GGGGQ
QRAGD WKCPN PTCEN MNFSW RNECN QCKAP KPDGP

GGGPG GSHMG GNYGD DRRGG RGGYD RGGYR GRGGD
RGGFR GGRGG GDRGG FGPGK MDSRG EHRQD RRERP
Y

The sequence of the FUS PLD region is shown below.

[residues 1–163 of UniProt sequence P35637-1]

FUS PLD MASND YTQQA TQSYG AYPTQ PGQGY SQQSS QPYGQ
QSYSG YSQST DTSGY GQSSY SSYGQ SQNTG YGTQS
TPQGY GSTGG YGSSQ SSQSS YGQQS SYPGY GQQPA
PSSTS GSYGS SSQSS SYGQP QSGSY SQQPS YGGQQ
QSYGQ QQSYN PPQGY GQQNQ YNS

We also considered LAF-1 RGG and two of its variants [84]:

[residues 1–168 of UniProt sequence D0PV95-1]

LAF-1 RGG MESNQ SNNGG SGNAA LNRGG RYVPP HLRGG DGGAA
AAASA GGDDR RGGAG GGGYR RGGGN SGGGG GGGYD
RGYND NRDDR DNRGG SGGYG RDRNY EDRGY NGGGG
GGGNR GYNNN RGGGG GGYNR QDRGD GGSSN FSRGG
YNNRD EGSDN RGSGR SYNND RRDNG GDG

Y→F MESNQ SNNGG SGNAA LNRGG RFVPP HLRGG DGGAA
AAASA GGDDR RGGAG GGGFR RGGGN SGGGG GGGFD
RGFND NRDDR DNRGG SGGFG RDRNW EDRGF NGGGG
GGGNR GFNNN RGGGG GGFNR QDRGD GGSSN FSRGG
FNNRD EGSDN RGSGR SFNND RRDNG GDG

R→K MESNQ SNNGG SGNAA LNKGG KYVPP HLKGG DGGAA
AAASA GGDDK KGGAG GGGYK KGGGN SGGGG GGGYD
KGYND NKDDK DNKGG SGGYG KDKNY EDKGY NGGGG
GGGNK GYNNN KGGGG GGYNK QDKGD GGSSN FSKGG
YNNKD EGSDN KGSGK SYNND KKDNG GDG

Finally, we give the sequences of DDX4-LCD and three of its variants [40]:

[residues 1–236 of UniProt sequence Q9NQI0-1]

DDX4-LCD MGDED WEAEI NPHMS SYVPI FEKDR YSGEN GDNFN
RTPAS SSEMD DGPSR RDHFM KSGFA SGRNF GNRDA
GECNK RDNTS TMGGF GVGKS FGNRG FSNSR FEDGD
SSGFW RESSN DCEDN PTRNR GFSKR GGYRD GNNSE
ASGPY RRGGR GSFRG CRGGF GLGSP NNDLD PDECM
QRTGG LFGSR RPVLS GTGNG DTSQS RSGSG SERGG
YKGLN EEVIT GSGKN SWKSE AEGGE S

CS MGDRD WRAEI NPHMS SYVPI FEKDR YSGEN GRNFN
DTPAS SSEMR DGPSE RDHFM KSGFA SGDNF GNRDA
GKCNE RDNTS TMGGF GVGKS FGNEG FSNSR FERGD
SSGFW RESSN DCRDN PTRND GFSDR GGYEK GNNSE
ASGPY ERGGR GSFDG CRGGF GLGSP NNRLD PRECM
QRTGG LFGSD RPVLS GTGNG DTSQS RSGSG SERGG
YKGLN EKVIT GSGEN SWKSE ARGGE S

F→A MGDED WEAEI NPHMS SYVPI AEKDR YSGEN GDNAN
RTPAS SSEMD DGPSR RDHAM KSGAA SGRNA GNRDA
GECNK RDNTS TMGGA GVGKS AGNRG ASNSR AEDGD
SSGAW RESSN DCEDN PTRNR GASKR GGYRD GNNSE
ASGPY RRGGR GSARG CRGGA GLGSP NNDLD PDECM
QRTGG LAGSR RPVLS GTGNG DTSQS RSGSG SERGG
YKGLN EEVIT GSGKN SWKSE AEGGE S

R→K MGDED WEAEI NPHMS SYVPI FEKDK YSGEN GDNFN
KTPAS SSEMD DGPSK KDHFM KSGFA SGKNF GNKDA
GECNK KDNTS TMGGF GVGKS FGNKG FSNSK FEDGD
SSGFW KESSN DCEDN PTKNK GFSKK GGYKD GNNSE
ASGPY KKGGK GSFKG CKGGF GLGSP NNDLD PDECM
QKTGG LFGSK KPVLS GTGNG DTSQS KSGSG SEKGG
YKGLN EEVIT GSGKN SWKSE AEGGE S

S3 ADDITIONAL PROTEIN POTENTIALS OF MEAN FORCE DATA

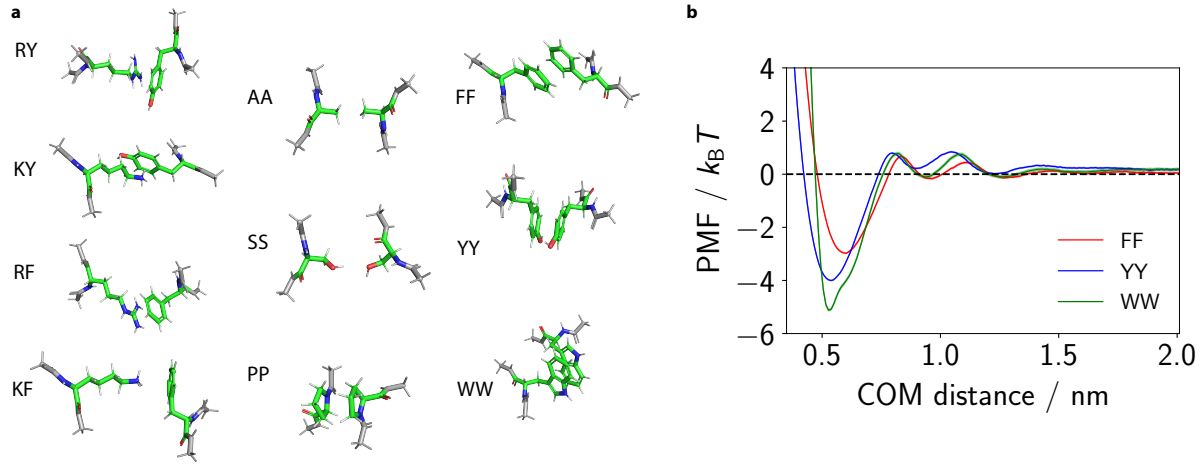

Figure S2. **Additional protein potentials of mean force.** **a** Orientation of amino-acid pairs for PMFs reported in Fig. 2 in the main text. Pairs are labelled using the one-letter codes for the amino acids in question (see SI Table S1). **b** PMFs for amino acids with aromatic side chains, specifically FF, YY and WW. PMFs are computed as described in the main text. Statistical errors (mean±s.d.) are given as error bands; they were computed via Bayesian bootstrapping [85] of 3 independent simulations.

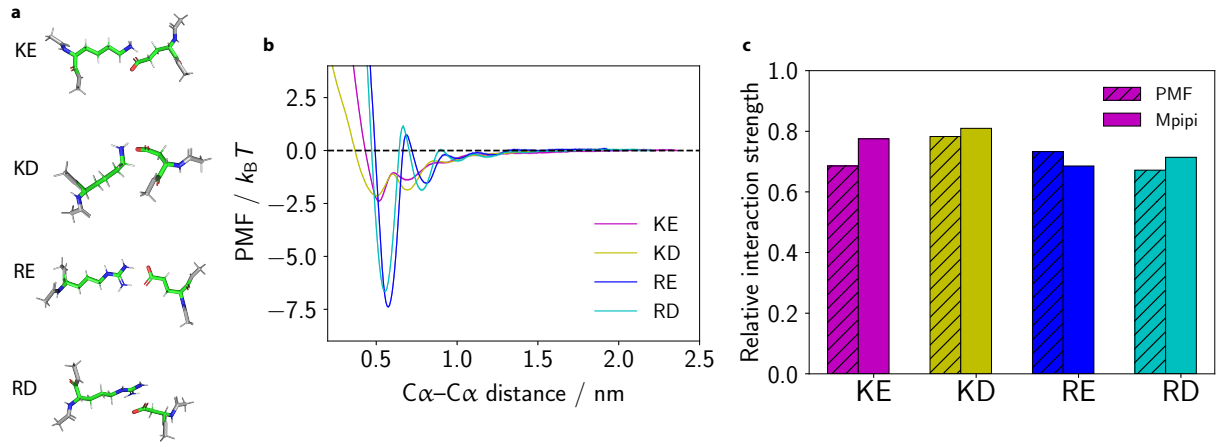

Figure S3. **Comparison of selected attractive charge–charge residue interactions.** **a** Orientation of selected charge–charge amino acid pairs. Pairs are labelled using the one-letter-codes for the amino acids in question (see SI Table S1). **b** PMFs for charge–charge residue pairs, specifically RE, RD, KE and KD. PMFs are computed as described in the main text. Statistical errors (mean±s.d.) are given as error bands; they were computed via Bayesian bootstrapping of 3 independent simulations. **c** Integrated interaction strengths for selected oppositely charged pairs, as implemented in the Mpipi model, normalised based on the RY interaction (as in the main text). The combined potential is fitted so that the ratio of the mean RE/KE and RD/KD strengths closely matches those at the atomistic resolution. When fitting for the Lys-based oppositely charged interactions, we include the first two minima of the PMF since the barrier between them is so small; in the high-temperature limit, these are therefore comparable to the Arg interaction strengths. Differences between Lys and Arg are mainly captured via their short-ranged pairwise contacts and via their homotypic contacts, with Arg parameterised as a stronger sticker and less self-repulsive than Lys (see Fig. 2e in the main text).

S4 RNA POTENTIALS OF MEAN FORCE

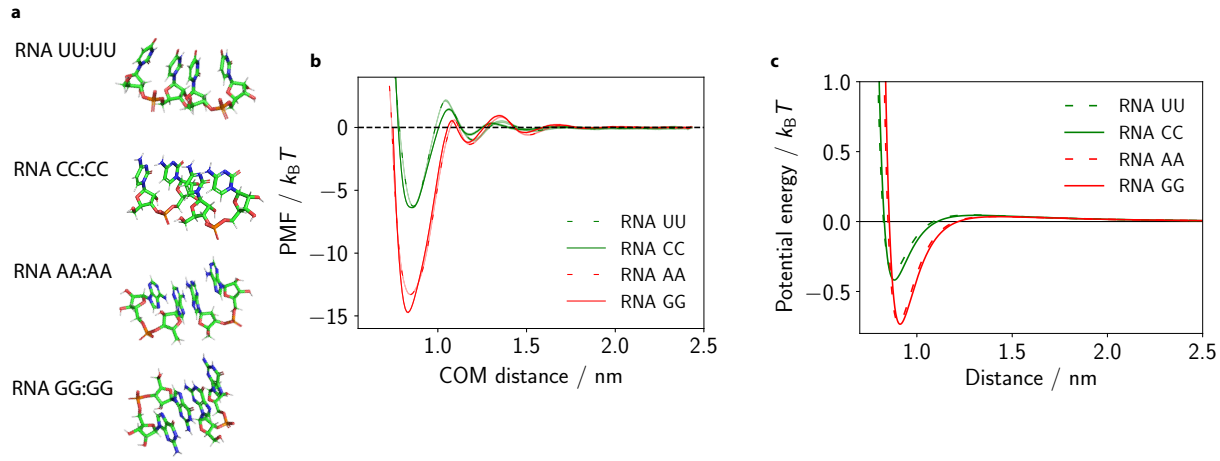

Figure S4. **Comparison of nucleic-acid dimer pair interactions.** **a** Orientation of dimers. Dimer pairs are labelled using the one-letter-codes: adenosine (A), cytidine (C), guanosine (G), uridine (U). **b** PMFs for RNA dimer pairs. PMFs are computed as described in the main text. Statistical errors (mean±s.d.) are given as error bands; they were computed via Bayesian bootstrapping of 3 independent simulations. **c** Pairwise interaction for selected RNA nucleic acid pairs, as implemented in the Mpipi model. Each curve represents the sum of the Wang–Frenkel and Debye–Hückel terms. Typically the atomistic PMF well depths are approximately an order of magnitude greater than the Mpipi model analogues. This disparity is mainly due to the constraints used when computing the PMFs, which allow enhanced sampling of the optimal interaction mode, ignoring other degrees of freedom. The use of explicit solvent in the atomistic calculations versus implicit solvent in the coarse-grained model leads to further differences in the well-depths. Here, the PMF well depths are about 20 times the coarse-grained model, which stems from the aforementioned factors as well as the use of dimer pairs instead of monomer pairs when computing the RNA PMFs.

S5 HPS+CATION–π(i) BENCHMARKS

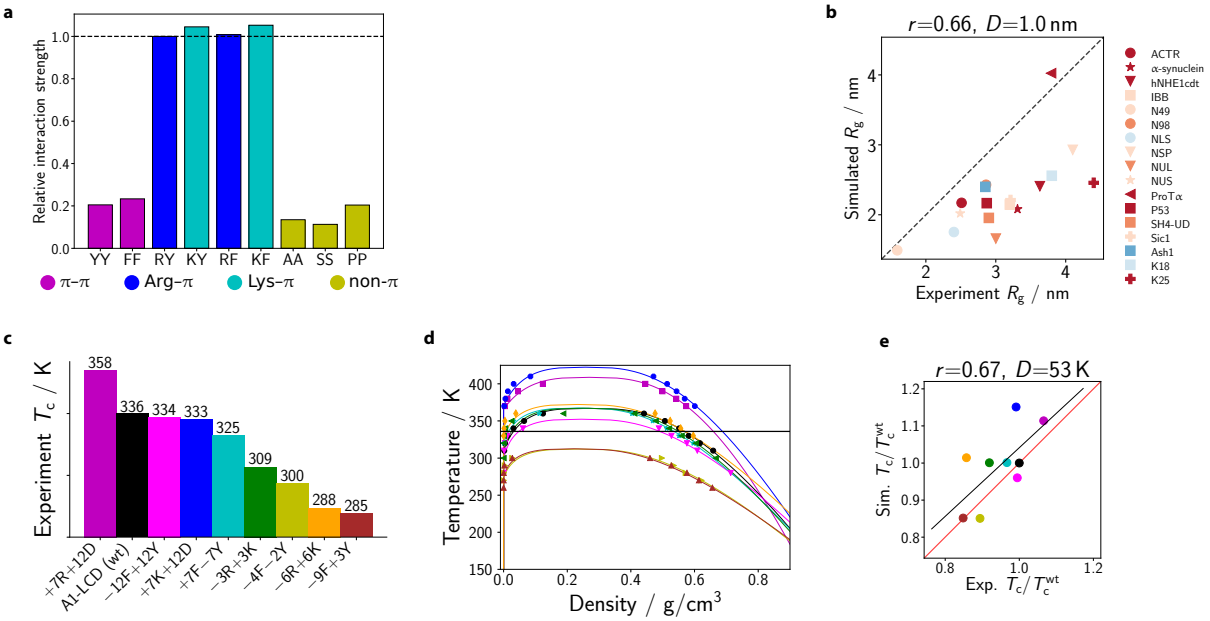

Figure S5. **Data obtained using the HPS+cation–π(i) model.** [40] Descriptions of individual panels are the same as for Supplementary Figure S6 immediately below.

S6 TSCL-M2 BENCHMARKS

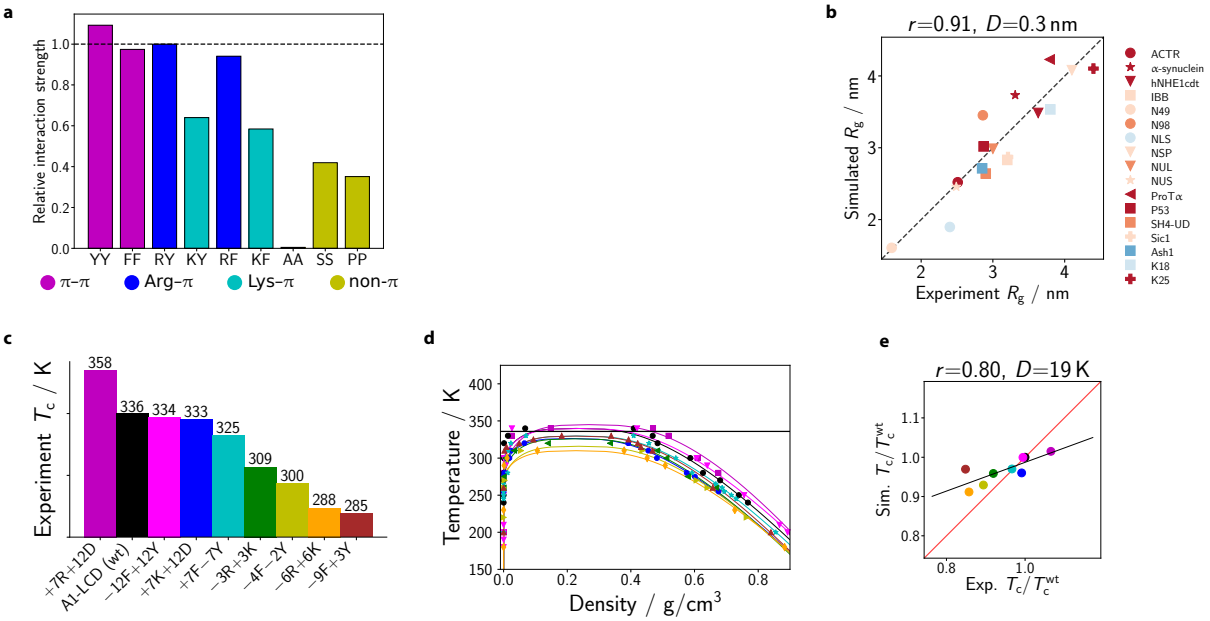

Figure S6. **Data obtained using the M2 parameter set of Tesei et al.** [27] **a** Relative interaction strengths for selected residue pairs, normalised relative to the Arg–Tyr (RY) interaction. A horizontal dashed line at the RY interaction strength is provided for comparison purposes (cf. Fig. 3 of the main text). **b** Comparison of simulated and experimental R_g. Each protein is coloured based on its dominant residue class (as categorised in Fig. 4a of the main text and excluding the 'neutral' class). The dashed line represents a 'perfect fit'. The Pearson correlation coefficient (r) and the root mean squared deviation from the experimental values (D) are reported in the plot title. **c** Experimental critical temperatures of hnRNPA1-LCD variants, reproduced from main text. The colour of each variant used in panel **c** is used in all remaining panels. **d** Phase diagrams for hnRNPA1-LCD variants obtained via direct-coexistence simulations. Estimation of critical points of phase diagrams is described in the main text. Curves are derived from empirical fits of the data to Eqs (6) and (7) of the main text. **e** 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 in the plot title.

S7 LLPS PROPENSITY OF OTHER PROTEINS

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 [84, 86]. In addition, the Absinth (self-Assembly of Biomolecules Studied by an Implicit, Novel, Tunable Hamiltonian) [69] 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 [4, 24]. 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 [84]. The Mpipi model correctly predicts this trend (Fig. 6a of the main text). 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; see black dotted line in Fig. 6a of the main text). We also employed Mpipi to compute the phase diagram for the FUS PLD (magenta curve in Fig. 6a of the main text) and estimated the temperature for single-molecule collapse via Absinth (magenta dotted line in Fig. 6a of the main text). Here, our critical temperature estimate (340 K) is 8 K higher than T_θ (~332 K). Besides an abundance of Tyr residues, about 65 % of FUS PLD is composed on Ser, Gln and Gly residues. As pointed out in the main text, 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 (Fig. 6b of the main text). 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). The LLPS propensities of the DDX4 NTD variants have been thoroughly characterised by Brady and colleagues [14]. They concluded that, although scrambling charges (i.e. the CS variant) reduces 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 [14]. Our computed phase diagrams agree qualitatively with these experimental predictions [14], with T_c decreasing in the order WT>CS≫F→A>R→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 (Fig. 6b of the main text). 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 useful to include a degree of anisotropic character for some interactions.

In addition to the preceding IDRs, we compute phase diagrams for the full-length FUS protein (FUS WT) and four additional FUS variants whose protein sequences are provided in Sec. S2. Wang *et al.* used fluorescence imaging to determine the saturation concentrations for various FUS mutants [11]. Compared to the WT protein, lower saturation concentrations were reported for the 27R and PLD 6D mutants, suggesting that these mutations enhance LLPS propensities. By contrast, the PLD Y→F and RBD R→G mutants both displayed higher saturation concentrations than the WT protein. We computed the critical temperatures for LLPS for each FUS variant and found that the order of T_c is consistent with the in vitro LLPS propensities, i.e. T_c values decrease in the order 27R>PLD 6D>WT>PLD Y→F>RBD R→G (Fig. 6c of the main text).

S8 SUPPLEMENTARY METHODS: FIGURES AND TABLES

S8.1 Finite-size scaling

Table IX. System sizes used in phase-diagram calculations. The box lengths L_x and L_y are always the same. For hnRNPA1 variants, all simulations were performed with two system sizes.

System	Chains	Beads	L_x / nm	L_z / nm
hnRNPA1 + variants	64	8640	11.3	34.0
	63	8505	10.0	44.0
FUS-LCD	52	8476	12.0	44.0
FUS + variants	24	12 624	13.8	76.9
LAF-1 RGG + variants	100	16 800	14.3	70.8
DDX4 + variants	24	5664	16.3	67.3

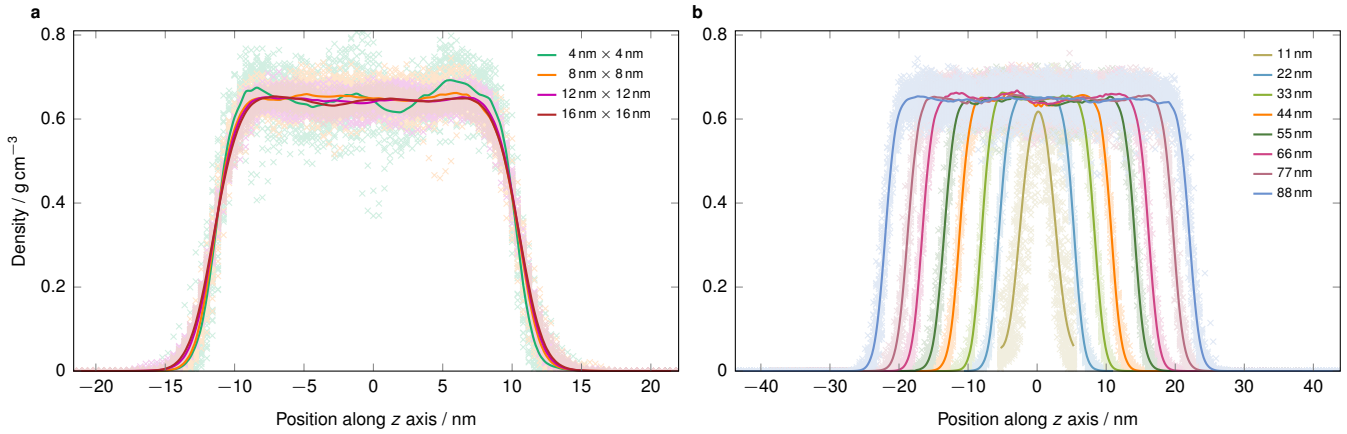

Figure S7. **Finite-size scaling analysis.** Plots of local density against the z-axis co-ordinate for FUS-LCD simulated with the Mpipi potential at 280 K and a constant number density of 1.85 nm^{-3}. The z axis is the longest of the three box dimensions. In panel **a**, the box size is fixed at 44 nm and the cross-sectional area of the interface changes as indicated in the legend. In panel **b**, the cross-sectional area of the interface is fixed at 8 nm × 8 nm, and the length of the box in the z direction changes as indicated in the legend. In each case, we show as pale symbols short-time averages of local densities calculated in 0.4-nm-wide bins along the z axis averaged over 80 000 measurements for each symbol, and repeated across long simulations, resulting in a spread of data points. We also show the mean value of these individual measurements as a solid line. In phase-diagram calculations, an average liquid-like density is then computed across the entire central portion of the density profiles shown.

S8.2 RNA–PolyR–PolyK systems

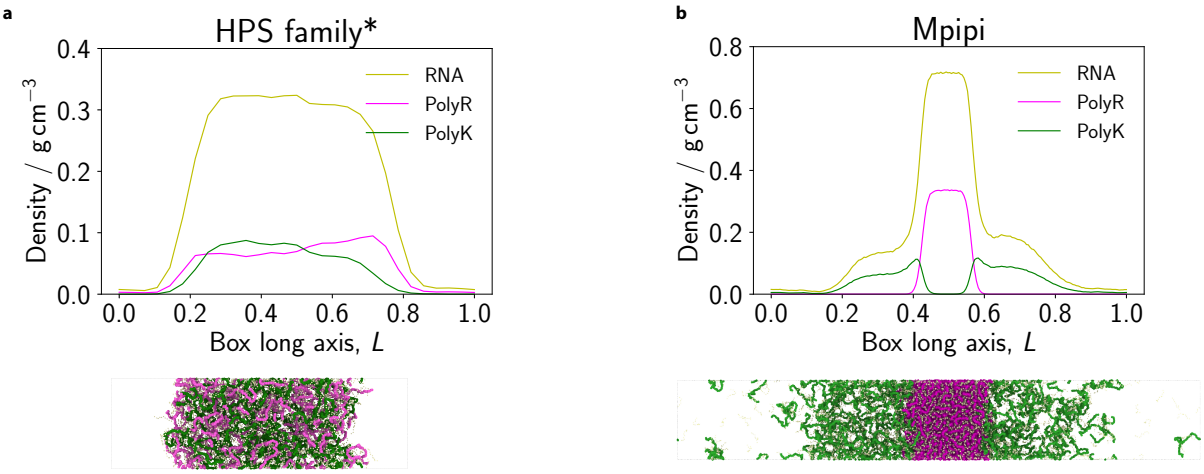

Figure S8. **Comparison of charge-matched RNA–PolyR–PolyK systems.** Plots of local density against the z-axis co-ordinate for RNA–PolyR–PolyK systems simulated with the **a** HPS-KR model and **b** Mpipi potential. **a** We simulate a mixture of PolyK (50 residues; 64 chains), PolyR (50 residues; 64 chains) and RNA (10 residues; 640 chains). For this simulation, we use the parameters for uridine, as proposed by Regy et al. [50]. We dubbed these simulations 'HPS family', which includes HPS-KR, HPS+cation–π(i) and HPS+cation–π(ii), since the Arg, Lys and RNA cross interactions are all the same in these three HPS-based models. **b** We simulate a mixture of PolyK (50 residues; 128 chains), PolyR (50 residues; 128 chains) and RNA (10 residues; 1280 chains). Here, we have extended the Mpipi model to include RNA parameters (see main text and Fig. S4). We also simulate a system containing 64, 64 and 640 chains of PolyK, PolyR and RNA, respectively, via the Mpipi potential and obtain similar multiphasic behaviour (see Fig. S9a). Simulation snapshots of each system are provided below the respective density plots. The colour code in the snapshot is consistent with that used in the density plots. The mixtures are simulated at $T/T_c \approx 0.8$, where T_c is the critical temperature for liquid–vapour phase separation.

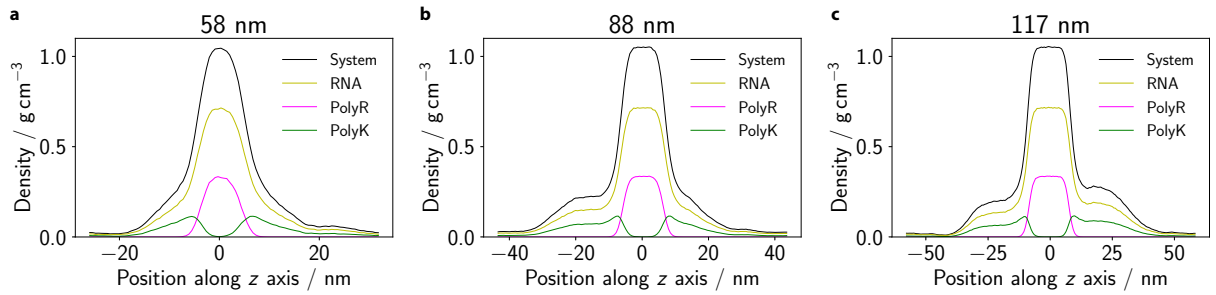

Figure S9. **Finite-size scaling analysis for the multiphase system.** Plots of local density against the longest (z) axis of the simulation box co-ordinate for RNA–PolyR–PolyK systems simulated with the Mpipi potential at $T/T_c \approx 0.8$, where T_c is the critical temperature for liquid–vapour phase separation. The lengths of the RNA, PolyR and PolyR chains are 10, 50, 50 beads, respectively. In **a**, 640 chains of RNA are used and 64 chains each of PolyR and PolyK. These numbers are increased by a factor of **b** 1.5 and **c** 2 in order to assess finite-size effects. The cross-sectional area of the interface in each simulation is fixed at 17.5 nm × 17.5 nm and the longest box dimension changes as indicated in the plot titles.

S8.3 Summary of potential parameters for all models

In SI Table X, we provide a summary of the parameters for each model we have considered using the same units throughout for ease of comparison. The spring constant refers to the harmonic spring constant k in the bonded term, $(k/2)(r-r_{\mathrm{ref}})^2$ (Eq. (2) of the main text), where r_{eff} is the equilibrium bond length for bonded amino-acid residues. Care must be taken not to confuse this quantity with the LAMMPS implementation of harmonic potentials in which the force constant is defined as $K = k/2$. In particular, as in our model, the HPS-KR and KH models [28] use $k = 8.03\,\mathrm{J\,mol^{-1}\,pm^{-2}}$, while the cation–π [40] and FB-HPS [26] reparameterisations use $k = 2.39\,\mathrm{kcal\,mol^{-1}\,Å^{-2}} = 1.0\,\mathrm{J\,mol^{-1}\,pm^{-2}}$ for the spring constant, and the HPS-Urry potential [29] uses $k = 4.0\,\mathrm{J\,mol^{-1}\,pm^{-2}}$. Although the value for the Coulomb cutoff is not reported in all studies, we have used a consistent value of 3.5 nm, except for the FB-HPS and TSCL-M2 models, where Coulomb cutoffs are specified. Although Dannenhoffer-Lafage and Best report a non-bonded potential cutoff of 1.0 nm [26], we presume this holds only for their calculation of the radii of gyration in single-molecule simulations, and we have used a 3σ cutoff, as for the remaining potentials, in order to observe any phase separation.

These parameter values have not always been reported in the literature very clearly, resulting in some confusion regarding both units and the inclusion, or otherwise, of a factor of $1/2$ in the spring constant. As a result, the HPS+cation–π models use the value for the spring constant as reported by Dignon and co-workers [28], although the intended parameterisation of the HPS-KR and KH potentials was as outlined in SI Table X. We have confirmed the latter point with one of the authors of Ref. 28.

Whilst our manuscript was under review, a pre-print with a further model ('TSCL') appeared on the bioRxiv [27]. For completeness, we also include the parameters for this potential in the table below. Interestingly, the dielectric constant in a temperature-dependent way using an empirical relation given by Akerlof and Oshry [87]. This in turn affects the Debye length. We found this to be an interesting and reasonably straightforward addition to this family of models; out of interest, we attempted a series of phase-diagram calculations for the Mpipi model with temperature-dependent dielectric constant and screening length. We have observed no appreciable difference in the behaviour of the A1-LCD variants, which suggests that within the relatively narrow range of temperatures at which we studied our systems, the relatively small reweighting of electrostatic interactions that this results in is not sufficient to change the phase behaviour significantly.

Finally, we remark that when representing non-bonded interactions with Lennard-Jones-based potentials, care must be taken in specifying and reporting the cutoffs used. The choice of the cutoff for non-bonded interactions can greatly affect the phase behaviour predicted by the model. For example, when we originally implemented the TSCL-M2 model, we used a cutoff of 3σ, as had been used for the remaining potentials, but this resulted in behaviour that was significantly different from that reported by the authors [27]. With their kind assistance, we eventually pinned down that the root cause of the discrepancy was this non-bonded cutoff, for which they used a value of 4 nm instead. The larger (4 nm) cutoff results in an increase in T_c of the order of 60 K. Similarly, in our initial implementation of the HPS-Urry potential [29], since no cutoff was reported in the manuscript, we initially used 3σ for the non-bonded cutoff; however, their supporting code suggests a cutoff of 2 nm may have been used instead, which again increases the critical temperatures, in this case by ∼20 K. Since the Wang–Frenkel potential is not truncated and therefore not subject to such abrupt changes with cutoff variations, this serves further to highlight yet another advantage of using it in preference to LJ-based potentials. For some of the LJ-based potentials we benchmark here, cutoff details were not specified in the corresponding manuscripts. Since all were derived from HPS, we have used a cutoff of 3σ in such cases, but we remark that, given the sensitivity of the phase behaviour to the details of the parameterisation, this may be a potential source of inconsistencies between implementations.

A full listing of all interaction parameters of the Mpipi potential is given in SI Tables XI and XII, and a LAMMPS implementation is provided as part of the supporting code.

Table X. Summary of model parameters. WF = Wang–Frenkel potential [19]. AH/LJ = Ashbaugh–Hatch-modified Lennard-Jones potential [88]. The spring constant refers to the harmonic spring constant k in the bonded term, $(k/2)(r - r_{\mathrm{ref}})^2$ (Eq. (2) of the main text), where r_{eff} is the equilibrium bond length for bonded amino-acid residues.

Model	Spring constant k / J mol^{-1} pm^{-2}	r_{eff} / nm	Charge on H / e	Short-ranged potential	Cutoff	Screening κ / nm^{-1}	Debye cutoff / nm
Mpipi	8.03	0.381	0.375	WF	3σ	1.26	3.5
KH[28]	8.03	0.381	0.5	AH/LJ	3σ	1.0	3.5
HPS-KR[28]	8.03	0.381	0.5	AH/LJ	3σ	1.0	3.5
FB-HPS[26]	1.0	0.38	0.5	AH/LJ	3σ*	1.0	3.0
HPS-Urry[29]	4.0	0.382	0	AH/LJ	2.0 nm	1.0	3.5
HPS+cation–π(i)[40]	1.0	0.38	0.5	AH/LJ	3σ	1.0	3.5
HPS+cation–π(ii)[40]	1.0	0.38	0.5	AH/LJ	3σ	1.0	3.5
TSCL-M2[27]	8.03	0.38	variable	AH/LJ	4.0 nm	variable	4.0

Table XI. Wang–Frenkel parameters for the Mpipi potential. For each residue, two lines are provided. The row highlighted in red lists ε/kJ mol^{-1} and the row highlighted in green lists σ/nm. The value of μ is 2 for all entries except the ones highlighted in blue [μ(V–I) = 4, μ(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	0.165536	0.284583	0.136758	0.147114	0.717619	0.186280	0.306152	0.324582	0.959705	0.094437	0.105776	0.502105	1.233991	0.902083	0.211635	0.851406	0.488298	0.247358	0.227764	0.083592
	0.646795	0.557618	0.656778	0.617823	0.664396	0.586850	0.614146	0.631818	0.659004	0.632057	0.648344	0.636519	0.675573	0.653821	0.593894	0.639251	0.618802	0.613446	0.609429	0.649639
G		0.403630	0.255806	0.266161	0.836666	0.305327	0.425199	0.443630	1.078752	0.213484	0.224823	0.621152	1.353039	1.021130	0.330682	0.970454	0.607345	0.366405	0.346812	0.202639
		0.469511	0.567134	0.528442	0.576639	0.498013	0.525925	0.543637	0.571317	0.541239	0.557893	0.548630	0.587924	0.566122	0.505252	0.551541	0.530902	0.524999	0.520872	0.558033
K			0.079986	0.118336	0.510189	0.157502	0.002063	0.002201	0.429701	0.065660	0.076998	0.473328	0.438466	0.482315	0.182858	0.365556	0.459520	0.218581	0.198987	0.054815
			0.667134	0.627940	0.673819	0.596669	0.623699	0.641358	0.668408	0.643354	0.659256	0.645969	0.684969	0.663228	0.603618	0.648660	0.628255	0.623106	0.619115	0.661800
T				0.128691	0.699197	0.167858	0.287729	0.306160	0.941283	0.076015	0.087354	0.483683	1.215565	0.883661	0.193213	0.832984	0.469876	0.228936	0.209342	0.065170
				0.588906	0.635202	0.557795	0.584987	0.602654	0.629806	0.603697	0.619809	0.607329	0.646378	0.624625	0.564801	0.610053	0.589610	0.584320	0.580321	0.621462
R					0.376209	0.738363	0.006368	0.006502	2.543119	0.646520	0.657859	1.054188	2.893893	2.280037	0.763718	0.519808	1.040381	0.799441	0.779847	0.635675
					0.683905	0.604909	0.632986	0.650709	0.678624	0.647780	0.664461	0.655828	0.695255	0.673424	0.612191	0.658837	0.638103	0.631985	0.627828	0.664491
A						0.207024	0.326896	0.345326	0.980449	0.115181	0.126520	0.522849	1.254736	0.922827	0.232379	0.872151	0.509046	0.268102	0.248509	0.104336
						0.527007	0.554510	0.572198	0.599537	0.571467	0.587918	0.576988	0.616122	0.594352	0.534119	0.579777	0.559268	0.553738	0.549687	0.588726
D							0.330938	0.344590	1.100321	0.235049	0.246387	0.642721	1.374603	1.042699	0.352247	0.007355	0.628914	0.387974	0.368380	0.224208
							0.582352	0.600058	0.627647	0.597874	0.614509	0.605004	0.644244	0.622456	0.561730	0.607876	0.587280	0.581455	0.577340	0.614699
E								0.358242	1.118751	0.253479	0.264818	0.661151	1.393034	1.061129	0.370677	0.007494	0.647344	0.406404	0.386811	0.242639
								0.617767	0.645370	0.615504	0.632146	0.622721	0.661969	0.640180	0.579422	0.625600	0.604999	0.599152	0.595036	0.632333
Y									1.753874	0.888606	0.899945	1.296274	2.028156	1.696252	1.005804	1.645576	1.282467	1.041527	1.021934	0.877761
									0.673363	0.642355	0.659039	0.650525	0.690005	0.668159	0.606828	0.653569	0.632798	0.626630	0.622465	0.659055
V										0.023338	0.034677	0.431006	1.162888	0.830984	0.140536	0.780308	0.417199	0.176259	0.156666	0.012493
										0.626600	0.639093	0.619973	0.658912	0.637178	0.578166	0.622610	0.602261	0.597433	0.593567	0.671801
L											0.046016	0.442345	1.174227	0.842323	0.151875	0.791646	0.428538	0.187598	0.168004	0.023832
											0.653407	0.636649	0.675594	0.653860	0.594697	0.639294	0.618936	0.614021	0.610125	0.660772
Q												0.838674	1.570556	1.238652	0.548204	1.187976	0.824867	0.583927	0.564334	0.420161
												0.627785	0.667141	0.645328	0.584255	0.630743	0.610059	0.604035	0.599885	0.636715
W													2.302443	1.970538	1.280086	1.919858	1.556753	1.315814	1.296220	1.152048
													0.706655	0.684798	0.623414	0.670209	0.649416	0.643217	0.639053	0.675603
F														1.638630	0.948182	1.587954	1.224845	0.983905	0.964312	0.820139
														0.662955	0.601640	0.648367	0.627603	0.621441	0.617279	0.653879
S															0.257734	0.897506	0.534397	0.293457	0.273864	0.129691
															0.541267	0.587063	0.566533	0.560932	0.556848	0.595253
H																0.113872	1.174168	0.933229	0.913635	0.769463
																0.633778	0.613018	0.606865	0.602702	0.639315
N																	0.811064	0.570124	0.550531	0.406358
																	0.592335	0.586308	0.582164	0.618999
P																		0.329185	0.309591	0.165419
																		0.580616	0.576515	0.614386
C																			0.289997	0.145825
																			0.572436	0.610587
I																				0.001653
																				0.692168

Table XII. Wang–Frenkel parameters for interactions with RNA bases in the Mpipi potential. For each residue, two lines are provided. The row highlighted in red lists ε/kJ mol^{-1} and the row highlighted in green lists σ/nm. The value of μ is 3 for all entries. All charged residues (both amino acids and nucleic-acid bases) have $q = \pm 0.75e$, as appropriate, except H, which has $q = 0.375e$, where e is the elementary charge. The nucleic-acid codes are shown in blue to distinguish them from corresponding amino-acid codes.

	A	C	G	U
M	1.153872	0.783588	1.162240	0.720828
	0.745398	0.734398	0.748897	0.731898
G	1.272919	0.902635	1.281287	0.839875
	0.656756	0.645756	0.660255	0.643255
K	0.666658	0.444487	0.671678	0.406831
	0.755567	0.744567	0.759067	0.742067
T	1.135450	0.765166	1.143818	0.702406
	0.716453	0.705453	0.719953	0.702953
R	2.518417	1.777849	2.535153	1.652329
	0.763953	0.752953	0.767453	0.750453
A	1.174616	0.804332	1.182984	0.741572
	0.685504	0.674504	0.689004	0.672004
D	1.236573	0.866289	1.244941	0.803529
	0.713176	0.702176	0.716676	0.699676
E	1.250225	0.879941	1.258593	0.817181
	0.730884	0.719884	0.734384	0.717384
Y	1.948041	1.577757	1.956409	1.514997
	0.758682	0.747682	0.762182	0.745182
V	1.082773	0.712489	1.091141	0.649729
	0.735300	0.724300	0.738800	0.721800
L	1.094112	0.723828	1.102480	0.661068
	0.748704	0.737704	0.752204	0.735204
Q	1.490441	1.120157	1.498809	1.057397
	0.735893	0.724893	0.739393	0.722393
W	2.222327	1.852043	2.230695	1.789279
	0.775327	0.764327	0.778827	0.761827
F	1.890419	1.520135	1.898787	1.457375
	0.753478	0.742478	0.756978	0.739978
S	1.199971	0.829687	1.208339	0.766927
	0.692634	0.681634	0.696134	0.679134
H	1.128040	0.757756	1.136408	0.694996
	0.738889	0.727889	0.742389	0.725389
N	1.476634	1.106354	1.485002	1.043590
	0.718168	0.707168	0.721668	0.704668
P	1.235698	0.865414	1.244066	0.802650
	0.712308	0.701308	0.715808	0.698808
C	1.216105	0.845821	1.224473	0.783061
	0.708218	0.697218	0.711718	0.694718
I	1.071932	0.701644	1.080300	0.638884
	0.768084	0.757084	0.771584	0.754584
A	2.142208	1.771924	2.150576	1.709164
	0.844000	0.833000	0.847500	0.830500
C		1.401640	1.780292	1.338880
		0.822000	0.836500	0.819500
G			2.158944	1.717532
			0.851000	0.834000
U				1.276120
				0.817000

S8.4 Estimation of error bars

Each individual simulation has a certain error associated with the density and the temperature; for example, in a typical thermostatted simulation with these models with system sizes we have studied, the mean temperature usually has a standard deviation of the order of 3 K. However, fluctuations about the mean are symmetrical for the temperature and the density, and so may not be reflected in the error associated with the value of the critical temperature. Rather than propagate the error from individual simulation data, we therefore estimate the error in our calculation of the critical temperature by running independent simulations to determine the phase diagram and in turn, using Eqs (6) and (7) of the main text, the upper critical solution temperature. For the hnRNPA1-LCD variants with the Mpipi potential, the standard deviation across 5 independent data points is given for each variant in SI Table XIII. The mean standard deviation is 1.8 K. The largest standard deviation corresponds to −6R+6K variant; the reason for this is that the density profile for −6R+6K exhibits considerable fluctuations within the high-density phase. These fluctuations are primarily caused by interactions between charged residues, and they make the density of the high-density phase more difficult to determine unambiguously. However, all estimates of T_c of this variant correspond to the lowest critical temperature of any variant. Moreover, the error in the critical temperature is not generally larger than the typical fluctuations in temperature for each simulation.

Table XIII. Standard deviation, including Bessel correction, for 5 independent estimates of the critical temperature for each hnRNPA1-LCD variant.

Variant	$\sigma(T_c)$ / K
WT	1.19
−3R+3K	2.58
−4F−2Y	0.74
−6R+6K	5.20
+7F−7Y	1.47
+7K+12D	0.86
+7R+12D	1.09
−9F+3Y	0.36
−12F+12Y	2.30

S8.5 List of algorithms and software

Table XIV. A summary of algorithms and software used in this work.

Method/software	Use
Umbrella sampling [89, 90]	probing interactions between pairs of amino acids and nucleic acids
WHAM [91]	obtaining PMF profiles from umbrella sampling runs
Bayesian bootstrap method [85]	estimating errors in PMF calculations
restrained electrostatic potential (RESP) [92]	refitting sidechain charges from quantum-mechanical calculations for cation–π pairs
Gaussian [93]	quantum-mechanical calculations on cation–π pairs
Absinth [69, 94]	estimating coil-to-globule transition temperature
Gromacs	PMF calculations
lincs [95]	rigid-molecule constraint algorithm in all-atom simulations
particle-mesh Ewald summation [96]	computation of long-ranged electrostatics
PyMol [97]	visualisation of nucleic-acid dimers in Fig. S4
direct-coexistence simulations [20, 98–100]	simulation set-up for computing phase diagrams
Lammps [101]	used to run simulations to obtain R_g values and to compute phase diagrams