Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal | View Issue

Cite this: *Chem. Sci.*, 2023, **14**, 1820

All publication charges for this article have been paid for by the Royal Society of Chemistry

Thermodynamic origins of two-component multiphase condensates of proteins†

Pin Yu Chew, [a] Jerelle A. Joseph, [abc] Rosana Collepardo-Guevara*[abc] and Aleks Reinhardt*[a]

Received 24th October 2022
Accepted 6th January 2023

DOI: 10.1039/d2sc05873a

rsc.li/chemical-science

Intracellular condensates are highly multi-component systems in which complex phase behaviour can ensue, including the formation of architectures comprising multiple immiscible condensed phases. Relying solely on physical intuition to manipulate such condensates is difficult because of the complexity of their composition, and systematically learning the underlying rules experimentally would be extremely costly. We address this challenge by developing a computational approach to design pairs of protein sequences that result in well-separated multilayered condensates and elucidate the molecular origins of these compartments. Our method couples a genetic algorithm to a residue-resolution coarse-grained protein model. We demonstrate that we can design protein partners to form multiphase condensates containing naturally occurring proteins, such as the low-complexity domain of hnRNPA1 and its mutants, and show how homo- and heterotypic interactions must differ between proteins to result in multiphasicity. We also show that in some cases the specific pattern of amino-acid residues plays an important role. Our findings have wide-ranging implications for understanding and controlling the organisation, functions and material properties of biomolecular condensates.

1 Introduction

Biomolecular condensates are involved in controlling many aspects of cell biology and pathology. By dynamically segregating particular biomolecules,[1,2] condensates help create intracellular micro-environments that contribute to the regulation of chemical reactions[3,4] and mediate a variety of fundamental biological processes, from cell signalling[5–9] to RNA metabolism,[10–12] response to stress,[13–17] regulation of transcription[18–23] and DNA repair.[24–26] Moreover, dysregulation of biomolecular phase separation has been associated with a growing list of diseases from cancer to neurodegeneration.[27–32] The link between aberrant liquid–liquid phase separation (LLPS) and disease highlights the desirability of developing new tools to manipulate the properties of condensates to bypass pathologies. Anti-cancer drugs have already been shown to concentrate preferentially into condensates,[33] and some small molecules can modulate LLPS,[34] making biomolecular condensates potential drug targets.

A number of biomolecular condensates *in vitro* and in cells have been observed to display multiphase architectures: structural heterogeneity over mesoscopic length scales, where immiscible phases with different compositions coexist within the same liquid droplet. *In vitro*, multiphase droplets have been observed in various multi-component mixtures, all involving RNA and at least two different proteins, *e.g.* those of poly(PR) and RNA homopolymers,[35] of polyR, polyK and polyU,[36] and of prion-like and arginine-rich polypeptides, and RNA.[37] Inside cells, a prime example is the nucleolus, a highly multi-component system, which exhibits a multilayered architecture[38] that is thought to be important for sequential processing of nascent rRNA transcripts.[10] Similar internal structuring is also found in stress[13] and P granules.[39–41] The presence of multiple condensed phases in a single condensate may reflect different biological processes taking place in physically separated regions within the same compartment.[42,43]

The emergence of a multiphase organisation has been associated with the physicochemical diversity of the various molecular components. One hypothesis is that multiphase condensates emerge in multicomponent systems when there is competition for a shared binding partner.[36,37] For instance, competing protein–protein and protein–RNA interactions can provide a regulatory mechanism for the organisation of multiphase condensates.[37] Simulations of multi-component systems[44,45] have further revealed that the phase boundary for demixed phases is sensitive to the variance of intermolecular interaction strengths: if it is sufficiently large, multiple distinct phases can form.[45] From the physicochemical point of view, the formation of an interface between two

[a]*Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge, CB2 1EW, UK. E-mail: rc597@cam.ac.uk; ar732@cam.ac.uk*

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

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

† Electronic supplementary information (ESI) available: Additional figures providing further data for the systems investigated in the main manuscript; data for and descriptions of analogous systems to the ones reported in the main text; a list of protein sequences used. See DOI: https://doi.org/10.1039/d2sc05873a

phases is thermodynamically unfavourable and must be compensated for by free-energetically favourable interactions in the demixed system. The extent to which an interface is unfavourable is quantified by the interfacial free-energy density, which gives the free-energy penalty per unit area of the interface. Unsurprisingly, the various forms of structuring and morphological patterns of multiphase condensates have been suggested to be modulated by the difference in the interfacial free-energy densities of the phases,[38,46–49] which in turn depend on the sequence-encoded molecular interactions of the components.[35–37,48] Phases with high interfacial free-energy densities are expected to be engulfed by those with lower ones,[50] while some phases form completely separate droplets due to high interfacial tensions.[51] These observations, together, might suggest that RNA is essential to sustain multiphase condensates, or that a large number of components are needed. But is that the case? To be able to identify the rules governing multiphasicity, here, we focus on systems that are rather simpler, with only two protein components. We develop a molecular-simulations approach that allows us to understand the physicochemical characteristics they must have to form multiphase condensates.

Although one could in principle speculate, based on physical intuition, which pairs of protein sequences might give rise to multiphase architectures, this strategy is likely only feasible for simple amino-acid sequences. Even then, the phase behaviour of multi-component systems with multiple coexisting condensed phases is far more challenging to predict than that of single-component condensates, especially given the complexity and diversity of biologically relevant proteins. Computational approaches, such as genetic algorithms, can help explore the vast size of protein sequence space by automating the design of protein sequence mutations. Broadly speaking, genetic algorithms use mechanisms inspired by biological evolution, such as crossovers and mutations, to optimise properties of a system,[52–54] and have long been used in optimisation problems in many fields,[55–65] including those with biological applications such as protein engineering[66–68] and drug design.[69,70] Genetic algorithms have recently been applied to evolve protein sequences to (de)stabilise their condensates.[71,72] However, in order to use genetic algorithms in the context of LLPS, we first need to quantify the protein properties we wish to optimise. Recently, computer simulations have connected features of individual biomolecules to their phase behaviour.[73] Depending on the question being addressed, models from atomistic[74–80] *via* residue-level[81–87] to minimal,[80,88–90] alongside other computational approaches such as predictive algorithms and machine-learning methods,[85–87,91–95] have all been used with success. Simple models for phase separation of multi-component mixtures and multiphase organisation have also been studied in detail with a combination of simulations and theory.[96–100]

Motivated by these ideas, here we develop an evolutionary algorithm that goes beyond manipulating the stability of condensates and allows us to enforce or inhibit a desired spatial organisation of biomolecules inside multi-component condensates. We couple molecular-dynamics simulations of a residue-resolution coarse-grained protein model that achieves near quantitative agreement with experiments[84] with a genetic algorithm[71] to evolve protein sequences towards increasing 'multiphasicity', which we define to be the difference in the compositions of the two coexisting phases of a multiphase condensate. The multiphasicity of a condensate increases with the purity of the two coexisting phases. We first demonstrate that we can increase the multiphasicity of a protein mixture using a genetic algorithm with an appropriate fitness function to evolve the amino-acid sequence of one of the two proteins (Section 2.1). We then show that we can design a protein sequence to act as a multiphase partner for some other protein of choice by coevolution (Section 2.2), including proteins of biological relevance such as the low-complexity domain (LCD) of heterogeneous nuclear ribonucleoprotein A1 (hnRNPA1). Finally, we analyse the changes in interaction energies (Section 2.3) and amino-acid patterning (Section 2.4) to probe the factors driving the formation of multilayered condensates.

2 Results

2.1 Genetic algorithms can improve the separation of two-protein multilayered systems

Genetic algorithms can effectively evolve amino-acid sequences to find mutations that give rise to desired changes in phase behaviour.[71,72] During the evolution, each protein in a population is associated with a fitness value, as computed with a fitness function, and the population is evolved with local mutations and crossovers between individual sequences towards a fitter – albeit not necessarily the fittest – solution. Rather than determining an optimal solution (however defined), the genetic algorithm follows local gradients in fitness space towards a better solution that is not drastically dissimilar to the starting sequence. The success and efficiency of the genetic algorithm depend on the quality of the fitness function, which should be relatively simple and inexpensive to compute, while at the same time sufficiently complex to act as a suitable proxy for the property being evolved. For simplicity, here we focus on evolving two-component protein condensates towards higher multiphasicity (as defined above). Nonetheless, our approach can readily be generalised to condensates with a larger number of components, which are in any case expected to behave similarly.[45]

To design our fitness function, we start with the simplest metric of multiphasicity we could conceive for a two-component system: the difference in the number densities between the two different proteins at the centre of the multilayered condensate. This quantity is small when the two-component mixture phase separates into a homogeneous condensate (*i.e.* low multiphasicity) and large when it phase separates into a multilayered condensate with each layer enriched in a different protein (*i.e.* high multiphasicity) [Fig. 1a]. However, if the two-component mixture phase separates into a homogeneous condensate that is depleted of one of the two proteins and a dilute phase that is enriched in the depleted protein, this metric is large even though the multiphasicity is low [Fig. 1a(iii)]. To avoid this, we introduce a second term in the fitness function that penalises the accumulation of either protein in the dilute phase, namely the sum of the number densities of the two components far away from the condensate [eqn (1)], scaled by a weighting parameter s. A large s might seem desirable to ensure the enrichment of both components inside the condensate; however, when its value is too large, it dominates the fitness

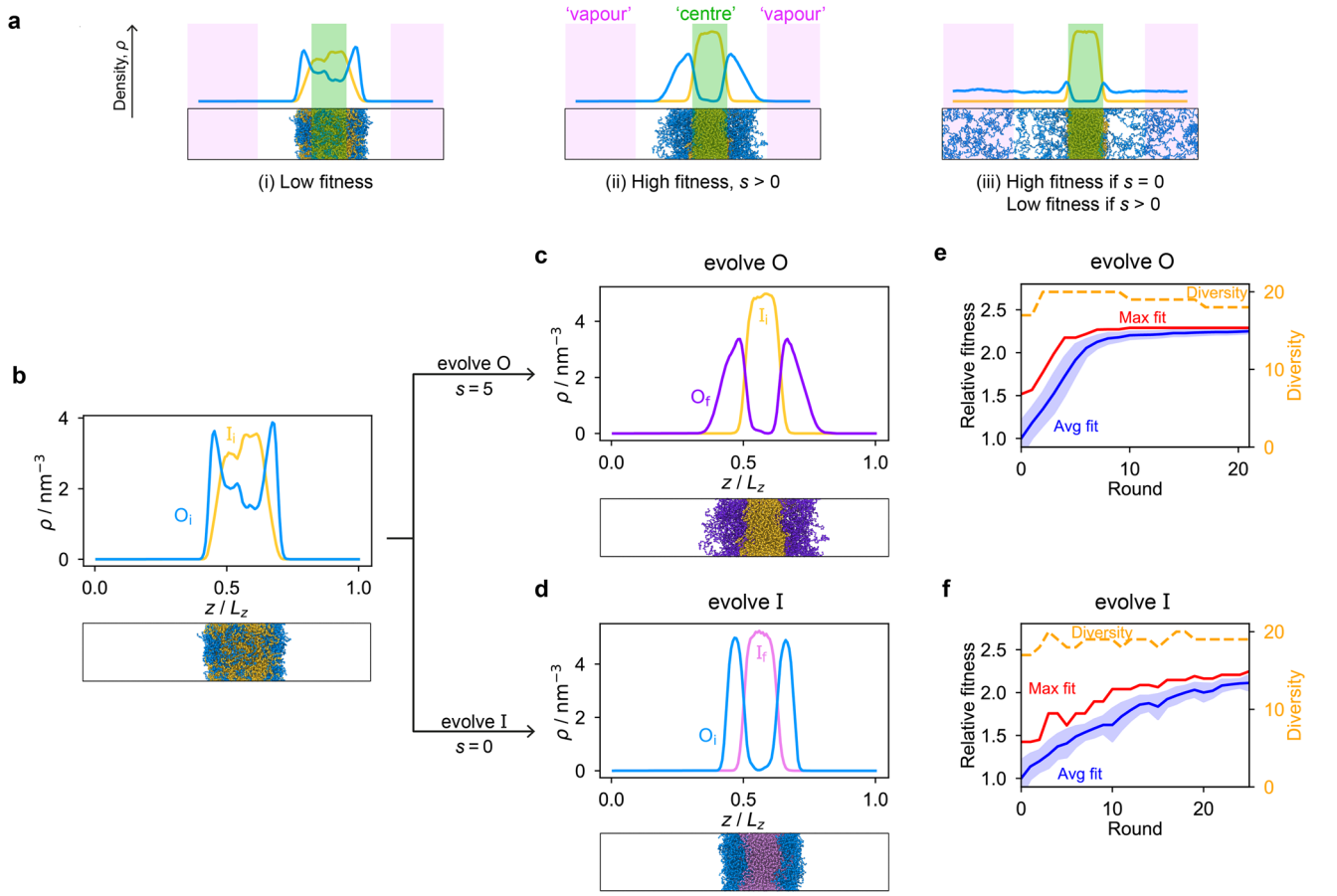

Fig. 1 The genetic algorithm can improve multiphasicity. (a) Examples of systems with high and low fitness values. The parameter s determines the trade-off between increasing the difference in compositions between the two phases and obtaining two stable liquid-like phases alongside a vapour-like phase. (b) Density profile (number density ρ against the long axis of the simulation box) of the initial two-component system considered. Protein I is enriched in the inner layer and protein O in the outer layer. Here $O_i = (FAFAA)_{10}$ and $I_i = F_{50}$ with random noise added by introducing mutations with probability 0.60 to the latter such that both sequences mix to give a system with low multiphasicity. (c) Density profile of the final evolved system with maximum fitness when protein O is evolved, and (d) when protein I is evolved in separate runs. Snapshots of the corresponding multiphase droplets are provided for each case. (e) and (f) Fitness (relative to the average fitness in round 0) as a function of genetic-algorithm progression. The shaded area corresponds to the standard deviation of the fitness across the population in each round. In both cases, a high population diversity is maintained.

function, making the first term irrelevant in magnitude, and can actually favour homogeneous condensates instead. In general, this weighting parameter can be tuned as necessary depending on the specific system of interest [see Fig. S1†].

When evolving a two-component protein system towards increasing multiphasicity, the goal is to obtain a set of mutations to the amino-acid sequences of the two proteins such that the mutated proteins form a condensate with a more segregated multilayered architecture than the starting pair. We refer to the protein enriched in the core of the multilayered condensate as the 'inner protein' or 'protein I', and the one concentrated in the outer layer as the 'outer protein' or 'protein O'. There are several routes one could take: one could evolve either the inner or the outer sequence in separate evolution runs, or even evolve both sequences, simultaneously or alternately, in the same run. In our evolution runs, we evolve either the inner sequence or the outer sequence whilst keeping the other sequence unchanged in order to simplify the subsequent analysis of driving forces. To evolve the sequence, we apply the genetic algorithm as described above (and in the Methods section), computing a sequence's fitness by performing direct-coexistence molecular-dynamics simulations of the two-component mixture in a slab geometry using our residue-resolution coarse-grained model, Mpipi,[84] at a fixed temperature.

As a preliminary test of our approach, we first consider a mixture of $(FAFAA)_{10}$ and F_{50} and add sufficient mutations (*i.e.* random noise) to the latter to ensure that the initial condensate has a low fitness. The mutations are added in a similar way to how the initial population of sequences is generated in the genetic algorithm (Methods), but with a replacement probability of 0.60. In this initial state, $(FAFAA)_{10}$ is slightly enriched at the interface and is deemed the outer sequence, but there is an appreciable degree of mixing with the inner sequence. We then perform two separate evolution runs, evolving (a) the inner sequence whilst keeping the outer unchanged, and (b) the outer sequence whilst keeping the inner sequence unchanged. A comparison of the density profiles of the initial system to that of the final systems with the maximum fitness [Fig. 1b–d] confirms that our fitness function can successfully guide the initial system towards increasing multiphasicity in

both cases. When the inner sequence is evolved, a stable multilayered condensate can be obtained with a weighting parameter $s = 0$ used in the fitness function; by contrast, when evolving the outer sequence, the final result is sensitive to the value of the parameter s and $s > 0$ must be used. We show in Fig. S1† the final evolved systems obtained with different values of s. Additionally, the genetic-algorithm progressions in each case are depicted [Fig. 1e and f]. We plot the mean fitness of the population (blue curve), the fitness of the fittest individual (red curve) and the number of distinct sequences in the population (*i.e.* its 'diversity'; yellow curve). Finally, we show the results for an evolution run starting from a different initial system with different sequences in Fig. S8a and b,† demonstrating that the behaviour observed is not sensitive to the initial sequence choice.

2.2 Multilayered condensates can be designed by coevolving a partner protein sequence alongside a protein of interest

Armed with an effective fitness function for multiphasicity, we next set out to use our genetic algorithm to guide the design of partner proteins that result in multilayered condensates alongside known phase-separating proteins of interest (*e.g.* a naturally occurring protein, such as hnRNPA1 LCD used below). There are many challenges involved in designing a partner protein for this purpose, such as ensuring that it phase separates under similar experimental conditions to the protein of interest (*e.g.* salt, pH, and temperature), and that it establishes suitable associative interactions with the protein of interest to form a single multilayered condensate. Thus, to facilitate convergence in this more complex scenario, we start our coevolution approach [Fig. 2a] from an initial reference system of two proteins, both different from our protein of interest, which phase separates into a multilayered structure with a high degree of multiphasicity. The initial reference systems used in the coevolution runs are designed using simple generic sequences of amino acids based on knowledge from previous experimental and theoretical studies showing that immiscible phases form when interaction strengths between their components are sufficiently different.[35,36,45,48,49,101] In particular, we choose the inner sequence to have a high aromatic sticker content [*e.g.* F_{135} or Y_{135}; see below], since aromatic residues exhibit strong favourable interactions. For the outer sequence, we choose a simple sequence that combines some sticker residues with spacer residues [*e.g.* $(FAFAA)_{10}$], so that the overall interaction is less strong compared to the inner sequence. To arrive at our protein of interest, we then systematically mutate one of the two reference sequences throughout the coevolution run. These systematic mutations are done gradually, once every 5 rounds, by randomly changing ~10% of the residues of this protein selected from those that have yet to be changed to the amino acid of the target protein. Simultaneously, we evolve the other sequence with the genetic algorithm using our fitness function [eqn (1)].

Since only a small proportion of the residues in the sequence that is systematically changed are modified each time, multiphasicity can be maintained at least to some degree throughout the process. This procedure ensures that there is a gradient of the fitness in sequence space in the direction of increasing multiphasicity which the genetic algorithm can evolve towards at every round during the coevolution. If one started from the reference multilayered system and changed one of the sequences to the target sequence in one go, this may result in full mixing of the two components within a homogeneous liquid-like phase and hence the loss of multiphasicity altogether. From combinatorics, there are numerous sequences that can form a homogeneous condensate with the target sequence; using the genetic algorithm starting from a fully mixed state would therefore be inefficient, as the initial random search for possible mutations that result in multiphasicity would be slow before there is a gradient in sequence space towards increasing multiphasicity that can be exploited by the genetic algorithm.

Of course in some cases, changing one sequence in the reference system to the target protein may still give a multilayered system, albeit likely with a lower degree of multiphasicity. Alternatively, one may also be able to propose, based on physical intuition and understanding of the intermolecular interactions that give rise to the formation of multiphase condensates, a protein sequence that can form a multilayered condensate with the target protein. In such situations, it is possible to use the genetic algorithm directly to evolve the system towards an increasing degree of multiphasicity, as discussed in Section 2.1. However, importantly, the coevolution approach we have outlined would find a possible solution much more efficiently even if an initial multilayered system with the target protein is unknown and difficult to predict by hand.

Here, to demonstrate the robustness of the coevolution approach, we select the initial reference systems and target sequences in the coevolution runs such that making the systematic change directly in one go results in complete mixing to give a single homogeneous liquid-like phase [*cf.* Fig. S19a†]. We show the results of the coevolution approach tested on systems with simple generic sequences in Section S12.

2.2.1 A multiphasic partner can be found for hnRNPA1 LCD and its variants. To demonstrate how the coevolution approach can be used to design multiphase condensates containing naturally occurring phase-separating proteins, we turn our attention to the LCD of hnRNPA1 (denoted here as A1 LCD). We first focus on designing a multilayered condensate with A1 LCD concentrated at the centre, and below we look at the converse case. To predict the partner sequence that forms a multilayered condensate with A1 LCD at the centre, we start our procedure with a mixture of $I = F_{135}$ and $O = (FAFAA)_{10}$, and then systematically change the inner protein to A1 LCD whilst evolving the outer protein using the genetic algorithm. We choose the initial inner protein such that its length is the same as that of A1 LCD. During the first 45 rounds of the coevolution procedure, the residues of the inner protein F_{135} are systematically and gradually changed to those of A1 LCD [Fig. S2e†], whilst the outer protein $(FAFAA)_{10}$ is evolved. We continue the genetic algorithm on the outer protein for an additional 20 rounds to increase the degree of multiphasicity of the system further. We show the density profiles and snapshots of the initial system and the final evolved system at the end of the coevolution run in Fig. S2b and c;† the final system exhibits two liquid-like phases of different composition with A1 LCD enriched in the centre, demonstrating that the coevolution

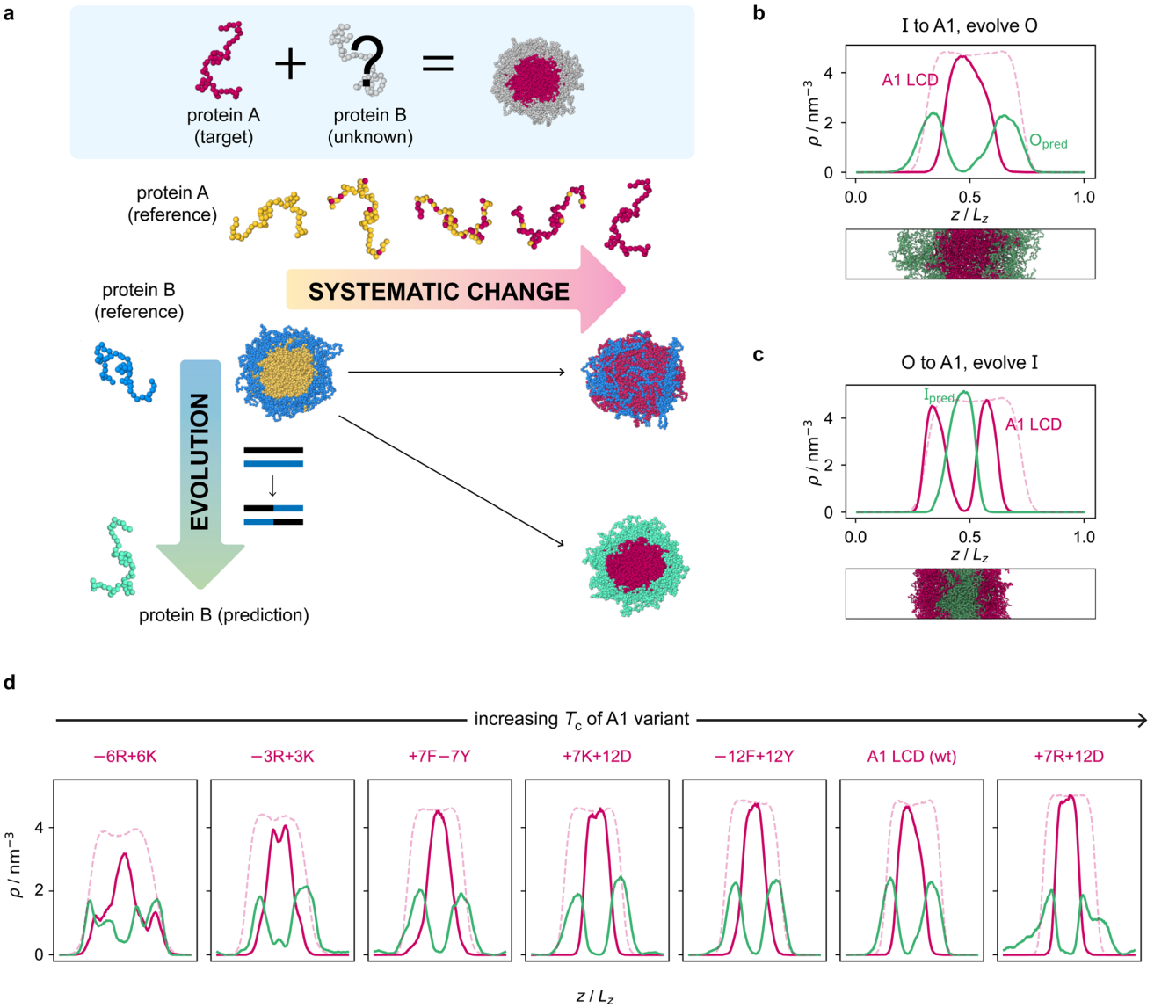

Fig. 2 Coevolution of a multilayered condensate partner for A1 LCD. (a) Illustration of the coevolution approach. Starting from an initial two-component reference system with high multiphasicity, we systematically change one sequence to the predetermined target protein (yellow to pink), while simultaneously evolving the other sequence (blue to green) to predict a partner sequence that forms a multilayered condensate with the target protein. Spherical droplets are shown for clarity; in direct-coexistence simulations, we use a slab geometry instead. (b) Density profile of the final evolved system with maximum fitness from the coevolution run where A1 LCD is designed to be at the centre and (c) where A1 LCD is designed to be on the outside of the condensate. The pink dashed line is the density profile of a single-component system of A1 LCD equilibrated at the same temperature (250 K). (d) Density profiles of the final evolved systems with maximum fitness from the coevolution runs with different variants of A1 LCD as the inner sequence. The densities of the A1 variant and the predicted outer partner sequence are plotted in pink and green, respectively. The pink dashed lines are the density profiles of a single-component system of the corresponding A1 variant equilibrated at the same temperature. The A1 variants are arranged in order of increasing upper critical solution temperature.

approach is able to predict a partner sequence that forms a multilayered condensate with A1 LCD.

The genetic-algorithm progression of this coevolution run is shown in Fig. S2d.† The average fitness and the maximum fitness suddenly decrease in rounds where systematic changes are made and the fitness of the entire population is recalculated. Although the outer sequence is evolved using the genetic algorithm for a considerable number of rounds after the inner sequence has been completely changed to A1 LCD, the maximum fitness does not improve in these rounds, suggesting that we have reached a local maximum in the fitness function. Changing the starting sequence of the protein being evolved does not appear to offer sufficient flexibility to support a higher degree of multiphasicity. We hypothesise that a higher multiphasicity might be achieved by increasing the length of the partner sequence that is evolved, since more sequence variations are possible with a longer sequence. To test this idea, we increase the length of the outer protein from 50 to 100 residues, but keeping the total number of protein residues

unchanged, and repeat our coevolution procedure. Specifically, we start the coevolution from a mixture of $I = F_{135}$ and $O = (FAFAA)_{20}$, and then systematically and gradually change the inner protein to A1 LCD whilst evolving the outer sequence [Fig. S4a†]. The density profile of the final evolved system with the longer partner sequence is shown in Fig. 2b [see also Fig. S6b and S5a†]. As hypothesised, the degree of multiphasicity of the final system is considerably improved with a longer protein partner, likely, as speculated, because of the greater flexibility in sequence choice with a longer protein.

In principle, the resulting partner sequences obtained from the coevolution run depend on the identity of the two proteins in the initial reference system, and it is not immediately obvious how to choose the reference system sensibly. Indeed, our work highlights that the solution to this problem is not unique and multiple different partner sequences can form diverse multilayered condensates with a specific target protein of interest. If we wish to find a possible solution, rather than a specific one, starting from any reference multilayered system should be feasible. To demonstrate this, we repeat the coevolution run starting from a different reference system with different protein sequences, namely a mixture of $O = N_{100}$ and $I = Y_{135}$. The latter, at the centre of the multilayered condensate, is then systematically changed to A1 LCD, while the outer protein is evolved using the genetic algorithm. Density profiles [Fig. S3b and c†] confirm that the coevolution approach is again successful; of course, unsurprisingly, the final evolved partner sequence is considerably different from before, since we expect it to retain at least some features from the initial reference sequence.

We now test the ability of our coevolution algorithm to predict the amino-acid sequence of a protein partner that forms a multilayered condensate with A1 LCD concentrated towards the interface of the condensate. To do so, we construct a reference multilayered system of $I = (FAFAA)_{20}$ and $O = (FIQII)_{27}$, but now we change the outer protein systematically and gradually to A1 LCD whilst evolving the inner one. As desired, we show [Fig. 2c] that the system forms a multilayered condensate with A1 LCD towards the interface [see also Fig. S6c and S5b† for further details].

Finally, to demonstrate the robustness of the approach to the target protein sequence, and to allow us to investigate if there are any overarching governing principles of multiphasicity that we can identify, we repeat the coevolution approach to find partner sequences for different variants of A1 LCD. In these cases, we choose the final multilayered condensates to have the A1 LCD variant concentrated in the centre. The phase diagrams of a set of A1 LCD variants have recently been determined both experimentally[102] and computationally using the same coarse-grained model that we have used in this work.[84] We consider sequences with similar, higher and lower upper critical solution temperatures compared to the wild type (WT), ensuring that sequences with distinct features are represented. In particular, we focus on two aromatic variants, +7F−7Y and −12F+12Y, two mixed-charged variants, +7R+12D and +7K+12D, and two arginine–lysine variants, −6R+6K and −3R+3K. As for the WT, we start coevolution runs from a mixture of $I = F_{135}$ and $O = (FAFAA)_{20}$. We show the density profiles of the final evolved systems in Fig. 2d and the corresponding genetic-algorithm progressions in Fig. S5c–h.† The coevolution approach successfully predicts a suitable partner sequence for each A1 variant. The overall trends in the change in fitness are similar for all of the variants, although the final systems have varying degrees of multiphasicity. In particular, the final evolved systems with the −6R+6K and −3R+3K variants, which have considerably lower critical temperatures than the WT, are less well separated. We speculate that this result can be improved with a longer partner sequence, as we have shown for the WT.

2.3 Multilayered condensates are driven by the difference in component interaction strengths

Predicting a partner sequence to design multilayered condensates containing a certain target sequence is of practical importance. However, molecular simulations allow us to go one step further and understand the underlying physical and molecular driving forces for the observed behaviours. To identify which properties are important for multiphasicity, we analyse the changes in the composition and patterning of the evolved sequence in the evolution runs with generic sequences, where one component remains unchanged throughout, and in the coevolution runs with systematic changes of one component to A1 LCD.

We summarise the changes in the amino-acid composition of the evolved sequences in Fig. 3. In the evolution runs with generic sequences where the outer sequence is evolved while the inner sequence is kept unchanged, the amino-acid composition of the evolved sequence changes to favour fewer aromatic residues [Fig. 3a]. By contrast, when the inner sequence is evolved while the outer sequence is kept unchanged, evolution favours a higher number of aromatics [Fig. 3b]. These observations suggest that within multilayered condensates, when other features are kept constant (*e.g.* protein charge, disorder and length), protein sequences with a higher aromatic content are likely to cluster towards the centre. The mean interaction strengths amongst amino acids change across the evolution runs; we use the parameter $\varepsilon_{i,\mathrm{Mpipi}}$ of the Mpipi model [see the Methods section] to estimate these changes. The average of $\varepsilon_{i,\mathrm{Mpipi}}$ across all residues in the evolved sequence decreases over the course of the genetic-algorithm run when the outer sequence is evolved [Fig. S7a†], but increases when the inner sequence is evolved [Fig. S7b†]. By comparing the final evolved sequences with maximum fitness across three independent evolution runs [Fig. 3], we note that residues become more strongly interacting on average when the inner sequence is evolved, and conversely, more weakly interacting when the outer sequence is evolved. As evidenced by Fig. 3, there are numerous ways of achieving such a change in interaction strengths, and the solution to this optimisation problem is unsurprisingly not unique. This is because for a given predetermined sequence, from combinatorics, there should exist multiple different protein sequences that can form a multilayered condensate with it, *i.e.* many sequences are similarly fit and the corresponding well in the fitness landscape is broad. This important observation puts forward the idea that the formation of multiphase condensates is a general phenomenon requiring only a generic set of interaction rules governed by the overall chemistry

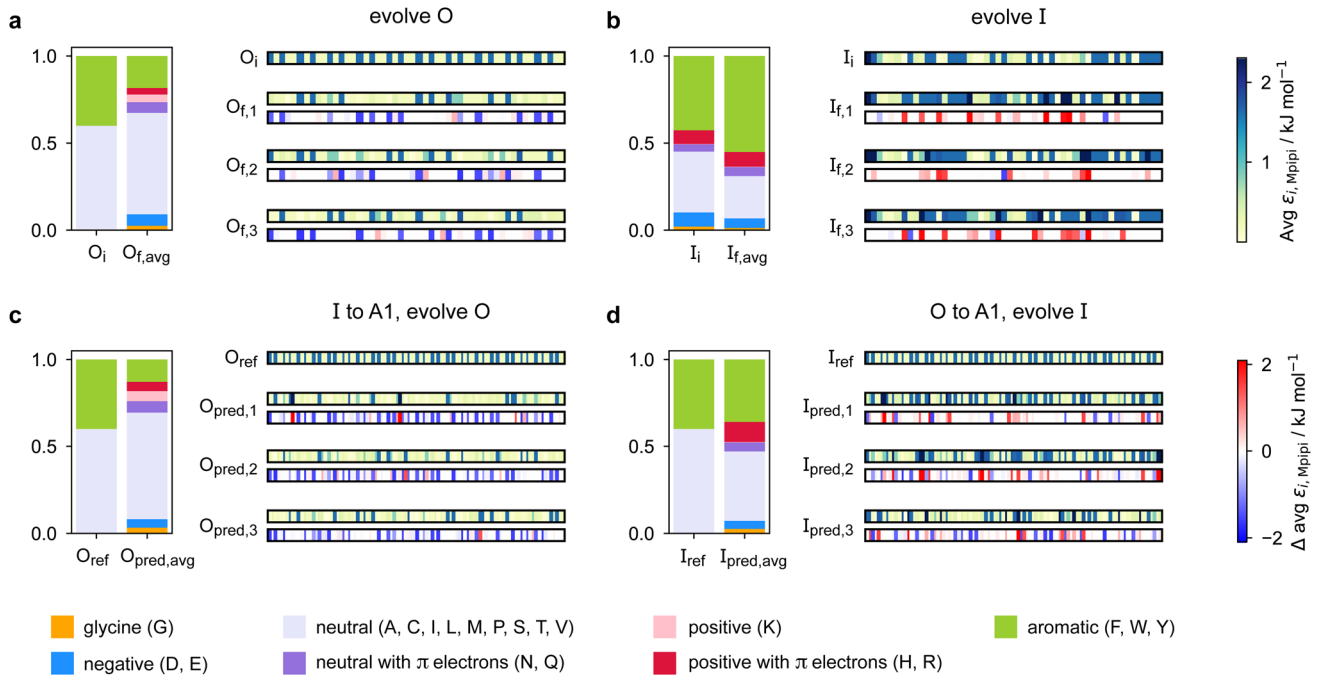

Fig. 3 Amino-acid composition and patterning on evolution. Comparison of the amino-acid composition and sequence patterning between the initial and final evolved sequences with maximum fitness in the evolution runs where (a) the outer sequence is evolved and the inner sequence is kept unchanged and (b) the inner sequence is evolved and the outer sequence is kept unchanged; and the coevolution runs where A1 LCD is designed to be (c) the inner sequence and (d) the outer sequence in the final multilayered system. For each case, we show the composition of the initial sequence and the final evolved sequence averaged across three independent runs. To illustrate the final evolved sequences in three independent runs in each case, we plot for each residue i along the sequence the absolute value of and the change in $\varepsilon_{i,\mathrm{Mpipi}}$ compared to the residue in that position in the initial sequence. The value of $\varepsilon_{i,\mathrm{Mpipi}}$ estimates the interaction strength of the residue in the coarse-grained model.

of the functional groups involved (*e.g.* π-rich, charged, or hydrophilic), rather than highly sequence-specific features.

When designing multiphase condensates that have the phase of A1 LCD proteins or its variants at the centre, we observe that the proportion of aromatic residues in the evolved partner protein decreases [Fig. 3c, S7c and S14a†]. The more strongly interacting residues are preferentially replaced by less strongly interacting ones throughout the evolved partner protein sequence. This change in composition of the evolved sequence is similar to the trend we observe in the evolution run with generic sequences where the outer sequence is evolved and the inner sequence is kept constant. The final evolved sequences in the coevolution runs with the different A1 variants are also similar in terms of the change in composition [Fig. S13a†], even though we selected variants with different features. Besides a decrease in the proportion of aromatic residues, we also observe that there is a slight upward trend in the average net charge per residue across the sequences in the population [Fig. S13b†]. The effect of increasing net charge per residue weakening the tendency to form condensates due to increasing repulsion or promoting solvation has been investigated by Bremer *et al.*[102] This may be another mechanism to decrease the stability of the condensates formed by the evolved outer protein in this case, although we note that the increase in net charge per residue is small. Altogether, the main driving force for the evolution towards increasing multiphasicity of condensates with A1 LCD at the centre is the decrease in the average interaction strength of the outer sequence. However, for the case where A1 LCD is designed to be on the outside of the multilayered condensate and the inner sequence is evolved, the proportion of aromatic residues in the final evolved sequences is similar to the initial inner sequence and it is less clear whether the residues become more or less strongly interacting throughout the sequence [Fig. 3d and S7d†]. This is not entirely surprising, since in the coevolution runs the two sequences in the system are being changed and evolved simultaneously, so we cannot necessarily expect the same trends as when only one sequence is evolved.

To rationalise why the compositional changes we observe in the evolved sequences favour multiphasicity, we compute the strengths of homotypic and heterotypic interactions between proteins forming the inner and the outer phases, and their changes throughout the evolution runs [Fig. 4]. Overall, our analyses reveal that the formation of two-component multilayered condensates depends on three crucial requirements. First, larger differences in the strengths of homotypic interactions of the different species (*i.e.* inner–inner *versus* outer–outer) favour demixing of the components into separate phases [Fig. 4d], similar to what has been observed in modelling work by Jacobs *et al.*[45] and Feric *et al.*,[38] as well as experimental work with minimal systems by Fisher *et al.*[36] Second, the proteins that establish the stronger homotypic interactions form the inner phase of the multilayered condensate. In other words, the inner–inner interaction is always the strongest, likely because such an arrangement guarantees saturation of binding sites that can form the most energetically favourable interactions. Third, the strength of heterotypic

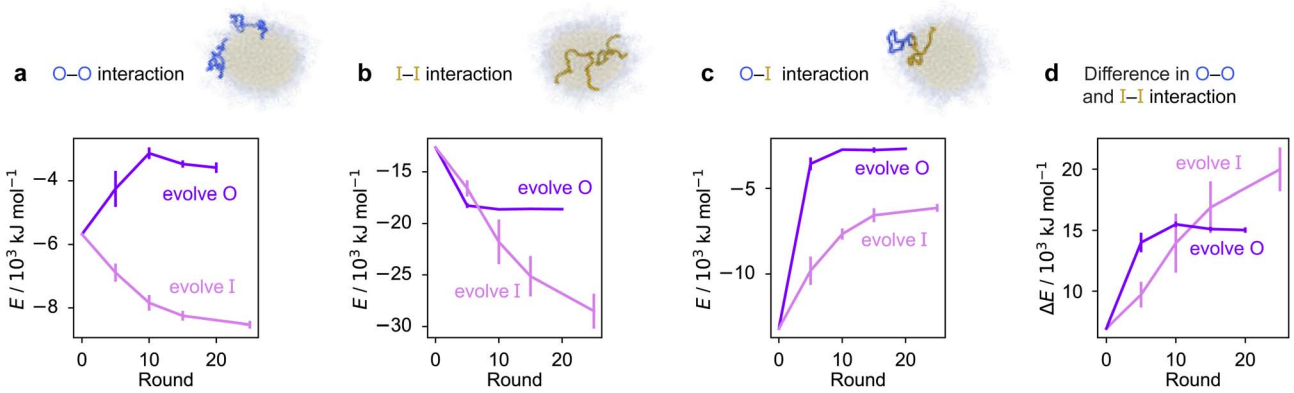

Fig. 4 Homo- and heterotypic interaction strengths control multiphasicity. The change in the interaction energies of intermolecular homotypic interactions between proteins enriched in the (a) outer and (b) inner layers, (c) heterotypic interactions between the two proteins and (d) the difference between the outer–outer and inner–inner homotypic interactions for the system with the fittest individual. Purple curves correspond to the system where the outer protein was evolved and the inner was unchanged, and pink curves correspond to the converse. Error bars correspond to the standard deviation across three independent runs. Spherical droplets are shown for clarity; we use a slab geometry in direct-coexistence simulations.

interactions should lie on a critical 'sweet spot': small enough to favour demixing into separate phases, but sufficiently large (*i.e.* comparable with the weaker outer–outer homotypic interactions) to keep the separate phases coexisting inside the same condensate.

Our evolution and co-evolution algorithms induce changes in the homotypic and heterotypic interactions across the evolution that depend on the properties of the starting condensates and which of the proteins is being evolved. For the evolution runs with generic sequences where one sequence is evolved and the other kept unchanged, our starting system exhibits relatively modest strengths for both the inner–inner and outer–outer homotypic interactions, and comparatively strong heterotypic interactions. These heterotypic interactions become weaker throughout the evolution runs when either the inner or the outer protein sequence is evolved [Fig. 4c]. The inner–inner homotypic interactions also become stronger in both cases, although the strengthening is rather more pronounced when the inner protein is being evolved [Fig. 4b]. This is expected when the two coexisting condensed phases become more pure and the two components become less well mixed.[103,104] This substantial strengthening in inner–inner interactions obtained when evolving the inner sequence indirectly results in the outer–outer interactions also becoming stronger as both phases become purer, even though the outer sequence is itself kept unchanged. By contrast, when the outer protein is evolved, the outer–outer homotypic interactions weaken even as the outer phase becomes purer [Fig. 4a]. Nevertheless, in both cases, we observe that the balance of interactions converges to the same behaviour across evolutionary runs: the difference in homotypic interactions becomes larger, while the heterotypic interactions become weaker and comparable in strength to the outer–outer interactions. The trends we observe for the changes in composition and interaction energies seem to be robust to the choice of the initial system [Fig. S8d–f and S9†]. The interaction energies in our co-evolved multiphase condensates with A1 LCD also meet the criteria we describe above [Fig. S11†]. That is, multiphasicity emerges for systems with sufficiently different homotypic interactions, where the inner–inner interactions are strongest and heterotypic interactions are small but comparable to the outer–outer interactions.

Previous work has shown that the excluded volume interactions of the residues, particularly of spacers, are important for multiphasicity.[99,105] In the evolution run with generic sequences where the inner sequence is evolved and the outer sequence is kept unchanged, we note that the average value of $\sigma_{i,\mathrm{Mpipi}}$ increases as the fitness function increases [Fig. S10†], and this may in part be related to the suggested increase in compositional demixing with increasing excluded volume. However, since only amino-acid residues are changed in the evolution runs and they all have relatively similar sizes, the changes in $\sigma_{i,\mathrm{Mpipi}}$ are much smaller than studied in previous work, and residues with larger $\sigma_{i,\mathrm{Mpipi}}$ may be favoured by the increase in interactions arising from the larger cutoff rather than the excluded volume itself.[71]

Finally, we compute the interfacial free-energy densities for the liquid–vapour interface for bulk A1 LCD and its final coevolved proteins [see the Methods section and Fig. S12†]. These results are shown in Table 1 and confirm the expectation that the protein with the largest surface tension with its vapour is most likely to be at the

Table 1 Interfacial thermodynamic parameters[a]

	γ/mJ m^{-2}	S_{int}/J m^{-2} K^{-1}	E_{int}/mJ m^{-2}
$\mathrm{I_{pred}}$	3.4(3)	17.1(1)	7.6(3)
A1 LCD	0.9(4)	9.4(7)	3.2(2)
$\mathrm{O_{pred}}$	0.03(18)	6.3(4)	1.61(8)

[a] Interfacial free-energy density γ at 250 K, interfacial entropy density S_{int} and interfacial energy density E_{int} for A1 LCD and its final coevolved proteins with maximum fitness when (i) the evolved protein is on the inside of the condensate ($\mathrm{I_{pred}}$) and A1 LCD is on the outside and (ii) the evolved protein is on the outside of the condensate ($\mathrm{O_{pred}}$) and A1 LCD is on the inside. Errors in brackets apply to the least significant digit and give the standard errors of the fitting parameters [Fig. S12].

centre[46] of the multilayered condensate. By computing the interfacial free-energy density at several different temperatures [Fig. S12†], we can extract the interfacial entropies[106] and in turn the interfacial energies; these are also shown in Table 1. The formation of the interface is energetically unfavourable; although it is in principle entropically favourable, since molecules in the liquid-like phase can gain considerable translational entropy at the interface, this contribution is relatively small. The difference in homotypic energies between species therefore also dominates the thermodynamic favourability of interface formation and determines the ordering of the layers in a multilayered condensate.

Overall, our analyses explain why residues that increase the difference in interaction strengths between the two sequences improve multiphasicity. These results support previous studies which found that multiple immiscible phases are formed when there is a sufficient difference in interaction strengths between the components in the two phases.[35,36,38,45,48,49,101]

2.4 Sequence patterning is only sometimes important for multiphasicity

The patterning of interacting amino-acid groups plays an important role in determining the phase behaviour of intrinsically disordered proteins (IDPs).[71,107] For example, the range of stability of A1 LCD condensates was shown to depend on the number and patterning of aromatic residues, which act as stickers in the 'stickers-and-spacers' framework.[107,108] More uniform distributions of stickers were found to promote the phase separation of A1 LCD and to decrease the propensity to form aggregates.[107] However, in our initial evolution runs with generic sequences and coevolution runs with A1 LCD, we have shown that the final evolved sequences in independent repeats of the same run, despite having similar overall compositions, can differ considerably in terms of the patterning of the more strongly interacting sticker residues.

To investigate the importance of the patterning of different residues in determining the degree of multiphasicity of these two-component systems, we shuffle the final evolved sequence with maximum fitness by rearranging residues of interest (*e.g.* stickers or spacers) whilst keeping the overall composition of the sequence unchanged, and compute the density profiles and fitness of shuffled sequences to examine the effect of shuffling on phase behaviour. In our analyses, we consider as stickers all the aromatic residues (F, Y, and W), the neutral residues with π electrons in the side chain (N and Q) and arginine (R), and the remaining residues as spacers.

We first do this for the final fittest systems resulting from the evolution runs with generic sequences (without A1 LCD) where one sequence is evolved and the other is kept unchanged. Unexpectedly, in these systems, when we shuffle (in multiple different ways) the evolved sequences stemming from runs where one sequence is evolved and the other is kept unchanged, the multiphasicity and hence the fitness are not notably altered [Fig. S15a and b†]. Even in the extreme cases where the evolved sequence is sorted such that the residues are rearranged in order of increasing $\varepsilon_{i,\mathrm{Mpipi}}$ values, with all the strongly interacting sticker residues clustered together at one end of the protein, the multilayered structure was still maintained, albeit with a drop in fitness indicating a lower degree of multiphasicity in some cases. This would suggest that for these sequences, the patterning of the stickers and spacers has a minimal effect on the formation of the two coexisting phases, and that it is only the overall composition of the sequence that determines whether the two proteins will mix into one homogeneous phase.

However, for the coevolution runs with A1 LCD, we do see patterning-dependent behaviour: a sorted sequence, with stickers at the ends of the protein molecules, results in rather different phase behaviour [Fig. 5a] compared to the original coevolved sequence [Fig. 2c]: in the sorted case, the sticker-rich ends interact so strongly that they form a locally crystalline structure. Interestingly, for the case where A1 LCD is at the centre of the multilayered condensate, shuffling just the positions of the spacers in the evolved sequence results in a lower degree of multiphasicity [Fig. 5b] relative to the system with the original evolved sequence [Fig. 2b]. The heterotypic interactions become stronger and the difference in homotypic interactions of the two sequences decreases after shuffling, consistent with the lower degree of

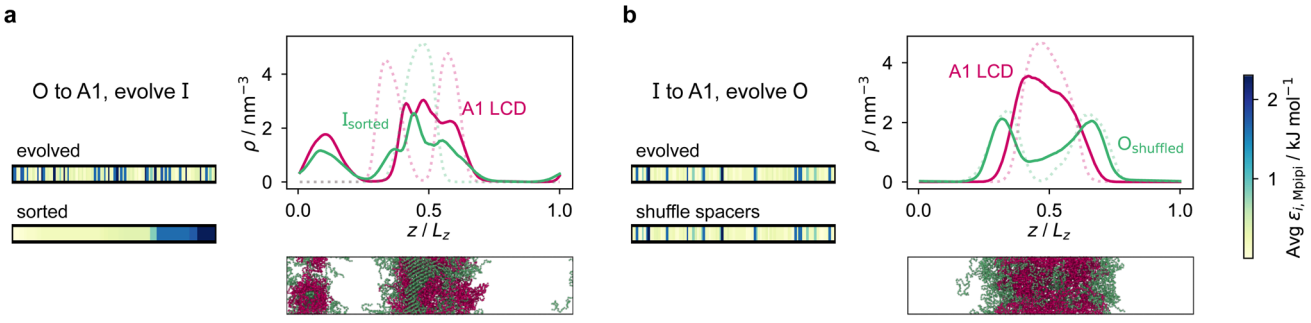

Fig. 5 Effect of sequence patterning on the formation of multilayered condensates. (a) Density profile of the final evolved system with maximum fitness in the coevolution run where A1 LCD is designed as the outer sequence, but with the residues in the final evolved sequence $\mathrm{I}_{\mathrm{pred}}$ rearranged in terms of increasing $\varepsilon_{i,\mathrm{Mpipi}}$. The sorted sequence with all the strongly interacting residues clustered at one end results in rather different phase behaviour. (b) Density profile of the final evolved system with maximum fitness in the coevolution run where A1 LCD is designed as the inner sequence, but with spacer residues (see the text) in the final evolved sequence $\mathrm{O}_{\mathrm{pred}}$ shuffled randomly. The sequence with the spacers shuffled resulted in a decrease in multiphasicity compared to that of the original evolved sequence predicted by the coevolution approach. Dotted lines represent the density profile of the original evolved system with the unshuffled partner sequence.

multiphasicity observed. The fact that different spacers do not give rise to identical phase behaviour has recently been investigated by Bremer *et al.*,[102] and similarly, in this case, we find that it is not just the distribution of stickers *versus* spacers that is important, but also both the identity of the spacers and their arrangement along the sequence.

Although in some cases patterning of amino-acid residues does not affect the phase behaviour much, it does in others. It would be helpful to anticipate the conditions where patterning is likely to be important. Our tests suggest that if the partner protein's sequence is repetitive with low compositional diversity, the relevant interactions can occur anywhere along the chain, reducing the need for a particular patterning of interactions to maintain phase separation. For example, for the case with generic sequences where we evolve the outer sequence and the inner protein is $I = (FAFAA)_{10}$, which is highly repetitive, the multiphasicity is unaffected by shuffling or sorting the outer protein [Fig. S15a†]; however, if we sort the inner sequence to give $I = F_{20}A_{30}$, thereby removing the repetition while maintaining the overall composition, this results in a substantial loss of multiphasicity [Fig. S16†]. By contrast, the protein partner of the analogue where the inner sequence is evolved is not especially repetitive [see protein I_i of Fig. 3b], but patterning is nevertheless not especially important [Fig. S15b†]. Another obvious difference between the sequences investigated is their length, and it may appear that with shorter proteins (such as those investigated in Fig. S15a and b†), all relevant residues are spatially sufficiently close that the same interactions dominate irrespective of their precise position in the sequence. However, the behaviour of systems where A1 LCD is partnered with a 50-residue strand (*cf.* Fig. S2†) is almost identical to the case of 100-residue strands shown in Fig. 5, S15c and d,† suggesting that the difference in length alone is not sufficient to rationalise the difference in behaviour.

The phase behaviour of IDPs has been observed to be affected by the patterning of not only aromatic residues, but also charged ones.[99,109–112] The role of charge patterning has been investigated with simpler polymer models containing charged segments in the context of demixing in both one-component[111,113] and two-component[97–99] systems. In the latter, a large mismatch in charge patterning between the two sequences has been found to favour compositional demixing.[97–99] We investigate the effect of charge patterning mismatch on the degree of multiphasicity in the shuffled systems with the system where shuffling amino-acid residues resulted in the largest variation in phase behaviour, namely the system where A1 LCD is concentrated at the centre. We show the variation of the sequence charge decoration (SCD), an order parameter quantifying charge patterning [see Section S10], in Fig. S17;† however, we observe little correlation between the mismatch in charge patterning and the fitness. The reason for the insensitivity to charge patterning in this case may simply be that phase separation is not principally charge-driven in the systems we have considered, compared to previous work where the polymers considered were entirely made up of charged residues. We expect that, if the proportion of charged residues in the relevant proteins was larger, there may be a stronger correlation between demixing (as quantified by the fitness function) and the charge pattern mismatch.

Since it is difficult to know *a priori* when the patterning of residues is likely to affect the phase behaviour of multilayered condensates, the use of a genetic algorithm and the coevolution approach as a predictive tool, where any relevant patterning is optimised alongside the interaction strengths, is especially attractive.

3 Discussion

We have developed a computational approach to design multi-component multilayered condensates that contain a target protein of interest. Our approach integrates a genetic algorithm, anchored in an innovative fitness function for automated evolution of multiphasicity, with our near-quantitative residue-resolution coarse-grained protein model, Mpipi.[84] We demonstrate the utility of our approach in a biological context by applying it to predict different protein partners capable of forming two-component multiphase condensates when mixed with A1 LCD or its variants. We show that our method can be adapted to produce condensates that concentrate the protein of interest (*e.g.* A1 LCD) either at the centre of the multilayered condensate or in the outer layer, as desired.

In addition to enabling the design of multiphase condensates, our approach helps uncover the biophysical mechanisms that drive the formation of complex multilayered organisations. In all cases, we find that multiphasicity in multi-component protein systems is favoured if the difference between homotypic interactions among different components is large, and the strength of heterotypic interactions is small but comparable with that of the weaker homotypic interactions in the mixture. In a two-component system, proteins that establish stronger homotypic interactions are concentrated at the core of the multiphase condensate, as saturating their bonds enhances the overall enthalpic gain for condensate formation. Similarly, the outer layer of the multiphase condensate is formed by the proteins that establish weaker homotypic interactions, as this reduces the overall interfacial free-energy density of the two-component system. Although the specific predicted partner sequences can differ across independent (co) evolution runs and several sequences can have similar fitness values, these general trends in interaction energies remain consistent in the cases we have tested, suggesting that the rules we have identified may be universal in driving the formation of multiphase condensates. The diversity of partner sequences obtained may also suggest that a given protein can form multiphase condensates with a wide range of different partners, rather than exclusively with a unique complementary sequence: multiphasic organisation thus appears to be a robust property of multi-component condensates.

Since the genetic algorithm is coupled to a residue-resolution coarse-grained model for proteins, the accuracy of our predictions is contingent on that of the model. Reassuringly, we have previously demonstrated that Mpipi reproduces the experimental phase diagrams of A1 LCD and its variants with near-quantitative accuracy, achieves excellent agreement with experiments probing the phase behaviour of naturally occurring proteins (*i.e.* FUS, Ddx4, LAF-1 and their variants) and of polyR/polyK/polyU mixtures, and predicts the experimental radii of gyration of a large set of IDPs

with high accuracy.[84] We are therefore hopeful that the predictions of our approach will be robust against experimental validation, which is an essential next step. An important factor to consider for such validation is that the specific amino-acid sequences we predict to from multiphase condensates are only applicable at the fixed temperature and salt concentration at which the simulations are run; however, these could of course be changed to design multilayered condensates that are stable under different conditions. One possible limitation of using residue-resolution coarse-grained models like Mpipi is that they are typically unable to consider the emergence of secondary or tertiary structural transitions from specific changes in amino-acid sequence. In this regard, our evolutionary approach is flexible enough to incorporate knowledge-based constraints to bypass selected patterns of amino-acid sequences known to favour, for instance, the folding of protein regions into α-helices or β-sheets in specific contexts, or limit the number of certain residues such as cysteine, which forms disulphide bridges. Our approach can also easily be modified to consider special requirements for each protein system by introducing further constraints in the algorithm: for instance to introduce tailored replacement probabilities and outcomes for different residues (*e.g.* to limit mutations to stickers, only allow mutations of charged to charged residues, or enforce a given pattern of aromatics) or protein regions (*e.g.* to avoid mutations at the N-terminus or to favour the concentration of aromatics at the centre) when proposing mutations.

While we have investigated multiphase condensates comprising only of two protein components, our evolutionary approach is transferable to multi-component systems with a larger number of components, and can also easily be extended to study the effect of RNA or post-translational modifications. In turn, our method expands the repertoire of tools available to gain molecular insight into LLPS in complex biological cellular functions. Our approach therefore presents new opportunities for designing multilayered condensates, probing more closely the underlying physicochemical factors that lead to their formation and, ultimately, deciphering the missing links to their function inside cells.

4 Methods

4.1 Genetic algorithm and fitness function

The basic framework of our genetic-algorithm approach is very simple:

(1) Choose an initial system with a degree of multiphasicity, such as one of the minimal systems we have outlined.

(2) Choose a target protein for which to design a partner.

(3) Decide whether the target protein should be on the inside or the outside of the multiphase condensate.

(4) Run a genetic algorithm on one protein and systematically change the other protein to the target protein.

In our implementation of the genetic algorithm, we maintain a population of 20 sequences in each round, alongside a partner protein sequence that is not being evolved. To generate the initial population, we mutate the initial sequence by replacing, with 0.05 probability, each residue with a new one chosen from the 20 canonical amino acids with uniform probability. Each individual sequence $\mathbf{x}$ is assessed with a fitness function,

$$f(\boldsymbol{x}) = \left|\rho^*_{\mathrm{A,\,centre}}(\boldsymbol{x}) - \rho^*_{\mathrm{B,\,centre}}(\boldsymbol{x})\right| - s\left(\rho^*_{\mathrm{A,vapour}}(\boldsymbol{x}) + \rho^*_{\mathrm{B,vapour}}(\boldsymbol{x})\right), \quad (1)$$

where $\rho^*_{\mathrm{A}}(\boldsymbol{x})$ and $\rho^*_{\mathrm{B}}(\boldsymbol{x})$ are the average dimensionless number densities of the two different protein sequences A and B in the two-component mixtures. $\rho^*_{\mathrm{A,centre}}(\boldsymbol{x})$ and $\rho^*_{\mathrm{A,vapour}}(\boldsymbol{x})$ denote the number density of protein A in the core of the multilayered condensate and in the dilute phase respectively, with analogous expressions for protein B [Fig. 1a]. Details of how these regions are determined are discussed in Section S1. The number densities are non-dimensionalised by dividing them with an appropriate unit, *e.g.* $\rho^* = \rho/\mathrm{nm}^{-3}$, although the choice of unit is immaterial, since only the relative ordering in fitness is important, not the absolute numerical value. Finally, as discussed in the main text, the parameter s determines the trade-off between obtaining two stable liquid-like phases to give a stable multilayered condensate and the difference in compositions between the two phases. The value of s can be tuned as necessary depending on the specific system of interest; here, we have used $s = 0$, 0.5, 1 and 5.

The fitness function we have introduced is facile to compute and works well even in relatively small systems, which is especially useful in genetic-algorithm simulations where the fitness must be evaluated for many systems. Other order parameters have also been shown to identify compositional demixing within the condensed phase well, such as by quantifying the compositional asymmetry by considering the fraction of the two sequences in the two coexisting phases.[98] However, in our simulations of relatively small system sizes, the structure and arrangement of the coexisting phases introduce extra complexity and it is difficult to determine accurately the density of the two components, especially in the outer layer of the condensate. Calculating the compositional asymmetry in this way would therefore likely not be straightforward enough to be used routinely in the genetic algorithm. Another approach uses intra- and inter-species pair correlation functions with demixing characterised by the intra-species pair correlation function dominating at small separations compared to the inter-species one.[99,100] We compute the pair correlation functions for a range of our simulation outputs and show in Fig. S18† that this approach is consistent with the simpler fitness function we outlined above.

Once the fitness of each individual is determined, we use the tournament selection algorithm[52,114] to select the fittest parents to cross over. We also apply a round of random mutations to explore previously unsampled regions of sequence space. Finally, we use a weak population replacement scheme to generate the population of sequences for the next round of the genetic-algorithm run. Our genetic-algorithm implementation is detailed in full in ref. 71.

The fitness of each individual in the population is computed when it is first encountered, *e.g.* following a mutation or crossover; however, when a systematic change is made to the partner sequence in coevolution runs, this too affects the phase behaviour and the fitness of all individuals in the population must therefore be recalculated. In our coevolution runs, we do this at 5-round intervals, at which we change ∼10% of the residues of this partner

protein to the target protein sequence, with residues to be changed chosen randomly from those that have not yet been changed with uniform probability.

4.2 Simulation details

To simulate protein chains, we use the Mpipi residue-resolution sequence-specific coarse-grained model,[84] combining (i) harmonic covalent bonds between residues, (ii) the Wang–Frenkel potential[115] to account for non-bonded interactions between amino acids and (iii) Debye–Hückel electrostatic interactions.[116] The Mpipi potential was shown to model LLPS of intrinsically disordered proteins well.[84] For each amino-acid pair ij, the Mpipi model defines a Wang–Frenkel well-depth (ε_{ij}), a characteristic length scale (σ_{ij}) and values for ν and μ that determine the steepness of the potential well.

We use direct-coexistence simulations[117,118] to model the vapour phase alongside the condensed phases in the same elongated tetragonal simulation box with explicit interfaces between phases. We use LAMMPS[119] to run molecular-dynamics simulations with a typical time step of 10 fs and a coupling to a Langevin thermostat with a relaxation time of 10 ps. To estimate the densities of the dilute and condensed phases, we run each simulation for 40 ns for equilibration and an additional 20 ns to compute the densities. Each simulation takes around an hour on 76 CPU cores.

We use 96 chains of each protein for the evolution runs with generic sequences with a box size of 11.4 nm × 11.4 nm × 56.9 nm. For coevolution runs, we use 45 chains of the protein that is changed to A1 LCD, and either 90 chains of 50 residues or 45 chains of 100 residues of the other protein, in a box of size 10.9 nm × 10.9 nm × 54.7 nm. Although finite-size effects were examined in ref. 84 with similar-sized systems, we simulate several systems obtained in coevolution runs with A1 LCD concentrated in the inner layer at larger system sizes to check that the density profiles are consistent. We do this separately for a system where the final evolved system is highly multiphasic and the two condensed phases are essentially pure [Fig. S21†], and a system where there is still a considerable degree of mixing of the two proteins in the condensed phases [Fig. S22†]. The results from increasing system sizes appear to suggest that the outer phase is not merely wetting the surface of the inner phase but forms a layer that scales with the system size and is likely to be a genuine immiscible phase. Multiphasic biomolecular condensates reported in the experimental literature are often relatively small, and so may be stabilised by interfacial considerations[120] rather than by forming truly immiscible bulk phases. If these phases are true thermodynamic phases, although they may of course have a preferred ordering in direct-coexistence simulations because of interfacial free-energy considerations, each of the three phases in question should be able to coexist independently with any one of the others under the same thermodynamic conditions. We test whether this holds for representative systems with different underlying multiphasic behaviour and confirm that each of the condensed phases coexists with the vapour-like phase under the same conditions as in the multiphasic regime [Fig. S23†]. For the systems investigated, as previously implied by finite-size scaling, these therefore appear to be genuine thermodynamic phases. Finally, we also perform a sanity check by verifying that if we double the system size, the ordering of fitness values is the same [Fig. S24†].

To ensure that our predictions are robust, we have confirmed that the final predicted sequences obtained from genetic-algorithm runs with the Mpipi potential exhibit similar multiphasic behaviour when simulated with another coarse-grained potential, namely Model 2 of ref. 87.

4.3 Interfacial free-energy densities

We compute interfacial free-energy densities for the interface between the vapour-like phase and the pure condensed phase for the final evolved maximum-fitness sequences in coevolution runs with A1 LCD for both cases, *i.e.* where the coevolved protein is the inner or the outer protein. To do this, we use the Kirkwood–Buff expression[121] to relate the interfacial free-energy density γ to the normal and tangential components of the pressure tensor, and the mean value theorem to simplify the result[122] for planar interfaces into $\gamma = (L_z/2)(P_{\mathrm{norm}} - P_{\mathrm{tang}})$, where γ is the interfacial free-energy density, L_z is the length of the simulation box along which the interface occurs, $P_{\mathrm{norm}} = P_{zz}$ is normal to the interface and $P_{\mathrm{tang}} = P_{xx} = P_{yy}$ is the tangential pressure, and the division by 2 accounts for the fact that there are two interfaces in our simulation set-up.[120] Although the pressure tensor has many possible definitions, from virial to mechanical expressions, the interfacial-free energy density is independent of this arbitrary choice;[123] we use the atomic virial pressure tensor in our calculation, which gives the same results as the molecular virial.[124]

We compute only the interfacial free-energy density for the interface between the dense and dilute phases (*i.e.* a surface tension using our coarse-grained model), since as long as a multilayered condensate forms, the interface between the two condensed phases of different compositions is always present and therefore does not affect the thermodynamics. We assume for simplicity that the resulting phases are pure and, in this back-of-the-envelope calculation, we do not consider the possible dependence of γ on the interface width or the curvature of the droplet. We determine the interfacial free-energy density for each system at several temperatures. The resulting data are well fitted with a linear function [Fig. S12a†]; since $(\partial\gamma/\partial T) = -S_{\mathrm{int}}$, the interfacial entropy,[106] this approach allows us to extract the interfacial energy and interfacial entropy for each component, as discussed in the main text. We test for finite-size effects in the interfacial free-energy density as a function of both the surface area and the bulk depth and see that the values are the same within error bars across different system sizes [Fig. S12b†].

Data availability

All relevant data are within the manuscript, its ESI† files and the Figshare data repository at **https://doi.org/10.6084/m9.figshare.21926154**.

Code availability

LAMMPS input scripts are available in the Figshare data repository at **https://doi.org/10.6084/m9.figshare.21926154**.

Author contributions

PYC, JAJ, RC-G and AR designed the research. PYC performed the research. PYC, JAJ, RC-G and AR analysed the results and wrote the paper.

Conflicts of interest

The authors declare no competing interests.

Acknowledgements

We acknowledge funding from the University of Cambridge Ernest Oppenheimer Fund [PYC], the Winton Programme for the Physics of Sustainability [PYC, RC-G], 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. 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 [RC-G, JAJ, AR]. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

References

1 F. Wippich, B. Bodenmiller, M. G. Trajkovska, S. Wanka, R. Aebersold and L. Pelkmans, Dual specificity kinase DYRK3 couples stress granule condensation/dissolution to mTORC1 signaling, *Cell*, 2013, **152**, 791–805.

2 T. Takahara and T. Maeda, Transient sequestration of TORC1 into stress granules during heat stress, *Mol. Cell*, 2012, **47**, 242–252.

3 S. Koga, D. S. Williams, A. W. Perriman and S. Mann, Peptide-nucleotide microdroplets as a step towards a membrane-free protocell model, *Nat. Chem.*, 2011, **3**, 720–724.

4 O. A. Saleh, B. J. Jeon and T. Liedl, Enzymatic degradation of liquid droplets of DNA is modulated near the phase boundary, *Proc. Natl. Acad. Sci. U. S. A.*, 2020, **117**, 16160–16166.

5 T. Schwarz-Romond, C. Merrifield, B. J. Nichols and M. Bienz, The Wnt signalling effector Dishevelled forms dynamic protein assemblies rather than stable associations with cytoplasmic vesicles, *J. Cell Sci.*, 2005, **118**, 5269–5277.

6 A. Grakoui, S. K. Bromley, C. Sumen, M. M. Davis, A. S. Shaw, P. M. Allen and M. L. Dustin, The immunological synapse: a molecular machine controlling T cell activation, *Science*, 1999, **285**, 221–227.

7 L. B. Case, X. Zhang, J. A. Ditlev and M. K. Rosen, Stoichiometry controls activity of phase-separated clusters of actin signaling proteins, *Science*, 2019, **363**, 1093–1097.

8 X. Su, J. A. Ditlev, E. Hui, W. Xing, S. Banjade, J. Okrut, D. S. King, J. Taunton, M. K. Rosen and R. D. Vale, Phase separation of signaling molecules promotes T cell receptor signal transduction, *Science*, 2016, **352**, 595–599.

9 P. Li, S. Banjade, H. C. Cheng, S. Kim, B. Chen, L. Guo, M. Llaguno, J. V. Hollingsworth, D. S. King, S. F. Banani, P. S. Russo, Q. X. Jiang, B. T. Nixon and M. K. Rosen, Phase transitions in the assembly of multivalent signalling proteins, *Nature*, 2012, **483**, 336–340.

10 F. M. Boisvert, S. Van Koningsbruggen, J. Navascués and A. I. Lamond, The multifunctional nucleolus, *Nat. Rev. Mol. Cell Biol.*, 2007, **8**, 574–585.

11 R. J. Ries, S. Zaccara, P. Klein, A. Olarerin-George, S. Namkoong, B. F. Pickering, D. P. Patil, H. Kwak, J. H. Lee and S. R. Jaffrey, m6A enhances the phase separation potential of mRNA, *Nature*, 2019, **571**, 424–428.

12 D. Updike and S. Strome, P granule assembly and function in Caenorhabditis elegans germ cells, *J. Androl.*, 2010, **31**, 53–60.

13 S. Jain, J. R. Wheeler, R. W. Walters, A. Agrawal, A. Barsic and R. Parker, ATPase-modulated stress granules contain a diverse proteome and substructure, *Cell*, 2016, **164**, 487–498.

14 A. Khong, T. Matheny, S. Jain, S. F. Mitchell, J. R. Wheeler and R. Parker, The Stress Granule Transcriptome Reveals Principles of mRNA Accumulation in Stress Granules, *Mol. Cell*, 2017, **68**, 808–820.

15 S. Markmiller, S. Soltanieh, K. L. Server, R. Mak, W. Jin, M. Y. Fang, E. C. Luo, F. Krach, D. Yang, A. Sen, A. Fulzele, J. M. Wozniak, D. J. Gonzalez, M. W. Kankel, F. B. Gao, E. J. Bennett, E. Lécuyer and G. W. Yeo, Context-Dependent and Disease-Specific Diversity in Protein Interactions within Stress Granules, *Cell*, 2018, **172**, 590–604.

16 T. M. Franzmann, M. Jahnel, A. Pozniakovsky, J. Mahamid, A. S. Holehouse, E. Nüske, D. Richter, W. Baumeister, S. W. Grill, R. V. Pappu, A. A. Hyman and S. Alberti, Phase separation of a yeast prion protein promotes cellular fitness, *Science*, 2018, **359**, eaao5654.

17 J. A. Riback, C. D. Katanski, J. L. Kear-Scott, E. V. Pilipenko, A. E. Rojek, T. R. Sosnick and D. A. Drummond, Stress-triggered phase separation is an adaptive, evolutionarily tuned response, *Cell*, 2017, **168**, 1028–1040.

18 A. Boija, I. A. Klein, B. R. Sabari, A. Dall'Agnese, E. L. Coffey, A. V. Zamudio, C. H. Li, K. Shrinivas, J. C. Manteiga, N. M. Hannett, B. J. Abraham, L. K. Afeyan, Y. E. Guo, J. K. Rimel, C. B. Fant, J. Schuijers, T. I. Lee, D. J. Taatjes and R. A. Young, Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains, *Cell*, 2018, **175**, 1842–1855.

19 D. Hnisz, K. Shrinivas, R. A. Young, A. K. Chakraborty and P. A. Sharp, A Phase Separation Model for Transcriptional Control, *Cell*, 2017, **169**, 13–23.

20 K. L. Zobeck, M. S. Buckley, W. R. Zipfel and J. T. Lis, Recruitment timing and dynamics of transcription factors at the Hsp70 loci in living cells, *Mol. Cell*, 2010, **40**, 965–975.
21 B. R. Sabari, A. Dall'Agnese, A. Boija, I. A. Klein, E. L. Coffey, K. Shrinivas, B. J. Abraham, N. M. Hannett, A. V. Zamudio, J. C. Manteiga, C. H. Li, Y. E. Guo, D. S. Day, J. Schuijers, E. Vasile, S. Malik, D. Hnisz, T. I. Lee, I. I. Cisse, R. G. Roeder, P. A. Sharp, A. K. Chakraborty and R. A. Young, Coactivator condensation at super-enhancers links phase separation and gene control, *Science*, 2018, **361**, eaar3958.
22 A. G. Larson, D. Elnatan, M. M. Keenen, M. J. Trnka, J. B. Johnston, A. L. Bdoiingame, D. A. Agard, S. Redding and G. J. Narlikar, Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin, *Nature*, 2017, **547**, 236–240.
23 A. R. Strom, A. V. Emelyanov, M. Mir, D. V. Fyodorov, X. Darzacq and G. H. Karpen, Phase separation drives heterochromatin domain formation, *Nature*, 2017, **547**, 241–245.
24 F. Pessina, U. Gioia, O. Brandi, S. Farina, M. Ceccon, S. Francia and F. d'Adda di Fagagna, DNA damage triggers a new phase in neurodegeneration, *Trends Genet.*, 2021, **37**, 337–354.
25 R. Oshidari, R. Huang, M. Medghalchi, E. Y. Tse, N. Ashgriz, H. O. Lee, H. Wyatt and K. Mekhail, DNA repair by Rad52 liquid droplets, *Nat. Commun.*, 2020, **11**, 695.
26 G. M. Harami, Z. J. Kovács, R. Pancsa, J. Pálinkás, V. Baráth, K. Tárnok, A. Málnási-Csizmadia and M. Kovács, Phase separation by ssDNA binding protein controlled *via* protein-protein and protein–DNA interactions, *Proc. Natl. Acad. Sci. U. S. A.*, 2020, **117**, 26206–26217.
27 S. Alberti and D. Dormann, Liquid–Liquid Phase Separation in Disease, *Annu. Rev. Genet.*, 2019, **53**, 171–194.
28 E. Chuang, A. M. Hori, C. D. Hesketh and J. Shorter, Amyloid assembly and disassembly, *J. Cell Sci.*, 2018, **131**, jcs189928.
29 N. B. Nedelsky and J. P. Taylor, Bridging biophysics and neurology: aberrant phase transitions in neurodegenerative disease, *Nat. Rev. Neurol.*, 2019, **15**, 272–286.
30 A. Molliex, J. Temirov, J. Lee, M. Coughlin, A. P. Kanagaraj, H. J. Kim, T. Mittag and J. P. Taylor, Phase Separation by Low Complexity Domains Promotes Stress Granule Assembly and Drives Pathological Fibrillization, *Cell*, 2015, **163**, 123–133.
31 A. Siegert, M. Rankovic, F. Favretto, T. Ukmar-Godec, T. Strohäker, S. Becker and M. Zweckstetter, Interplay between tau and α-synuclein liquid–liquid phase separation, *Protein Sci.*, 2021, **30**, 1326–1336.
32 N. M. Kanaan, C. Hamel, T. Grabinski and B. Combs, Liquid–liquid phase separation induces pathogenic tau conformations *in vitro*, *Nat. Commun.*, 2020, **11**, 2809.
33 I. A. Klein, A. Boija, L. K. Afeyan, S. W. Hawken, M. Fan, A. Dall'Agnese, O. Oksuz, J. E. Henninger, K. Shrinivas, B. R. Sabari, I. Sagi, V. E. Clark, J. M. Platt, M. Kar, P. M. McCall, A. V. Zamudio, J. C. Manteiga, E. L. Coffey, C. H. Li, N. M. Hannett, Y. E. Guo, T. M. Decker, T. I. Lee, T. Zhang, J. K. Weng, D. J. Taatjes, A. Chakraborty, P. A. Sharp, Y. T. Chang, A. A. Hyman, N. S. Gray and R. A. Young, Partitioning of cancer therapeutics in nuclear condensates, *Science*, 2020, **368**, 1386–1392.
34 W. M. Babinchak, B. K. Dumm, S. Venus, S. Boyko, A. A. Putnam, E. Jankowsky and W. K. Surewicz, Small molecules as potent biphasic modulators of protein liquid–liquid phase separation, *Nat. Commun.*, 2020, **11**, 5574.
35 S. Boeynaems, A. S. Holehouse, V. Weinhardt, D. Kovacs, J. Van Lindt, C. Larabell, L. V. D. Bosch, R. Das, P. S. Tompa, R. V. Pappu and A. D. Gitler, Spontaneous driving forces give rise to protein–RNA condensates with coexisting phases and complex material properties, *Proc. Natl. Acad. Sci. U. S. A.*, 2019, **116**, 7889–7898.
36 R. S. Fisher and S. Elbaum-Garfinkle, Tunable multiphase dynamics of arginine and lysine liquid condensates, *Nat. Commun.*, 2020, **11**, 4628.
37 T. Kaur, M. Raju, I. Alshareedah, R. B. Davis, D. A. Potoyan and P. R. Banerjee, Sequence-encoded and composition-dependent protein-RNA interactions control multiphasic condensate morphologies, *Nat. Commun.*, 2021, **12**, 872.
38 M. Feric, N. Vaidya, T. S. Harmon, D. M. Mitrea, L. Zhu, T. M. Richardson, R. W. Kriwacki, R. V. Pappu and C. P. Brangwynne, Coexisting Liquid Phases Underlie Nucleolar Subcompartments, *Cell*, 2016, **165**, 1686–1697.
39 J. T. Wang, J. Smith, B. C. Chen, H. Schmidt, D. Rasoloson, A. Paix, B. G. Lambrus, D. Calidas, E. Betzig and G. Seydoux, Regulation of RNA granule dynamics by phosphorylation of serine-rich, intrinsically disordered proteins in *C. elegans*, *eLife*, 2014, **3**, e04591.
40 A. Hubstenberger, S. L. Noble, C. Cameron and T. C. Evans, Translation repressors, an RNA helicase, and developmental cues control RNP phase transitions during early development, *Dev. Cell*, 2013, **27**, 161–173.
41 U. Sheth, J. Pitt, S. Dennis and J. R. Priess, Perinuclear P granules are the principal sites of mRNA export in adult *C. elegans* germ cells, *Development*, 2010, **137**, 1305–1314.
42 A. S. Holehouse and R. V. Pappu, Functional Implications of Intracellular Phase Transitions, *Biochemistry*, 2018, **57**, 2415–2423.
43 I. A. Sawyer, D. Sturgill and M. Dundr, Membraneless nuclear organelles and the search for phases within phases, *Wiley Interdiscip. Rev.: RNA*, 2019, **10**, e1514.
44 W. M. Jacobs and D. Frenkel, Predicting phase behavior in multicomponent mixtures, *J. Chem. Phys.*, 2013, **139**, 024108.
45 W. M. Jacobs and D. Frenkel, Phase Transitions in Biological Systems with Many Components, *Biophys. J.*, 2017, **112**, 683–691.
46 D. C. Sundberg, A. P. Casassa, J. Pantazopoulos, M. R. Muscato, B. Kronberg and J. Berg, Morphology development of polymeric microparticles in aqueous dispersions. I. Thermodynamic considerations, *J. Appl. Polym. Sci.*, 1990, **41**, 1425–1442.

47 L. D. Zarzar, V. Sresht, E. M. Sletten, J. A. Kalow, D. Blankschtein and T. M. Swager, Dynamically reconfigurable complex emulsions *via* tunable interfacial tensions, *Nature*, 2015, **518**, 520–524.
48 G. A. Mountain and C. D. Keating, Formation of Multiphase Complex Coacervates and Partitioning of Biomolecules within them, *Biomacromolecules*, 2020, **21**, 630–640.
49 T. Lu and E. Spruijt, Multiphase Complex Coacervate Droplets, *J. Am. Chem. Soc.*, 2020, **142**, 2905–2914.
50 N. N. Deng, W. Wang, X. J. Ju, R. Xie, D. A. Weitz and L. Y. Chu, Wetting-induced formation of controllable monodisperse multiple emulsions in microfluidics, *Lab Chip*, 2013, **13**, 4047–4052.
51 Y. Shin and C. P. Brangwynne, Liquid phase condensation in cell physiology and disease, *Science*, 2017, **357**, eaaf4382.
52 L. Chambers, *Practical handbook of genetic algorithms: applications*, CRC Press Inc., 1995.
53 M. Mitchell, *An Introduction to Genetic Algorithms*, MIT Press, 1998.
54 J. H. Holland, *Adaptation in Natural and Artificial Systems*, University of Michigan Press, 1975.
55 R. S. Judson and H. Rabitz, Teaching lasers to control molecules, *Phys. Rev. Lett.*, 1992, **68**, 1500–1503.
56 A. Assion, T. Baumert, M. Bergt, T. Brixner, B. Kiefer, V. Seyfried, M. Strehle and G. Gerber, Control of chemical reactions by feedback-optimized phase-shaped femtosecond laser pulses, *Science*, 1998, **282**, 919–922.
57 D. M. Deaven and K. M. Ho, Molecular geometry optimization with a genetic algorithm, *Phys. Rev. Lett.*, 1995, **75**, 288–291.
58 S. M. Woodley, P. D. Battle, J. D. Gale and C. R. A. Catlow, The prediction of inorganic crystal structures using a genetic algorithm and energy minimisation, *Phys. Chem. Chem. Phys.*, 1999, **1**, 2535–2542.
59 G. J. Pauschenwein and G. Kahl, Clusters, columns, and lamellae—minimum energy configurations in core softened potentials, *Soft Matter*, 2008, **4**, 1396–1399.
60 J. Fornleitner and G. Kahl, Lane formation *vs.* cluster formation in two-dimensional square-shoulder systems — A genetic algorithm approach, *Europhys. Lett.*, 2008, **82**, 18001.
61 L. Filion and M. Dijkstra, Prediction of binary hard-sphere crystal structures, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.*, 2009, **79**, 046714.
62 I. G. Johnston, S. E. Ahnert, J. P. K. Doye and A. A. Louis, Evolutionary dynamics in a simple model of self-assembly, *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.*, 2011, **83**, 066105.
63 M. Z. Miskin and H. M. Jaeger, Adapting granular materials through artificial evolution, *Nat. Mater.*, 2013, **12**, 326–331.
64 J. C. Forster, J. Krausser, M. R. Vuyyuru, B. Baum and A. Šarić, Exploring the design rules for efficient membrane-reshaping nanostructures, *Phys. Rev. Lett.*, 2020, **125**, 228101.
65 M. G. Wessels and A. Jayaraman, Machine learning enhanced computational reverse engineering analysis for scattering experiments (CREASE) to determine structures in amphiphilic polymer solutions, *ACS Polym. Au*, 2021, **1**, 153–164.
66 T. Dandekar and P. Argos, Potential of genetic algorithms in protein folding and protein engineering simulations, *Protein Eng., Des. Sel.*, 1992, **5**, 637–645.
67 R. Unger and J. Moult, Genetic algorithms for protein folding simulations, *J. Mol. Biol.*, 1993, **231**, 75–81.
68 W. P. Stemmer, Rapid evolution of a protein *in vitro* by DNA shuffling, *Nature*, 1994, **370**, 389–391.
69 R. V. Devi, S. S. Sathya and M. S. Coumar, Evolutionary algorithms for *de novo* drug design – A survey, *Appl. Soft Comput.*, 2015, **27**, 543–552.
70 J. O. Spiegel and J. D. Durrant, AutoGrow4: An open-source genetic algorithm for *de novo* drug design and lead optimization, *J. Cheminf.*, 2020, **12**, 25.
71 S. M. Lichtinger, A. Garaizar, R. Collepardo-Guevara and A. Reinhardt, Targeted modulation of protein liquid–liquid phase separation by evolution of amino-acid sequence, *PLoS Comput. Biol.*, 2021, **17**, e1009328.
72 X. Zeng, C. Liu, M. J. Fossat, P. Ren, A. Chilkoti and R. V. Pappu, Design of intrinsically disordered proteins that undergo phase transitions with lower critical solution temperatures, *APL Mater.*, 2021, **9**, 021119.
73 G. L. Dignon, R. B. Best and J. Mittal, Biomolecular phase separation: From molecular driving forces to macroscopic properties, *Annu. Rev. Phys. Chem.*, 2020, **71**, 53–75.
74 W. Zheng, G. L. Dignon, N. Jovic, X. Xu, R. M. Regy, N. L. Fawzi, Y. C. Kim, R. B. Best and J. Mittal, Molecular details of protein condensates probed by microsecond long atomistic simulations, *J. Phys. Chem. B*, 2020, **124**, 11671–11679.
75 Z. Jing, C. Liu, S. Y. Cheng, R. Qi, B. D. Walker, J.-P. Piquemal and P. Ren, Polarizable Force Fields for Biomolecular Simulations: Recent Advances and Applications, *Annu. Rev. Biophys.*, 2019, **48**, 371–394.
76 M. Paloni, R. Bailly, L. Ciandrini and A. Barducci, Unraveling molecular interactions in liquid–liquid phase separation of disordered proteins by atomistic simulations, *J. Phys. Chem. B*, 2020, **124**, 9009–9016.
77 H. Liu, H. Fu, X. Shao, W. Cai and C. Chipot, Accurate description of cation-π interactions in proteins with a nonpolarizable force field at no additional cost, *J. Chem. Theory Comput.*, 2020, **16**, 6397–6407.
78 T. J. Welsh, G. Krainer, J. R. Espinosa, J. A. Joseph, A. Sridhar, M. Jahnel, W. E. Arter, K. L. Saar, S. Alberti, R. Collepardo-Guevara and T. P. J. Knowles, Surface electrostatics govern the emulsion stability of biomolecular condensates, *Nano Lett.*, 2022, **22**, 612–621.
79 G. Krainer, T. J. Welsh, J. A. Joseph, J. R. Espinosa, S. Wittmann, E. de Csilléry, A. Sridhar, Z. Toprakcioglu, G. Gudiškytė, M. A. Czekalska, W. E. Arter, J. Guillén-Boixet, T. M. Franzmann, S. Qamar, P. S. George-Hyslop, A. A. Hyman, R. Collepardo-Guevara, S. Alberti and T. P. Knowles, Reentrant liquid condensate phase of proteins is stabilized by hydrophobic and non-ionic interactions, *Nat. Commun.*, 2021, **12**, 1085.

80 J. A. Joseph, J. R. Espinosa, I. Sanchez-Burgos, A. Garaizar, D. Frenkel and R. Collepardo-Guevara, Thermodynamics and kinetics of phase separation of protein–RNA mixtures by a minimal model, *Biophys. J.*, 2021, **120**, 1219–1230.

81 G. L. Dignon, W. Zheng, Y. C. Kim, R. B. Best and J. Mittal, Sequence determinants of protein phase behavior from a coarse-grained model, *PLoS Comput. Biol.*, 2018, **14**, e1005941.

82 R. M. Regy, G. L. Dignon, W. Zheng, Y. C. Kim and J. Mittal, Sequence dependent phase separation of protein-polynucleotide mixtures elucidated using molecular simulations, *Nucleic Acids Res.*, 2020, **48**, 12593–12603.

83 S. Das, Y. H. Lin, R. M. Vernon, J. D. Forman-Kay and H. S. Chan, Comparative roles of charge, π, and hydrophobic interactions in sequence-dependent phase separation of intrinsically disordered proteins, *Proc. Natl. Acad. Sci. U. S. A.*, 2020, **117**, 28795–28805.

84 J. A. Joseph, A. Reinhardt, A. Aguirre, P. Y. Chew, K. O. Russell, J. R. Espinosa, A. Garaizar and R. Collepardo-Guevara, Physics-driven coarse-grained model for biomolecular phase separation with near-quantitative accuracy, *Nat. Comput. Sci.*, 2021, **1**, 732–743.

85 A. P. Latham and B. Zhang, Consistent Force Field Captures Homologue-Resolved HP1 Phase Separation, *J. Chem. Theory Comput.*, 2021, **17**, 3134–3144.

86 T. Dannenhoffer-Lafage and R. B. Best, A Data-Driven Hydrophobicity Scale for Predicting Liquid–Liquid Phase Separation of Proteins, *J. Phys. Chem. B*, 2021, **125**, 4046–4056.

87 G. Tesei, T. K. Schulze, R. Crehuet and K. Lindorff-Larsen, Accurate model of liquid–liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties, *Proc. Natl. Acad. Sci. U. S. A.*, 2021, **118**, e2111696118.

88 V. Nguemaha and H. X. Zhou, Liquid–liquid phase separation of patchy particles illuminates diverse effects of regulatory components on protein droplet formation, *Sci. Rep.*, 2018, **8**, 6728.

89 J. R. Espinosa, J. A. Joseph, I. Sanchez-Burgos, A. Garaizar, D. Frenkel and R. Collepardo-Guevara, Liquid network connectivity regulates the stability and composition of biomolecular condensates with many components, *Proc. Natl. Acad. Sci. U. S. A.*, 2020, **117**, 13238–13247.

90 H. Liu, S. K. Kumar and F. Sciortino, Vapor-liquid coexistence of patchy models: Relevance to protein phase behavior, *J. Chem. Phys.*, 2007, **127**, 084902.

91 S. Li, K. Yu, Q. Zhang, Z. Liu, J. Liu, H.-Q. Ju, Z. Zuo, X. Li, Z. Wang, H. Cheng and Z.-X. Liu, dSCOPE: a software to detect sequences critical for liquid–liquid phase separation, bioRxiv, 2021, preprint.

92 G. van Mierlo, J. R. Jansen, J. Wang, I. Poser, S. J. van Heeringen and M. Vermeulen, Predicting protein condensate formation using machine learning, *Cell Rep.*, 2021, **34**, 108705.

93 K. L. Saar, A. S. Morgunov, R. Qi, W. E. Arter, G. Krainer, A. A. Lee and T. P. J. Knowles, Learning the molecular grammar of protein condensates from sequence determinants and embeddings, *Proc. Natl. Acad. Sci. U. S. A.*, 2021, **118**, e2019053118.

94 K. M. Ruff, T. S. Harmon and R. V. Pappu, CAMELOT: A machine learning approach for coarse-grained simulations of aggregation of block-copolymeric protein sequences, *J. Chem. Phys.*, 2015, **143**, 243123.

95 S. Kosuri, C. H. Borca, H. Mugnier, M. Tamasi, R. A. Patel, I. Perez, S. Kumar, Z. Finkel, R. Schloss, L. Cai, M. L. Yarmush, M. A. Webb and A. J. Gormley, Machine-assisted discovery of chondroitinase ABC complexes toward sustained neural regeneration, *Adv. Healthcare Mater.*, 2022, **11**, 2102101.

96 Y. Zhang, B. Xu, B. G. Weiner, Y. Meir and N. S. Wingreen, Decoding the physical principles of two-component biomolecular phase separation, *eLife*, 2021, **10**, e62403.

97 Y.-H. Lin, J. D. Forman-Kay and H. S. Chan, Theories for Sequence-Dependent Phase Behaviors of Biomolecular Condensates, *Biochemistry*, 2018, **57**, 2499–2508.

98 Y.-H. Lin, J. P. Brady, J. D. Forman-Kay and H. S. Chan, Charge pattern matching as a 'fuzzy' mode of molecular recognition for the functional phase separations of intrinsically disordered proteins, *New J. Phys.*, 2017, **19**, 115003.

99 T. Pal, J. Wessén, S. Das and H. S. Chan, Subcompartmentalization of polyampholyte species in organelle-like condensates is promoted by charge-pattern mismatch and strong excluded-volume interaction, *Phys. Rev. E*, 2021, **103**, 042406.

100 Y.-H. Lin, J. Wessén, T. Pal, S. Das and H. S. Chan, Numerical Techniques for Applications of Analytical Theories to Sequence-Dependent Phase Separations of Intrinsically Disordered Proteins, in *Methods in Molecular Biology*, ed. H.-X. Zhou, J.-H. Spille and P. R. Banerjee, 2023, vol. 2563, pp. 51–94.

101 I. Sanchez-Burgos, J. R. Espinosa, J. A. Joseph and R. Collepardo-Guevara, Valency and binding affinity variations can regulate the multilayered organization of protein condensates with many components, *Biomolecules*, 2021, **11**, 278.

102 A. Bremer, M. Farag, W. M. Borcherds, I. Peran, E. W. Martin, R. V. Pappu and T. Mittag, Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains, *Nat. Chem.*, 2021, **14**, 196–207.

103 E. J. Williams, The effect of thermal agitation on atomic arrangement in alloys-III, *Proc. R. Soc. London, Ser. A*, 1935, **152**, 231–252.

104 G. Beaucage, R. S. Stein and R. Koningsveld, Tacticity effects on polymer blend miscibility. 1. Flory–Huggins–Staverman analysis, *Macromolecules*, 1993, **26**, 1603–1608.

105 T. S. Harmon, A. S. Holehouse and R. V. Pappu, Differential solvation of intrinsically disordered linkers drives the formation of spatially organized droplets in ternary systems of linear multivalent proteins, *New J. Phys.*, 2018, **20**, 045002.

106 A. Reinhardt and J. P. K. Doye, Note: Homogeneous TIP4P/2005 ice nucleation at low supercooling, *J. Chem. Phys.*, 2013, **139**, 096102.
107 E. W. Martin, A. S. Holehouse, I. Peran, M. Farag, J. J. Incicco, A. Bremer, C. R. Grace, A. Soranno, R. V. Pappu and T. Mittag, Valence and patterning of aromatic residues determine the phase behavior of prion-like domains, *Science*, 2020, **367**, 694–699.
108 J. M. Choi, A. S. Holehouse and R. V. Pappu, Physical Principles Underlying the Complex Biology of Intracellular Phase Transitions, *Annu. Rev. Biophys.*, 2020, **49**, 107–133.
109 T. J. Nott, E. Petsalaki, P. Farber, D. Jervis, E. Fussner, A. Plochowietz, T. D. Craggs, D. P. Bazett-Jones, T. Pawson, J. D. Forman-Kay and A. J. Baldwin, Phase Transition of a Disordered Nuage Protein Generates Environmentally Responsive Membraneless Organelles, *Mol. Cell*, 2015, **57**, 936–947.
110 Y.-H. Lin, J. Song, J. D. Forman-Kay and H. S. Chan, Random-phase-approximation theory for sequence-dependent, biologically functional liquid–liquid phase separation of intrinsically disordered proteins, *J. Mol. Liq.*, 2017, **228**, 176–193.
111 R. K. Das and R. V. Pappu, Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, 13392–13397.
112 R. K. Das, Y. Huang, A. H. Phillips, R. W. Kriwacki and R. V. Pappu, Cryptic sequence features within the disordered protein p27Kip1 regulate cell cycle signaling, *Proc. Natl. Acad. Sci. U. S. A.*, 2016, **113**, 5616–5621.
113 S. Das, A. N. Amin, Y. H. Lin and H. S. Chan, Coarse-grained residue-based models of disordered protein condensates: Utility and limitations of simple charge pattern parameters, *Phys. Chem. Chem. Phys.*, 2018, **20**, 28558–28574.
114 B. L. Miller and D. E. Goldberg, Genetic Algorithms, Tournament Selection, and the Effects of Noise, *Complex Syst.*, 1995, **9**, 193–212.
115 X. Wang, S. Ramírez-Hinestrosa, J. Dobnikar and D. Frenkel, The Lennard-Jones potential: when (not) to use it, *Phys. Chem. Chem. Phys.*, 2020, **22**, 10624–10633.
116 P. Debye and E. Hückel, Zur Theorie der Elektrolyte. I. Gefrierpunktserniedrigung und verwandte Erscheinungen, *Phys. Z.*, 1923, **24**, 185–206.
117 A. Opitz, Molecular dynamics investigation of a free surface of liquid argon, *Phys. Lett. A*, 1974, **47**, 439–440.
118 A. J. Ladd and L. V. Woodcock, Triple-point coexistence properties of the Lennard-Jones system, *Chem. Phys. Lett.*, 1977, **51**, 155–159.
119 S. Plimpton, Fast parallel algorithms for short-range molecular dynamics, *J. Comput. Phys.*, 1995, **117**, 1–19.
120 I. Sanchez-Burgos, J. A. Joseph, R. Collepardo-Guevara and J. R. Espinosa, Size conservation emerges spontaneously in biomolecular condensates formed by scaffolds and surfactant clients, *Sci. Rep.*, 2021, **11**, 15241.
121 J. G. Kirkwood and F. P. Buff, The statistical mechanical theory of surface tension, *J. Chem. Phys.*, 1949, **17**, 338–343.
122 E. de Miguel and G. Jackson, The nature of the calculation of the pressure in molecular simulations of continuous models from volume perturbations, *J. Chem. Phys.*, 2006, **125**, 164109.
123 P. Schofield and J. R. Henderson, Statistical mechanics of inhomogeneous fluids, *Proc. R. Soc. London, Ser. A*, 1982, **379**, 231–246.
124 J. G. Harris, Liquid–vapor interfaces of alkane oligomers: structure and thermodynamics from molecular dynamics simulations of chemically realistic models, *J. Phys. Chem.*, 1992, **96**, 5077–5086.