bioRxiv preprint doi: https://doi.org/10.1101/2023.10.22.563461; this version posted October 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Design of intrinsically disordered protein variants with diverse structural properties

Francesco Pesce[1], **Anne Bremer**[2], **Giulio Tesei**[1], **Jesse B. Hopkins**[3], **Christy R. Grace**[2], **Tanja Mittag**[2], **Kresten Lindorff-Larsen**[1*]

***For correspondence:** lindorff@bio.ku.dk (KLL)

[1]Structural Biology and NMR Laboratory, The Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, Copenhagen, Denmark; [2]Department of Structural Biology, St. Jude Children's Research Hospital, Memphis, TN 38105, USA; [3]BioCAT, Department of Physics, Illinois Institute of Technology, Chicago, IL, USA.

Abstract Intrinsically disordered proteins (IDPs) perform a wide range of functions in biology, suggesting that the ability to design IDPs could help expand the repertoire of proteins with novel functions. Designing IDPs with specific structural or functional properties has, however, been difficult, in part because determining accurate conformational ensembles of IDPs generally requires a combination of computational modelling and experiments. Motivated by recent advancements in efficient physics-based models for simulations of IDPs, we have developed a general algorithm for designing IDPs with specific structural properties. We demonstrate the power of the algorithm by generating variants of naturally occurring IDPs with different levels of compaction and that vary more than 100 fold in their propensity to undergo phase separation, even while keeping a fixed amino acid composition. We experimentally tested designs of variants of the low-complexity domain of hnRNPA1 and find high accuracy in our computational predictions, both in terms of single-chain compaction and propensity to undergo phase separation. We analyze the sequence features that determine changes in compaction and propensity to phase separate and find an overall good agreement with previous findings for naturally occurring sequences. Our general, physics-based method enables the design of disordered sequences with specified conformational properties. Our algorithm thus expands the toolbox for protein design to include also the most flexible proteins and will enable the design of proteins whose functions exploit the many properties afforded by protein disorder.

Introduction

Intrinsically disordered proteins and regions (from here collectively termed IDPs) (***Uversky and Dunker, 2010***) represent a diverse class of proteins that carry out a wide range of functions (***Van Der Lee et al., 2014***) and are characterized by extreme but often non-random structural heterogeneity. Their distinct amino acid composition and sequences (***Uversky et al., 2000***) differ from those of natively folded proteins, and prevent the formation of stable folded conformations. Thus, IDPs are best described by ensembles of heterogeneous conformations that interconvert rapidly (***Mittag and Forman-Kay, 2007***; ***Thomasen and Lindorff-Larsen, 2022***). The disordered and dynamic nature of IDPs is often central for their biological and biochemical functions. They can be linkers separating functional domains, regulating the interaction between the latter (***Li et al., 2018***), or they

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.22.563461; this version posted October 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

can play roles as spacers that impair undesirable protein-protein interactions (***Santner et al., 2012***; ***Jamecna et al., 2019***). IDPs are often involved in mediating molecular interactions including via so-called short-linear motifs (***Davey et al., 2012***), and their large capture radius may give rise to faster binding kinetics compared to that of folded proteins (***Shoemaker et al., 2000***). Thus, IDPs are for example commonly found in signaling molecules (***Wright and Dyson, 2015***) and transcription factors (***Liu et al., 2006***). Furthermore, the interactions within and between IDPs and other biomolecules have emerged as an important factor in the spatial organization of cellular matter. Through their ability to form multivalent interactions, IDPs can aid in or drive the formation of membraneless organelles, which typically consist of a wide range of biomolecules and compartmentalize many biological processes (***Banani et al., 2017***; ***Mittag and Pappu, 2022***). In vitro, many IDPs have been shown to undergo a phase separation (PS) process that leads to the co-existence of a protein-rich dense phase that separates from a dilute phase when the concentration of the protein reaches the so-called saturation concentration (c_{sat}) (***Mittag and Pappu, 2022***). Thus, at concentrations above c_{sat}, the protein may be found both in a dilute phase, and a co-existing dense phase that macroscopically may appear liquid-like and at the molecular level may behave as a viscoelastic fluid (***Mittag and Pappu, 2022***; ***Alshareedah et al., 2023***).

Similarly to the long-lasting quest for predicting the native structure of folded proteins from their sequences (***Kuhlman and Bradley, 2019***), a field which has recently witnessed substantial advances (***Jumper et al., 2021***; ***Baek et al., 2021***; ***Lin et al., 2023***), there is interest in understanding the sequence determinants for the conformational properties of IDPs (***Uversky et al., 2000***; ***Marsh and Forman-Kay, 2010***; ***Das et al., 2015***; ***Cohan et al., 2019***) and how these are related to their function (***Zarin et al., 2021***; ***Tesei et al., 2023***). For both folded and disordered proteins, the ability to predict structure(s) from sequences may help infer its functional properties. Accurate structure prediction may also support or sometimes replace the need for experimental studies of protein structure. Finally, rapid structure prediction enables proteome-wide analyses and can aid in protein design.

In parallel with our continuously improving ability to predict structures of folded proteins, there has been substantial development in our ability to design sequences that fold into specific three-dimensional folded structures (***Pan and Kortemme, 2021***; ***Woolfson, 2021***; ***Goverde et al., 2023***). Given the multitude of functions and properties of IDPs, there would be a great potential in designing IDPs with targeted properties. Such proteins could potentially find applications in designing linkers in multi-domain enzymes (***Van Rosmalen et al., 2017***), signalling molecules, or using IDPs as biomaterials (***Dzuricky et al., 2018***). In contrast to the developments for folded proteins, our ability to design IDPs with specific properties remains more limited. This is because characterizing and predicting the structural properties of IDPs is a complicated task, and because we know less about the sequence-ensemble relationships for IDPs. The native structure of folded proteins can be experimentally determined at atomic resolution, and the availability of many high-resolution structures has been one key driving force to understand and predict how sequences encode structures (***Jumper et al., 2021***). On the other hand, characterizing the ensemble of conformations that an IDP adopts generally requires the integration of experiments and simulation methods (***Mittag and Forman-Kay, 2007***; ***Thomasen and Lindorff-Larsen, 2022***). Collecting and interpreting such data is, however, difficult and often ambiguous, and as a consequence there are only limited examples of detailed structural characterizations (***Lazar et al., 2021***). Thus, there are still many open questions about how the sequence of an IDP translates into a structural ensemble and function (***Lindorff-Larsen and Kragelund, 2021***). Despite these limitations, a number of rules have emerged that govern the local and global conformational properties of IDPs. For example, the content (***Müller-Späth et al., 2010***) and patterning (***Das and Pappu, 2013***) of charged residues has been related to the global expansion of an IDP in solution (***Tesei et al., 2023***; ***Lotthammer et al., 2023***), as well as their propensity to undergo PS (***Lin and Chan, 2017***; ***Schuster et al., 2020***; ***Bremer et al., 2022***). Similarly, hydrophobicity, and in particular the number and patterning of aromatic residues, influences the compaction of an IDP and its propensity to phase separate (***Zheng et al., 2020***; ***Martin***

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.22.563461; this version posted October 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

et al., 2020; *Holehouse et al., 2021*).

A number of different approaches have recently enabled the development of accurate, yet highly computationally-efficient models for molecular simulations of the global conformational properties of IDPs (*Shea et al., 2021*; *Tesei et al., 2021*; *Dannenhoffer-Lafage and Best, 2021*; *Regy et al., 2021*; *Joseph et al., 2021*; *Tesei and Lindorff-Larsen, 2022*). These simulation methods make it possible to use a physics-based coarse-grained model to predict conformational ensembles from sequences on time-scales that are compatible with screening large number of sequences, e.g. all IDPs in the human genome (*Tesei et al., 2023*). Building on these developments, we here present an algorithm to generate sequences of IDPs with pre-defined conformational properties. The central idea is to search sequence space and to use efficient coarse-grained simulations to link each sequence to conformational properties. Specifically, we use the CALVADOS model, that has been optimized by targeting small-angle X-ray scattering (SAXS) and paramagnetic relaxation enhancement NMR experiments on IDPs in solution (*Tesei et al., 2021*), and which has been extensively validated using independent experimental data (*Tesei et al., 2023*). In some aspects our work builds on previous work using genetic algorithms (*Zeng et al., 2021*; *Lichtinger et al., 2021*), but we show how our design method enables large-scale exploration of the sequence-structure space and validate the results experimentally.

We begin by studying four IDPs with different sequence compositions and characteristics. Starting from each sequence, we design new sequences with different levels of compaction while keeping the amino acid composition constant. The results show that—even with the restriction of having a fixed amino acid composition—it is possible to achieve conformational ensembles with highly diverse properties. We show that this is mainly, but not solely, due to differences in the patterning of charges. We used the low complexity domain of hnRNPA1 (hereafter A1-LCD), to study the relationship between sequence patterning, single-chain properties, and the propensity to undergo PS. We selected five variants of A1-LCD for experimental characterization, and find good agreement between the experiments and predictions. Together, our results show that the algorithm that we have developed is efficient and can be used to design IDP sequences with novel properties. The algorithm is fully general, and can therefore also be used to design sequences with varying amino acid composition and for other target properties than chain expansion.

Results

Algorithm to design novel IDPs

To design IDP sequences with specific conformational properties, it is necessary to be able to predict these properties from sequences accurately and rapidly. Therefore, the first question that we address is whether it is possible to use state-of-the-art simulation-based approaches to develop a generalizable method for IDP design. Very recent work has established efficient machine-learning-based methods to predict average conformational properties from sequences (*Tesei et al., 2023*; *Lotthammer et al., 2023*), but these methods do not predict full conformational ensembles and have not been tested experimentally on novel sequences. Instead, we used a simulation-based approach where we employ a coarse-grained model to generate a conformational ensemble for a given sequence (Fig. 1).

We combine coarse-grained molecular dynamics (MD) simulations using the CALVADOS model (*Tesei et al., 2021*) with alchemical free-energy calculations in an algorithm that sequentially generates new sequences and characterizes their conformational ensembles in a time-efficient manner. While MD simulations with a coarse-grained model can rapidly produce conformational ensembles from which structural features can be directly calculated, screening a large number of different IDPs sequentially with only MD simulations would still be computationally difficult. Alchemical free-energy calculations, on the other hand, can predict conformational properties of newly proposed sequences from conformational ensembles generated by simulations of different sequences. Our algorithm thus combines simulations and alchemical free-energy calculations in an optimization

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.22.563461; this version posted October 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

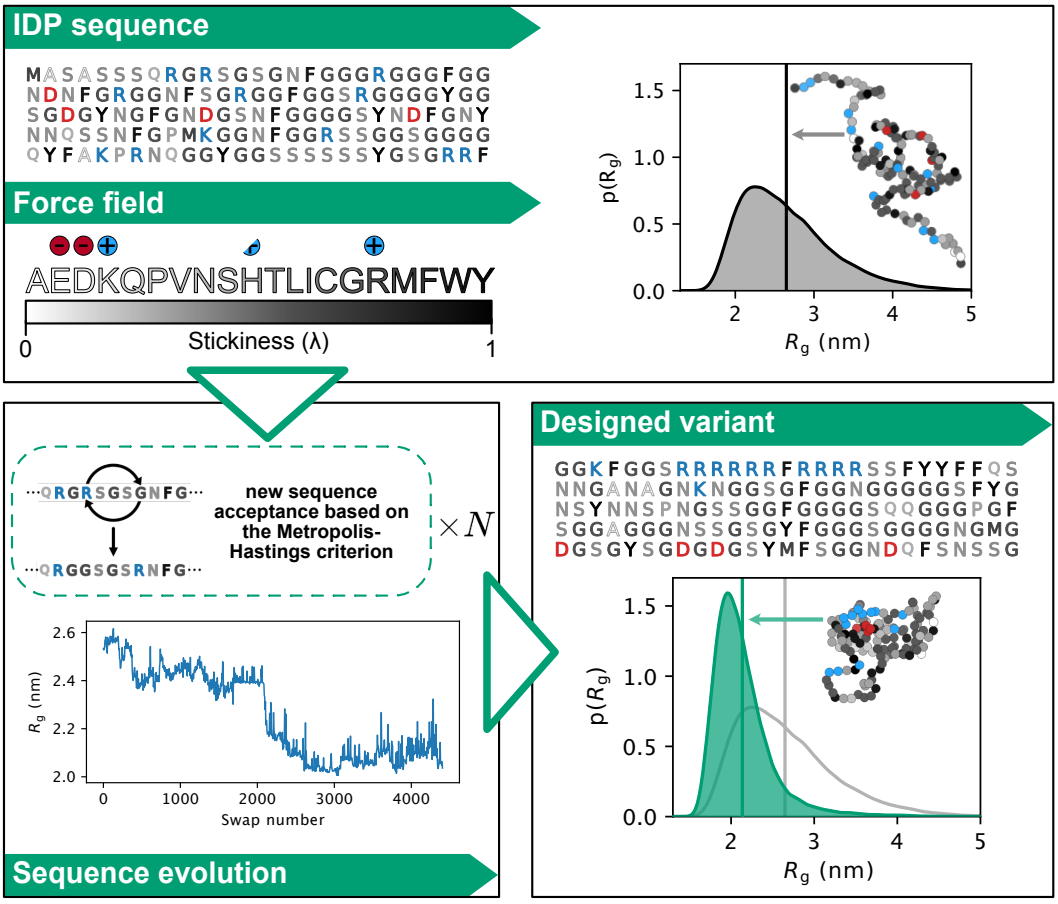

Figure 1. Outline of our algorithm for designing sequences of IDPs with targeted conformational properties. As starting point, we here use naturally occurring IDP sequences, though this is not a requirement of the approach. We use MD simulations with the coarse-grained CALVADOS force field to describe the IDPs and to generate a conformational ensemble. New sequences are proposed through a Markov chain Monte Carlo scheme. We evolve the sequences by consecutive swaps in positions between two randomly selected residues, and evaluate whether the sequences get closer or further away from the design target—here chain compaction. During sequence optimization, we calculate the conformational properties for a given sequence either by direct simulations or through alchemical calculations that rely on conformational ensembles of previously sampled sequences. The conformations shown have the same radius of gyration as the average of the conformational ensemble.

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.22.563461; this version posted October 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

process that in some ways is analogous to what has been proposed in the context of force field optimization (***Norgaard et al., 2008***; ***Orioli et al., 2020***; ***Köfinger and Hummer, 2021***).

While the overall sequence composition of an IDP is known to affect its conformational properties (***Tesei et al., 2023***), we here aimed at exploring the more subtle and difficult-to-extract effects of sequence patterning (***Das and Pappu, 2013***; ***Das et al., 2015***; ***Sherry et al., 2017***; ***Beveridge et al., 2019***; ***Martin et al., 2020***; ***Cohan et al., 2021***). Therefore, we apply our design algorithm to generate sequences of IDPs with diverse structural properties while preserving the overall amino acid composition. In this way we also test and possibly expand our understanding of how the patterning of specific residues in a sequence influences its conformational properties. Early pioneering work focused on the role of charge patterning on conformational properties and propensity to phase separate (***Das and Pappu, 2013***; ***Das et al., 2016***; ***Lin and Chan, 2017***; ***Schuster et al., 2020***). Other studies have linked the number and patterning of amino acids, in particular aromatic and arginine residues, to both conformational and phase properties (***Wang et al., 2018***; ***Martin et al., 2020***; ***Holehouse et al., 2021***; ***Bremer et al., 2022***).

Nonetheless, even restricting the sequence space to sequences of fixed composition, the number of possible sequences is enormous; for example, there are ca. 1.8×10^{127} unique sequences with the amino acid composition of the disordered domain of the fused in sarcoma (FUS) protein. Thus, sampling even a tiny part of this space is unfeasible. To circumvent this problem, our algorithm drives the exploration of the sequence space towards sequences resulting in the target conformational property. This is achieved via a Markov chain Monte Carlo (MCMC) sampling scheme that iteratively generates sequence variants and predicts their conformational properties (through MD simulations and alchemical free-energy calculations) in search of specific arrangements of amino acids that determine a certain structural feature (see Methods for a more detailed description of the algorithm and its components).

To exemplify and demonstrate the power of our algorithm we generate variants of IDPs with either increased or decreased chain expansion, measured by their radius of gyration (R_g), while keeping a fixed amino acid composition. To this aim, at each iteration the algorithm swaps the positions of two randomly selected residues to generate a variant (from hereon called a swap variant). We compare the R_g before and after the swap (evaluated either from MD simulations or alchemical free-energy calculations), and the Monte Carlo move is accepted or rejected based on the Metropolis-Hastings criterion (Fig. 1). Although we here have focused on the difficult problem of changing conformational properties while keeping a fixed amino acid composition, the algorithm is versatile and other criteria can be used to propose changes in the sequences (e.g. single point mutations without keeping a fixed amino acid composition) as well as selecting for other structural features than the R_g.

Design of IDPs with conformational ensembles that vary in compaction

The second question that we address is: Starting from a natural IDP, how much more compact or expanded can it become when only changing the positions of the amino acids in its sequence? To answer this question, we selected four IDPs with different sequence compositions: α-Synuclein (αSyn), and the low complexity domain from hnRNPA1 (A1-LCD), the prion-like domain in FUS (FUS-PLD) and the R-/G-rich domain of the P granule protein LAF-1 (LAF-1-RGG) (Fig. 2a). We used our sequence design algorithm in a simulated annealing protocol to let the sequences evolve in search of amino acid arrangements that result in more compact ensembles. The results show that we can generate sequence permutations of αSyn, A1-LCD and LAF-1-RGG, that are substantially more compact than the wild-type sequence (Fig. 2b, green lines). In contrast, for FUS-PLD we only find variants that are modestly more compact than the wild-type protein. To demonstrate that the algorithm can also find sequences of increased expansion, we began from the compact designs and instead targeted greater R_g values. For αSyn, A1-LCD and LAF-1-RGG we find that the algorithm quickly generates sequences with wild-type-like dimensions (Fig. 2b, orange lines). Interestingly, in all cases the algorithm only finds sequences that are modestly more expanded than the wild-

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.22.563461; this version posted October 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

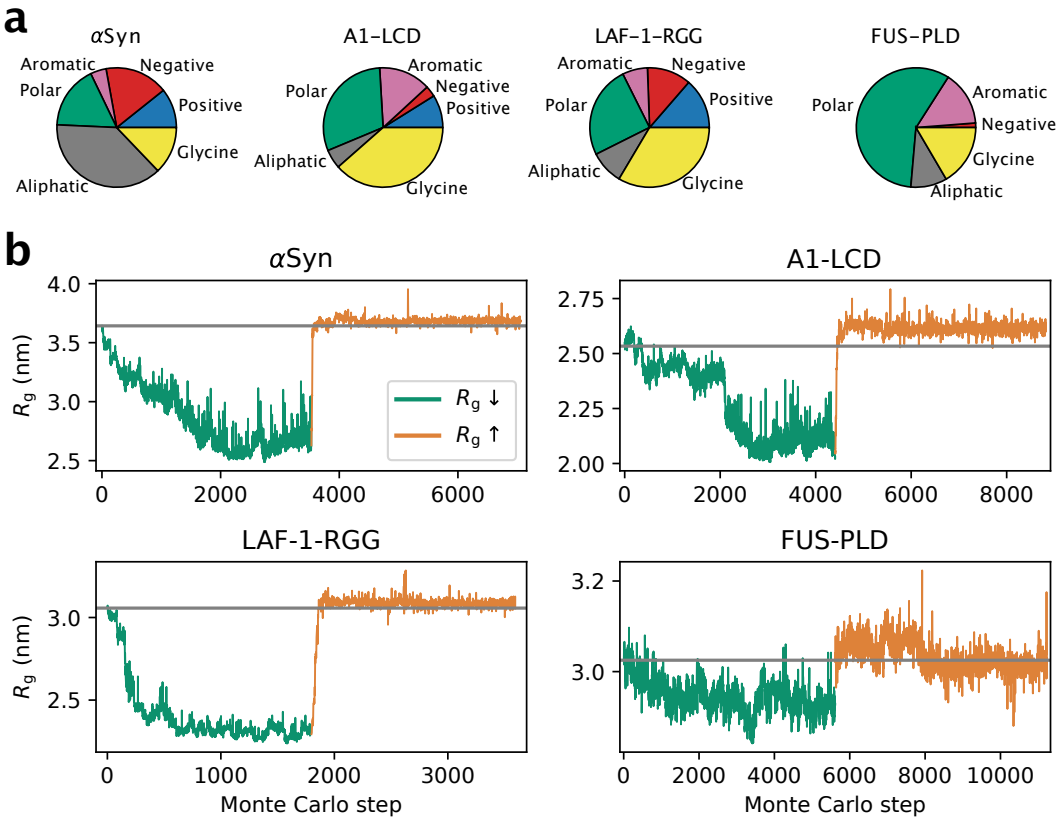

Figure 2. (a) Pie chart of the sequence composition of αSyn, A1-LCD, LAF-1-RGG and FUS-PLD. Amino acids are grouped as negative (D, E), positive (R, K), aromatic (Y, W, F), polar (S, T, N, Q, H, C), aliphatic (A, V, I, L, M, P) and glycine. (b) Design of compact (green lines) and expanded (orange lines) variants for αSyn, A1-LCD, LAF-1-RGG and FUS-PLD. Each accepted Monte Carlo step thus gives rise to a sequence that differs from the previous by the position of the two swapped residues. Each Monte Carlo step therefore corresponds to a different sequence, whose ensemble averaged R_g is evaluated by either MD simulations or alchemical free-energy calculations. The grey horizontal line indicates the R_g of the wild-type sequence.

type sequence although the algorithm was tuned to expand the protein as much as possible. We repeated these calculations starting also from the wild-type sequences and reached similar results (Fig. S1).

Sequence features that determine the compaction of the designs

In the calculations above, we observed that while thousands of swap moves are required for the algorithm to reach the most compact ensembles, a much smaller number of moves was required to recover sequences with wild-type-like dimensions (Fig. 2b). As the moves swap two randomly selected positions, we speculate that there is an entropic barrier in sequence space in finding the arrangement of amino acids that determines compact ensembles. This suggests compaction is driven by some kind of specific ordering of the amino acid sequences. The next question we addressed was therefore: What are the sequence determinants of IDP compaction in the generated sequences? As described above, we were able to generate substantially more compact variants for αSyn, A1-LCD and LAF-1-RGG, but not for FUS-PLD. We therefore aimed to identify which sequence features led to this compaction, and assessed if the same features were responsible in all three cases. We calculated a number of sequence features for the variants of αSyn, A1-LCD and LAF-1-RGG and examined the correlation with the R_g (Figs. 3a and S2). In all cases, we observe a strong correlation between the patterning of the charged amino acid residues, as captured by the κ parameter (***Das and Pappu, 2013***) (Fig. 3a), and chain dimensions. The κ parameter captures

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.22.563461; this version posted October 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

whether the positively and negatively charged residues are well mixed together (low κ) or whether they tend to be found in blocks of like charges (high κ) (***Das and Pappu, 2013***). For all three proteins we observe that the positively charged residues tend to be clustered in the N-terminal third of the sequence and the negatively charged residues in the C-terminal third as the sequences get increasingly compact during the sequence design (Fig. 3b). Since the N-terminus carries a positive charge, and the C-terminus carries a negative charge, it is likely that the termini contribute to the overall charge segregation. We stress that we did not directly drive this charge clustering during the sequence design algorithm, but that the analysis shows that clustering of the charges occurs as the algorithm explores sequence space to generate compact structures. The formation of charge-clustered sequences is in line with the hypothesis above of an 'entropic bottleneck' during the sequence design, and that it is easier to disrupt such patterns than to generate them by randomly swapping amino acid residues.

We also examined other sequence features including patterning of aromatic and hydrophobic residues, and found that they generally have a weaker correlation with the R_g (Fig. S2). For LAF-1-RGG we, however, found that the patterning of hydrophobic residues may also contribute to compaction similarly to the patterning of charges (Fig. S2). This suggests that while charge patterning captures most of the variation in compaction of the permuted sequences, it is difficult to find individual sequence descriptors that fully explain the chain dimensions of IDPs, and that combinations of features may be needed to predict compaction (***Cohan et al., 2021***; ***Tesei et al., 2023***; ***Lotthammer et al., 2023***; ***Chao et al., 2023***). The importance of charge patterning also helps to explain why we were not able to obtain swap variants of FUS-PLD that are more compact than the wild-type, since FUS-PLD has only two negatively charged and no positively charged residues (Fig. 2a).

Relating sequence, compaction and propensity to phase separate for the designs

Theory, simulations and experiments show that the compaction of an IDP is related to its propensity to self-associate and to undergo different forms of phase transitions (***Choi et al., 2020***). Conceptually, this can be understood by the fact that the intramolecular interactions that drive sequence compaction are the same as the intermolecular interactions that drive self-association and phase separation. It would be useful to be able to design proteins with predefined propensities to undergo phase separation and participate in the formation of biomolecular condensates. Building on previous work in this area (***Zeng et al., 2021***; ***Lichtinger et al., 2021***), the fourth question that we sought to answer is: Are the changes in single-chain compaction of the designed swap variants accompanied by a change in their propensity to phase separate? To examine this question we chose to study A1-LCD in more detail because the relationship between sequence and phase separation of A1-LCD has been studied extensively by experiments, theory and simulations (***Martin et al., 2020***; ***Tesei et al., 2021***; ***Bremer et al., 2022***; ***Maristany et al., 2023***).

To improve statistics, we performed nine additional runs of the design algorithm to generate a larger and more diversified pool of A1-LCD variants with different levels of compaction (Fig. S3). We then grouped these sequences by their R_g (in bins of 0.05-nm width), clustered the sequences (see Supplementary material), and use the centroid of each cluster for further analyses. In this way we remove sequences that are very similar to each other (there are many similar sequences within each run of sequence design since the design algorithm evolves sequences by consecutive position swaps of two residues) and only use one representative sequence for each cluster. We then performed 1- μs simulations of each centroid sequence to re-evaluate their R_g. We do this to validate the accuracy of the alchemical free-energy calculations in predicting the R_g of variants proposed by the design algorithm. In line with preliminary tests (Fig. S4, see Methods), we find an average error on the predicted R_g values of 1.5% (Fig. S5). We then re-binned the centroids based on the R_g from simulations, and for each bin we selected up to 15 sequences that are diverse in the patterning of charged and aromatic residues. In this way, we selected 120 A1-LCD swap variants (including the wild type) with diverse sequence features and compaction (Fig. 4a,b). Of the 119

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.22.563461; this version posted October 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

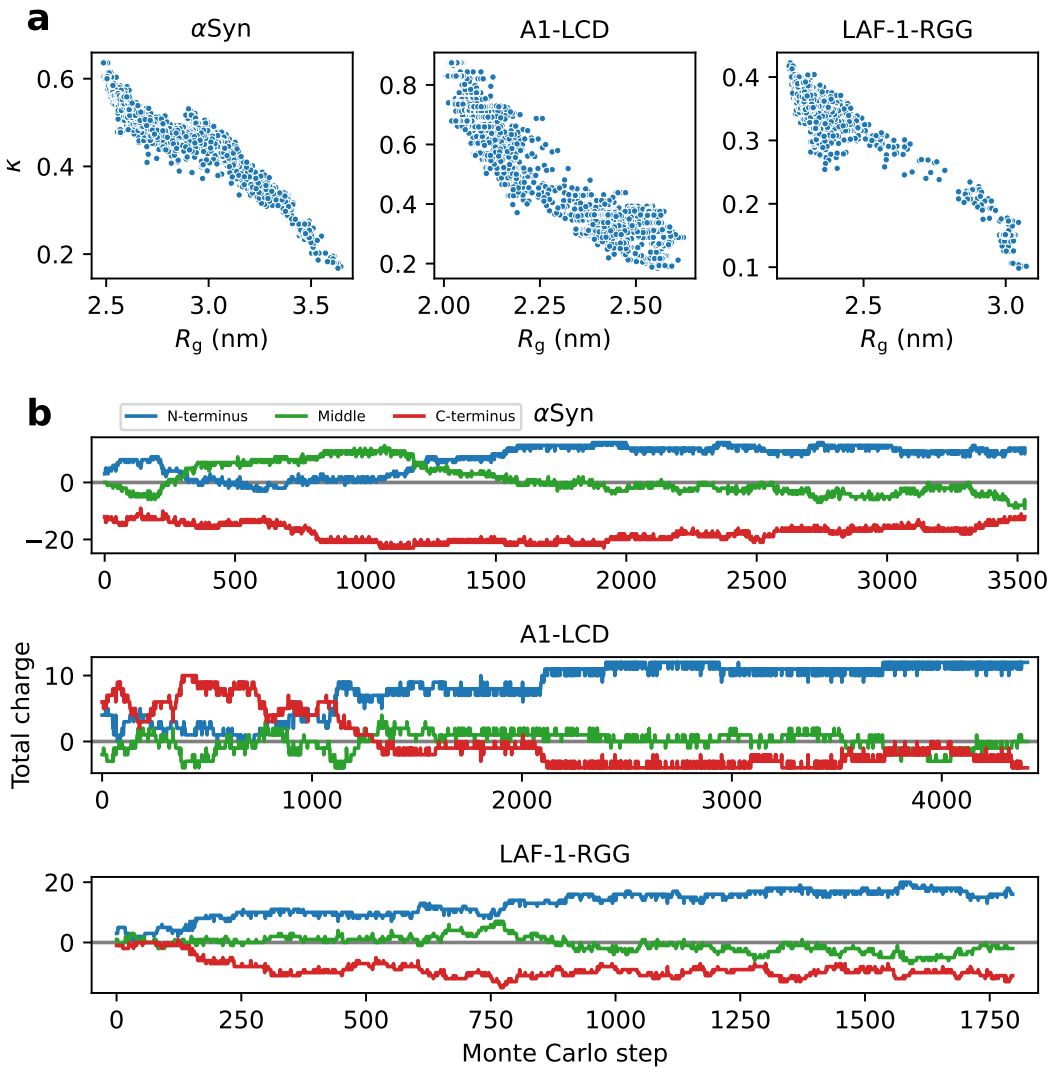

Figure 3. (a) Correlation between R_g and κ (a high κ indicates segregated clusters of residues with the same charge, a low κ indicates that charges are well mixed along the sequence). (b) We divided the sequences of αSyn, A1-LCD and LAF-1-RGG into three sections covering the N-terminal third (blue), the middle third (green), and the C-terminal third (red) of the sequence, and calculated the total charge in each of these sections.

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.22.563461; this version posted October 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

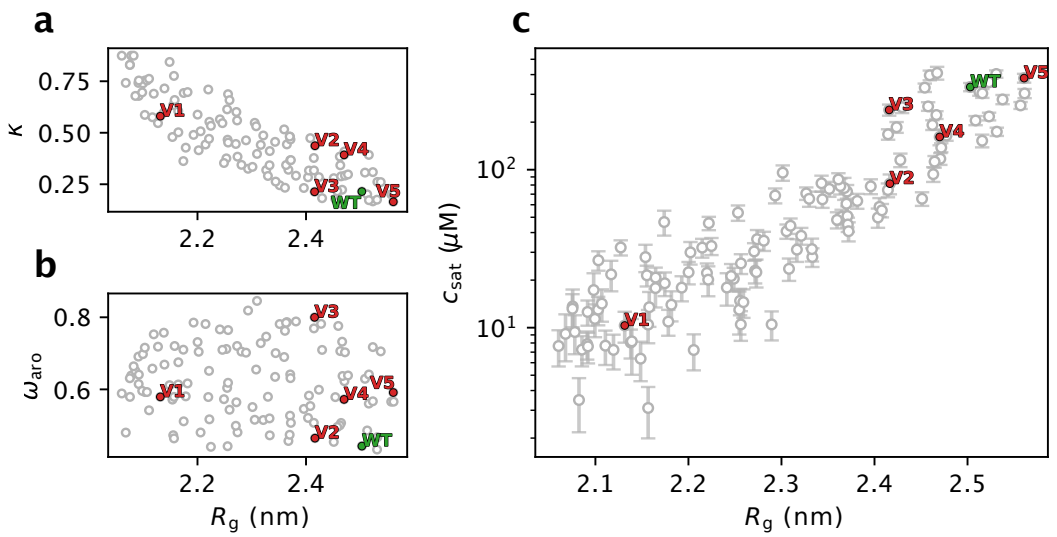

Figure 4. Characterization of the 119 A1-LCD swap variants selected by designing for more compact conformational ensembles and the wild-type (WT) A1-LCD. We show the relationship between R_g and (a) κ, (b) ω_{aro} (patterning of aromatic residues; a high ω_{aro} indicates clustering of aromatic residues), (c) the c_{sat} calculated from simulations of 100 chains in slab geometry. We highlight the WT sequence of A1-LCD in green and five variants selected for experimental characterization in red. Error bars of the average R_g are not shown as their size is negligible.

swap variants, 113 have less than 30% sequence identity to the wild-type protein (Fig. S6).

To examine the propensity of the designed A1-LCD variants to phase separate, we ran simulations of these variants (one at a time) consisting of 100 copies in a 'slab' geometry and estimate their c_{sat} from the concentration of the dilute phase in the simulation box (***Dignon et al., 2018***). As previously observed for a model system (***Lin and Chan, 2017***), we find a logarithmic relationship between R_g and c_{sat}, with compact variants showing a stronger propensity to PS (low c_{sat}), and expanded variants showing a weaker propensity for PS (high c_{sat}) (Fig. 4c). Despite this expected correlation between single-chain properties and the propensity to phase separate, we find some sequences with similar R_g values whose c_{sat} values differ by up to one order of magnitude. This observation suggests that while the single chain behaviour can be very similar, other features encoded in the sequences can cause diversity in the PS properties. Overall, this correlation between R_g and c_{sat} further supports a strong link between single-chain properties and PS propensity that can be used to extrapolate PS propensity from single chain compaction, but also suggests that other sequence features that do not substantially change the single-chain R_g might have a role in PS.

Experimental characterization of A1-LCD variants

Above we have described an approach to design IDPs and examine how the arrangement of amino acids in the primary sequences can influence their behaviour. While the coarse-grained model that we use in our algorithm (***Tesei et al., 2021***) has been extensively validated on naturally occurring proteins and variants thereof (***Tesei et al., 2023***), it has not been used as a generative model and tested on novel, designed sequences. We thus asked whether the accuracy of CALVADOS for predicting R_g and c_{sat} for natural proteins also extends to sequences that show little sequence identity to natural proteins and, for example, show substantial charge patterning. Thus, a fifth question that we asked was: How accurate are our computational predictions of chain compaction and propensity to phase separate for the designed variants?

We therefore sought to test our predictions by experiments. We focused our experiments on

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.22.563461; this version posted October 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

fifteen swap variants of A1-LCD, selected from the 120 sequences analysed above, that represent a range of compaction and sequence properties. We focused on A1-LCD since the wild-type protein is already relatively compact and because its propensity to phase separate is rather strong for a protein of its length (***Martin et al., 2020***; ***Bremer et al., 2022***). Thus, we speculated that the ability to make it even more compact and endow it with lower c_{sat} without changing the amino acid composition would be a powerful test of our design algorithm and the CALVADOS model.

Out of the fifteen variants that we selected, we successfully expressed and purified five variants (red points in Fig. 4 and S7) and the wild-type A1-LCD protein. We ran new simulations of the selected variants under the conditions of the experiments and including a glycine-serine pair at the N-terminus that is present in the experimental constructs (Table S1). We name these variants V1 to V5, sorted by their calculated R_{g}, with V1 predicted to be the most compact and most strongly phase separating variant, with a strong segregation of positive and negative charges at the termini (Fig. 5a). We induced phase separation by adding 150 mM NaCl and visualized the resulting condensates by differential interference contrast (DIC) microscopy. We observed that all variants form condensates, and show some diversity in their morphology (Fig. 5b). We measured the c_{sat} of the five variants and the wild-type and compared the experimental results with those predicted from multi-chain simulations. We find a high correlation between predicted and observed values of c_{sat} (Fig. 5c), with the only outlier being V5, which is the sole variant expected to be more expanded than the WT (Fig. 5b). To investigate possible reasons for the discrepancy in PS propensity of V5 we ran additional simulations. The calculated c_{sat} values that we compare to experiments (Fig. 5c) are averages over the c_{sat} values calculated from three independent simulations. We obtained comparable results from the three independent replicates, demonstrating that the differences are not due to lack of convergence of the simulations (Fig. S8). We also ran simulations with different setups: one with twice as many chains to address potential finite size effects, and another with the updated CALVADOS 2 model (***Tesei and Lindorff-Larsen, 2022***). All three simulation setups gave comparable values for c_{sat} (Fig. S8).

We used previously described methods to measure SAXS data for proteins close to the solubility limit (***Martin et al., 2021***) to test our predictions of sequence compaction. Like for c_{sat}, we find a high correlation between the R_{g} values derived from SAXS and those from simulations (Fig. 5d), and a good agreement between the experimental and calculated SAXS curves with χ_r^2 values around 1–2 (Fig. S9). Given the low c_{sat} of V1 (15 μM), we were not able to obtain a sufficient signal-to-noise ratio at a protein concentration below c_{sat}. We instead turned to diffusion NMR experiments at low protein concentrations to measure the hydrodynamic radius (R_{h}) of V1 and wild-type A1-LCD. We thus acquired NMR data for wild type A1-LCD and V1 at 307 K, where the measured c_{sat} of V1 is 34 μM (compared to 15 μM at 298 K). At this temperature, we find that V1 is substantially more compact than wild-type A1-LCD (Fig. 5e). We note that for both R_{g} and R_{h} there appears to be a small, but systematic, offset between the predicted and experimentally determined values. Some of these differences may indicate remaining errors in the CALVADOS force field, but may also reflect uncertainty in how R_{g} and R_{h} are estimated from experiments and simulations (***Henriques et al., 2018***; ***Pesce and Lindorff-Larsen, 2021***; ***Pesce et al., 2022***; ***Tranchant et al., 2023***), and we also note the high agreement between calculated and experimental SAXS data (Fig. S9).

We find that both simulations and experiments show that V3 is more compact than V4 (Fig. 5d), while V4 has a lower c_{sat} than V3 (Fig. 5c). Previously it has been shown that changes in the formal net charge may break the correlation between R_{g} and c_{sat} (***Tesei et al., 2021***; ***Bremer et al., 2022***), but the case of V3 and V4 show that certain sequence features can break this symmetry even without changing the amino acid composition, and that this is captured by CALVADOS. Examining the sequence features of V3 and V4, we note that V4 has a greater value of κ (indicating that negatively and positively charged residues are not well mixed) (Fig. 4a), while the high value of ω_{aro} in V3 show that the aromatic residues are highly segregated (Fig. 4b); a feature that has previously been correlated with an increased propensity to form amorphous aggregates (***Martin et al., 2020***). Whether these or other sequence features cause the 'symmetry breaking' between R_{g} and c_{sat} for V3 and V4

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.22.563461; this version posted October 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

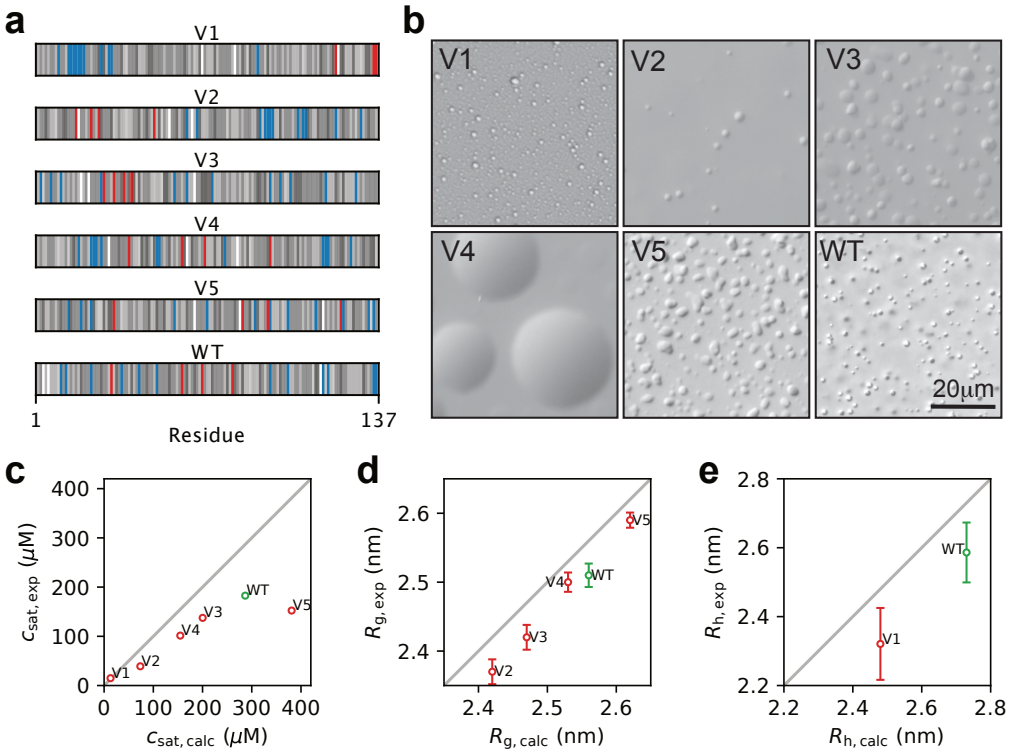

Figure 5. Experimental characterization of wild-type A1-LCD and five designed variants. (a) Diagram of the arrangement of amino acids in A1-LCD and the five design variants. Negative and positive charges are coloured respectively in red and blue. The neutral residues are coloured by a grey scale that reflects their hydrophobicity (corresponding to the λ parameter in CALVADOS), with the least hydrophobic residues in white and the most hydrophobic residues in black. (b) Visualization of condensates of wild-type A1-LCD and the five variants by DIC microscopy; these images are only meant to illustrate the formation of condensates and not necessarily differences between the variants. (c) Comparison of experimental and calculated values of c_{sat} at 298 K. (d) Comparison of experimental and calculated values of R_g for wild-type A1-LCD and V2–V5. (e) Comparison of experimental and calculated values of R_h at 304 K for wild-type A1-LCD and V1. Error bars whose sizes are comparable to that of the markers are not shown.

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.22.563461; this version posted October 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

will be an interesting topic for future analyses.

Designed variants in the context of the human disordered proteome

The results described above show that we can design IDPs with specific levels of compaction and that charge segregation emerges as an important determinant of compaction of the designed sequences. This result is in line with previous observations from theory, simulation and experiments (***Das and Pappu, 2013***; ***Sherry et al., 2017***; ***Choi et al., 2020***). Recently, we have performed simulations of all IDPs from the human proteome (the IDRome), and found that chain compaction of this broad range of natural sequences is governed by a complex interplay between average hydrophobicity, net charge and charge patterning (***Tesei et al., 2023***). Motivated by these observations we examined the results of the sequences generated by our design algorithm in the context of the properties of natural disordered sequences in the human proteome.

The first aspect which we examined was inspired by our observation that we could generate more compact variants of αSyn, A1-LCD and LAF-1-RGG, but not expand these proteins much (Fig. 2). As discussed above, we speculated that this observation was due to the fact that the charged residues in these proteins are already well-mixed so that it is easier to compact them by segregating positive and negative charges than to expand them by further mixing these charged residues. Similarly, we hypothesized that the small number of charged residues in FUS-PLD was the cause of the inability to change the compaction substantially. These observations led us to hypothesize that it would be possible to increase the compaction of natural proteins with stronger charge segregation. We therefore turned to calculations of the $z(\delta_{+-})$ score, which is analogous to the κ score for charge segregation, but is defined in a way that makes it more appropriate for comparisons across sequences of different lengths and compositions (***Cohan et al., 2021***). We thus examined the distribution of $z(\delta_{+-})$ scores across the human IDRome (***Tesei et al., 2023***) and find that, for example, A1-LCD has a well-mixed arrangement of charges as indicated by $z(\delta_{+-}) \approx 0$ (Fig. 6a).

To examine whether charge patterning and compaction of the designed variants reflect the same rules as for natural proteins we turned to the calculation of scaling exponents (ν) as a length-independent measure of compaction. For a so-called 'ideal-chain' polymer, protein–protein, protein–water, and water-water interactions are balanced, and $\nu = 0.5$; smaller values of ν indicate more compact sequences, and an expanded, excluded-volume random-coil has $\nu \approx 0.6$. We calculated ν for the designed A1-LCD variants and find that they follow the overall general relationship between charge segregation ($z(\delta_{+-})$) and sequence compaction (ν) observed for natural proteins (Fig. 6b).

To explore these aspects further, we selected three naturally occurring human IDPs (the disordered domains of HSFX4, FRAT2 and SFMBT1) whose compaction can be explained by their strong segregation of positively and negatively charged residues (Fig. 6c). Building on our hypothesis of why we could not expand the well-mixed sequences of αSyn, A1-LCD and LAF-1-RGG (Fig. 2), we asked whether we could design sequences resulting in more expanded conformational ensembles if we started from these charge segregated sequences. Indeed, when we applied our design algorithm with the wild-type sequences of HSFX4, FRAT2 and SFMBT1 as starting points, we were able to obtain substantially more expanded sequences as well as also modestly more compact sequences (Fig. 6d). Together, these results support the notion that—for fixed sequence composition—modulation of the distribution of the positively and negatively charged residues is a key determinant of compaction and our ability to change this.

While charge segregation is important for fixed sequence composition, we previously found a more complex interplay between a wider range of sequence properties and chain compaction (***Tesei et al., 2023***). These observations in turn enabled us to train a support vector regression (SVR) machine-learning model to predict scaling exponents from sequences (ν_{SVR}). Given that the SVR model was trained on natural sequences, we asked how well our machine learning model was able to predict chain compaction for designs that have properties that are less common in

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.22.563461; this version posted October 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

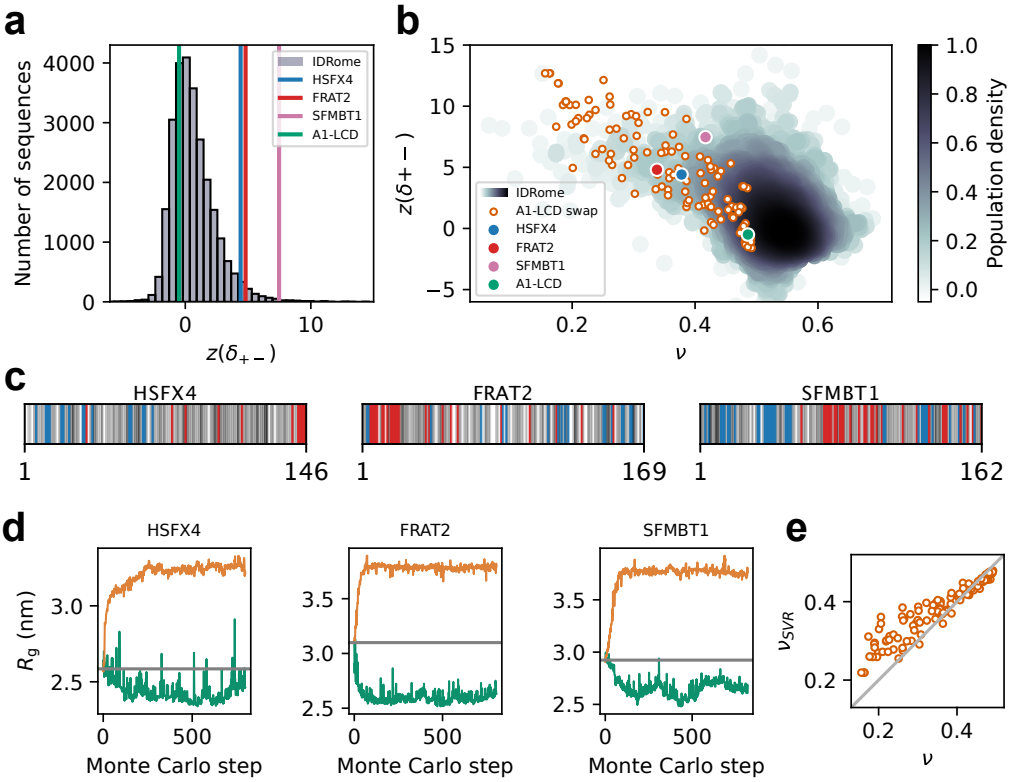

Figure 6. Designed swap variants in the context of the IDRome. (a) Histogram of the sequences in the IDRome grouped based on their charge clustering. We use $z(\delta_{+-})$ to compare the degree of charge clustering for sequences of different lengths and composition, with high values of $z(\delta_{+-})$ indicating high segregation (***Cohan et al., 2021***). $z(\delta_{+-})$ for the wild-type A1, HSFX4, FRAT2, SFMBT1 are indicated respectively in green, blue, red and pink. (b) Comparison of 120 swap variants of A1-LCD (orange) with the IDRome by compaction (ν) and charge clustering ($z(\delta_{+-})$). (c) Diagram of the sequences of disordered regions in HSFX4, FRAT2 and SFMBT1 that we extracted from the IDRome as representative naturally occurring IDPs that show strong charge clustering. Negative and positive charges are coloured respectively in red and blue. The neutral residues are coloured by a grey scale that reflects their hydrophobicity (corresponding to the λ parameter in CALVADOS), with the least hydrophobic residues in white and the most hydrophobic residues in black. (d) Design of more expanded and more compact swap variants starting from the wild-type sequences of the disordered domains of HSFX4, FRAT2 and SFMBT1. (e) Comparison of ν calculated from MD simulations (with CALVADOS 2 (***Tesei and Lindorff-Larsen, 2022***)) and predicted via an SVR machine-learning model (ν_{SVR}) (***Tesei et al., 2023***) for 120 representative A1-LCD variants.

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.22.563461; this version posted October 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

natural sequences. Overall, we find a high correlation between predicted (ν_{SVR}) scaling exponents and those obtained directly from simulations (ν) of the 120 A1-LCD variants (Fig. 6e). The average absolute error of the predictions (14%) is somewhat greater than the value found across the IDRome (2.3%; ***Tesei et al.*** (***2023***)), though these values are not fully comparable due to the different ranges of scaling exponents in the two data sets. We note that defining and calculating scaling exponents is most robust for proteins that behave more like long homopolymers, and that the specific structural properties in the most compact sequences make the average scaling exponent less representative of the conformational ensemble.

Conclusions

Intrinsically disordered proteins and regions play important roles in a range of biological processes and convey functions that complement those of folded proteins. Thus, the ability to design disordered sequences could substantially expand our ability to design proteins with novel functions and properties, in the same way as biology exploits combinations of order and disorder. Combinations of experiments and simulations has led to an improved understanding of the conformational properties of IDPs, which in turn has enabled improved models to generate conformational ensembles directly from sequence via molecular simulations (***Vitalis and Pappu, 2009***; ***Shea et al., 2021***). These models have enabled previous work on design of IDPs (***Zeng et al., 2021***; ***Lichtinger et al., 2021***) and genome-wide studies of sequence-ensemble relationships (***Tesei et al., 2023***; ***Lotthammer et al., 2023***).

Here, we describe a general approach for designing IDPs that exploits a computationally efficient simulation model. Our design algorithm is based on MCMC sampling of sequence space, where each sequence is structurally characterized by combining CALVADOS-based MD simulations (***Tesei et al., 2021***) and alchemical free-energy calculations (***Shirts and Chodera, 2008***). The MCMC sampling guides the sequence towards a design target, and uses the MD simulations and alchemical calculations to predict the conformational ensembles of candidate sequences. Together, this leads to an efficient algorithm that we have successfully used to generate a wide range of sequences with diverse structural features.

We selected five variants of A1-LCD for experimental characterization and find good agreement between experiments and simulations both in terms of the target property (compaction) as well as the propensity of the sequences to undergo phase separation. These findings are in our view important. First, we selected A1-LCD because it is one of the more compact IDPs that have been characterized experimentally, and thus making it even more compact is non-trivial. Second, we restricted our optimization algorithm to maintain sequence composition, and show that we can find substantially more compact sequences even with this restriction. Third, the high correlation between the experimental and calculated radii of gyration demonstrates that CALVADOS remains accurate even for highly unnatural sequences whose properties are well outside those it has previously been trained and benchmarked on. This is a strong validation of our approach of using a physics-based model to drive the sequence design algorithm. We note, however, that the CALVADOS force field we used could have been readily reparameterized to improve predictions of single-chain compaction, in case our experiments had revealed discrepancies with simulation predictions (***Norgaard et al., 2008***; ***Tesei et al., 2021***). Fourth, we show that our designs not only match the experiments for the design target (compaction), but also have phase separation properties that generally match the predictions from simulations. We note, however, that V5 appears to be an outlier since its experimental c_{sat} value is lower than the prediction from CALVADOS and deviates from the observed trend of increasing c_{sat} with increasing R_g. The origin of the discrepancy for the c_{sat} value is unclear and we note again that we accurately predict the R_g of V5.

In addition to developing an algorithm to design IDPs with different levels of compaction, our work also sheds light on sequence-ensemble relationships that can help us understand how natural evolution shapes IDPs. We found that we could generate more compact structures for proteins with the same composition as αSyn, A1-LCD and LAF-1-RGG, but not for FUS-PLD, and that we

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.22.563461; this version posted October 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

could not generate substantially more expanded conformations based on any of these compositions. Our results show that these effects are mainly due to the number and patterning of charged residues in these proteins. Thus, while global sequence composition may be an important factor in the evolution of IDPs (***Hansen et al., 2006***; ***Tompa and Fuxreiter, 2008***; ***Moesa et al., 2012***) our results support the notion that patterning also plays a key role. The results from these analysis are in line with previous bioinformatics analyses that show that most natural IDPs have relatively high mixing of positively and negatively charged residues (***Holehouse et al., 2017***). Nevertheless, we and others have previously shown that some natural IDPs are compact due to strong segregation of positively and negatively charged residues (***Das and Pappu, 2013***; ***Sawle and Ghosh, 2015***; ***Tesei et al., 2023***; ***Lotthammer et al., 2023***), and we show that for sequences such as the disordered domains of HSFX4, FRAT2 and SFMBT1 we can indeed generate more expanded sequences by disrupting this charge patterning. Whether the high mixing of charged residues is due to entropic effects of many tolerated mutations in IDPs (***Nilsson et al., 2011***; ***Schlessinger et al., 2011***; ***Pajkos et al., 2012***; ***Forman-Kay and Mittag, 2013***) or is due to effects e.g. on solubility or preventing erroneous interactions is an interesting question for future studies.

Looking ahead, our results show that the accuracy of CALVADOS appears to extrapolate also outside the realm of the natural proteins, and variants thereof, on which the model was trained. This suggests that even more extensive sampling of sequence space might be useful. While our MCMC-based approach enables a fine-grained and substantial sampling of the sequence space, it may be combined with or replaced by other approaches to guide the sequence design. We and others have recently shown that it is possible to encode the sequence-ensemble relationships from coarse-grained simulations in machine learning methods (***Tesei et al., 2023***; ***Lotthammer et al., 2023***; ***Chao et al., 2023***); we suggest that such methods for predicting properties from sequences may be used together with, for example, reinforcement learning (***Angermueller et al., 2020***; ***Wang et al., 2023***) or Bayesian optimization (***Yang et al., 2022***) to explore sequence space even more efficiently. This would in particular be important when designing for structural observables that are more complex than single-chain compaction, where simulations could be more expensive and alchemical free-energy calculations might be less efficient. Indeed, our algorithm is general and can be applied to design for other structural features than compaction, and can be adapted to other ways of sampling sequence space. The range of applications can therefore be extended to studies focused on understanding the effect of the patterning of specific residues or groups of residues, or to designing for e.g. binders for disordered therapeutic targets.

In summary, we have developed, applied and validated an algorithm for designing disordered sequences with specified conformational properties. We show that we can design IDPs with substantially increased compaction even with fixed amino acid composition, and find that our algorithms mostly exploits the relationship between charge patterning and compaction. We also explain why some sequences are difficult to expand when the positively and negatively charged residues are well-mixed. Our experimental validation highlights the accuracy of the coarse-grained model with prospective testing of novel sequences. Together, our results show that it is now possible to design sequences of disordered proteins, thus expanding our toolbox for designing proteins with novel or improved functions.

Methods

Markov chain Monte Carlo sampling for IDP design

We employed a MCMC algorithm to generate sequences of IDPs. We here targeted the compaction of the chain (as quantified by the R_g) and kept the composition constant during the sequence sampling by using swaps of a randomly selected pair of residues as our MCMC move. We evaluated the R_g of the new sequence, either by running an MD simulation or by reweighting (see below), and

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.22.563461; this version posted October 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

used the Metropolis-Hastings criterion to evaluate the probability of acceptance ($A_{k-1\to k}$):

$$A_{k-1\to k} = \begin{cases} \exp\left[-\frac{|\Delta R_{g,k}|-|\Delta R_{g,k-1}|}{c}\right], & |\Delta R_{g,k}| > |\Delta R_{g,k-1}| \\ 1, & |\Delta R_{g,k}| \le |\Delta R_{g,k-1}| \end{cases} \quad (1)$$

Here, $|\Delta R_{g,k}|$ is the cost function that quantifies the absolute difference between the R_g of the sequence at the MCMC step k and a target R_g ($|\Delta R_{g,k}| = |R_{g,k} - R_{g,target}|$), and c is a control parameter. $R_{g,target}$ is set to 0 nm to design for more compact IDPs and to 10 nm to design for more expanded IDPs. The starting value for c is 0.014, corresponding to $A_{k-1\to k}$=0.5 for $|\Delta R_{g,k}| - |\Delta R_{g,k-1}|$=0.01 nm. We apply simulated annealing using an approach where c is decreased by 1% every $2l$ MCMC steps, where l is the number of amino acids in the IDP sequence.

Although in this work we focus on the specific application of generating variants with fixed amino acid composition, the algorithm and our software accommodates other user-specified MCMC moves (e.g. single- or multi-site amino acid substitutions, substitutions restricted to specific positions and specific residue types). Furthermore, other observables that can be calculated from the simulations can be used as design target. A scheme of the design algorithm is shown in Fig. S10.

Molecular dynamics simulations

We ran coarse-grained molecular dynamics simulations using the CALVADOS M1 (***Tesei et al., 2021***) C_α-based model. Instead, when comparing ν from simulations to ν predicted with the SVR model, we used the CALVADOS 2 (***Tesei and Lindorff-Larsen, 2022***) model since the SVR model was trained on CALVADOS 2 simulations. Single chain simulations in the design algorithm were run for 500 ns with a 10 fs time step. Simulation conditions were set to reproduce 298 K, 150 mM ionic strength and pH 7. Other single chain simulations that are not in the context of the design were run for 1 μs and, when simulations are compared to experiments, at the experimental conditions.

Multi-chain simulations to study the PS propensity of the A1-LCD variants were performed in slab geometry with the CALVADOS M1 model. One hundred chains were assembled in a simulation box 150 nm long and with a cross-section of 15 nm×15 nm. Multi-chain simulations were run for 20 μs. For multi-chain simulations of experimental constructs, three replicates were run for a total simulation time of 120 μs (one replicate 20 μs long and two replicates 50 μs long).

The cut-off used for nonbonded non-ionic interactions was 4 nm for single-chain simulations and 2 nm for multi-chain simulations (***Tesei and Lindorff-Larsen, 2022***). Charge-charge interactions were truncated and shifted at a cut-off of 4 nm in all simulations.

Alchemical free-energy calculations with MBAR

When proposing a new sequence, the design algorithm attempts to predict the R_g by reweighting simulations generated at previous steps of the MCMC algorithm using the Multistate Bennett Acceptance Ratio (MBAR) method (***Shirts and Chodera, 2008***). Since the simulations are performed with a C_α-based coarse-grained model, changing the amino acid type in a position of the sequence simply means changing the force field parameters and possibly the charge of the bead representing the residue at that position. Thus, it is easy to evaluate the per-frame potential energy of a new sequence of conformations sampled with another protein sequence. MBAR takes as input an energy matrix defined by frames coming from n simulations of different sequences (MBAR pool) and the potential energy functions from each sequence. We calculate the potential energies of the frames of the simulations for a new sequence proposed by the MCMC algorithm, and use MBAR to obtain the Boltzmann weights to estimate the weighted average of the R_g of the new sequence without running a new simulation.

The reweighting is most accurate when there is substantial overlap between the potential energy functions of the simulations in the MBAR pool and that of the new sequence. We quantify how much the energies of the frames from the simulations in the MBAR pool are compatible with

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.22.563461; this version posted October 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the potential energy function of the new sequence by calculating the number of effective frames (N_{eff}) that contributes to the averaging:

$$N_{\text{eff}} = N \exp\left[-\sum_{i}^{N} w_i \ln(w_i N)\right] \quad (2)$$

where N is the total number of frames from the simulations in the MBAR pool and w_i is the weight of the i^{th} frame obtained from MBAR to calculate the R_g of the new sequence. By generating test data sets where we compare the simulated R_g with the predicted R_g from MBAR weights, we assessed the relationship between N_{eff} and the accuracy of the predicted R_g (Fig. S4). In light of this analysis, we set a threshold for N_{eff} to 20000. When the weights obtained by MBAR result in a N_{eff} below this threshold, the algorithm initiates a new simulation and uses the R_g from this simulation when evaluating the acceptance probability.

The ability to estimate the R_g of new sequences by reweighting makes the design algorithm more efficient as it decreases the number of MD simulations that are needed. Due to the large size of the energy matrix, we still need to keep the number of simulations in the MBAR pool relatively low, so that the calculations are efficient. With a test data set, we also assessed how the efficiency of the algorithm would change varying the size of the MBAR pool. In general, the larger the pool, the less simulations are required by the algorithm (*i.e.* it occurs less frequently that the N_{eff} drops below 20000). In light of these observations, we set the maximum size of the MBAR pool to 10 (Fig. S4). When the size of the pool is at its maximum and the N_{eff} drops below the threshold, a new simulation is performed and added to the pool, while the oldest simulation is discarded from the MBAR pool.

Small-angle X-ray scattering

SAXS (Fig. S11 and Table S2) was performed at BioCAT (beamline 18ID at the Advanced Photon Source, Chicago) with in-line size exclusion chromatography (SEC-SAXS) to separate sample from aggregates, contaminants and storage buffer components, thus ensuring optimal sample quality (Fig. S12) as previously reported (***Bremer et al., 2022***; ***Martin et al., 2020***, ***2021***). Samples were loaded onto a Superdex 75 Increase 10/300 GL column (Cytiva), which was run at 0.6 mL/min by an AKTA Pure FPLC (GE) and the eluate, after passing through the UV monitor, was flown through the SAXS flow cell. The flow cell consisted of a 1.0 mm ID quartz capillary with ~20 μm walls. All protein solutions were measured at room temperature in 20 mM HEPES (pH 7.0), 150 mM NaCl, 2 mM DTT. A co-flowing buffer sheath was used to separate the sample from the capillary walls, helping prevent radiation damage (***Kirby et al., 2016***). Scattering intensity was recorded using an Eiger2 XE 9M (Dectris) detector which was placed 3.685 m from the sample giving us access to a q-range of 0.0029–0.42 Å^{-1}. 0.5 s exposures were acquired every 1 s during elution and data were reduced using BioXTAS RAW 2.1.4 (***Hopkins et al., 2017***). Buffer blanks were created by averaging regions flanking the elution peak and subtracted from exposures selected from the elution peak to create the $I(q)$ vs q curves (scattering profiles) used for subsequent analyses. RAW was used for buffer subtraction, averaging, and Guinier fits. Scattering profiles were additionally fit using an empirically derived molecular form factor (MFF) (***Riback et al., 2017***) (used to calculate the experimental R_g values in Fig. 5).

Diffusion Ordered NMR Spectroscopy

We carried out diffusion ordered spectroscopy (DOSY) experiments (***Wu et al., 1995***) at 307 K to measure translational diffusion coefficients for WT A1-LCD and the V1 variant, by fitting intensity decays of individual signals selected between 0.5 ppm and 2.5 ppm (***Leeb and Danielsson, 2020***) with the Stejskal-Tanner equation (***Stejskal and Tanner, 1965***). We used 1,4-dioxane (0.10% v/v) as internal reference for the R_h (2.27 ± 0.04 Å, (***Tranchant et al., 2023***)). We acquired 80 scans for A1-LCD and 480 scans for V1. Spectra were recorded on a Bruker 600 MHz spectrometer equipped with a cryoprobe and Z-field gradient, and were obtained over gradient strengths from 5 to 95% (32

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.22.563461; this version posted October 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

points) for A1-LCD and from 5% to 75% (16 points) for V1 (γ = 26752 rad s^{-1} $Gauss^{-1}$) with a diffusion time (Δ) of 50 ms and gradient length (δ) of 6 ms. Translational diffusion coefficients were fitted in Dynamics Center v2.5.6 (Bruker) and were used to estimate the R_h for the proteins (***Prestel et al., 2018***), with error propagation using the diffusion coefficients of both the protein and dioxane.

Data and code availability

Data and code used and produced by this study are available on GitHub. MD simulations of 120 A1-LCD variants and of the six experimental constructs of A1-LCD variants and wild-type, both as single-chain and multi-chains in slab geometry, are available on the Electronic Research Data Archive. SAXS data are deposited in SASDB (***Kikhney et al., 2020***) (Table S2).

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.22.563461; this version posted October 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Author contributions

F.P, T.M. and K.L.-L. designed the study. F.P, G.T. and K.L-L. handled all computational and theoretical aspects. F.P. and A.B. expressed and purified proteins, measured c_{sat} and acquired DIC microscopy images. C.R.G. measured NMR data. F.P. and C.R.G. analyzed NMR data. J.B.H. measured SAXS data. F.P., A.B. and J.B.H. analyzed SAXS data. F.P. and K.L.-L. analyzed the data and wrote the paper with input from all authors.

Acknowledgments

We thank Wade Borcherds and Emil Tranchant for helpful discussions, and George Campbell for assistance with DIC microscopy. This work was supported by the Lundbeck Foundation BRAINSTRUC structural biology initiative (R155-2015-2666, to K.L.-L.) and the PRISM (Protein Interactions and Stability in Medicine and Genomics) centre funded by the Novo Nordisk Foundation (NNF18OC0033950, to K.L.-L.). We acknowledge access to computational resources from the Danish National Supercomputer for Life Sciences (Computerome). This work was supported by the US National Institutes of Health through grant R01NS121114 (T.M.), the St. Jude Research Collaborative on the Biology and Biophysics of RNP granules (T.M.), and the American Lebanese Syrian Associated Charities (to T.M.). We acknowledge use of the Cell and Tissue Imaging Center - Light Microscopy Facility at St. Jude Children's Research Hospital. This research used resources of the Advanced Photon Source, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under Contract No. DE-AC02-06CH11357. BioCAT was supported by grant P30 GM138395 from the National Institute of General Medical Sciences of the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily reflect the official views of the National Institute of General Medical Sciences or the National Institutes of Health.

References

Alshareedah I, Borcherds WM, Cohen SR, Farag M, Singh A, Bremer A, Pappu RV, Mittag T, Banerjee PR. Sequence-encoded grammars determine material properties and physical aging of protein condensates. bioRxiv. 2023; p. 2023–04.

Angermueller C, Dohan D, Belanger D, Deshpande R, Murphy K, Colwell L. Model-Based Reinforcement Learning for Biological Sequence Design. In: *International Conference on Learning Representations (eds A. Rush)*; 2020. .

Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science. 2021; 373(6557):871–876.

Banani SF, Lee HO, Hyman AA, Rosen MK. Biomolecular condensates: organizers of cellular biochemistry. Nature reviews Molecular cell biology. 2017; 18(5):285–298.

Beveridge R, Migas LG, Das RK, Pappu RV, Kriwacki RW, Barran PE. Ion mobility mass spectrometry uncovers the impact of the patterning of oppositely charged residues on the conformational distributions of intrinsically disordered proteins. Journal of the American Chemical Society. 2019; 141(12):4908–4918.

Bremer A, Farag M, Borcherds WM, Peran I, Martin EW, Pappu RV, Mittag T. Deciphering how naturally occurring sequence features impact the phase behaviours of disordered prion-like domains. Nature Chemistry. 2022; 14(2):196–207.

Chao TH, Rekhi S, Mittal J, Tabor DP. Data-Driven Models for Predicting Intrinsically Disordered Protein Polymer Physics Directly from Composition or Sequence. Molecular Systems Design & Engineering. 2023; .

Choi JM, Holehouse AS, Pappu RV. Physical principles underlying the complex biology of intracellular phase transitions. Annual review of biophysics. 2020; 49:107–133.

Cohan MC, Ruff KM, Pappu RV. Information theoretic measures for quantifying sequence–ensemble relationships of intrinsically disordered proteins. Protein Engineering, Design and Selection. 2019; 32(4):191–202.

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.22.563461; this version posted October 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Cohan MC, Shinn MK, Lalmansingh JM, Pappu RV. Uncovering non-random binary patterns within sequences of intrinsically disordered proteins. Journal of Molecular Biology. 2021; p. 167373.

Dannenhoffer-Lafage T, Best RB. A data-driven hydrophobicity scale for predicting liquid–liquid phase separation of proteins. The Journal of Physical Chemistry B. 2021; 125(16):4046–4056.

Das RK, Huang Y, Phillips AH, Kriwacki RW, Pappu RV. Cryptic sequence features within the disordered protein p27Kip1 regulate cell cycle signaling. Proceedings of the National Academy of Sciences. 2016; 113(20):5616–5621.

Das RK, Pappu RV. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. Proceedings of the National Academy of Sciences. 2013; 110(33):13392–13397.

Das RK, Ruff KM, Pappu RV. Relating sequence encoded information to form and function of intrinsically disordered proteins. Current opinion in structural biology. 2015; 32:102–112.

Davey NE, Van Roey K, Weatheritt RJ, Toedt G, Uyar B, Altenberg B, Budd A, Diella F, Dinkel H, Gibson TJ. Attributes of short linear motifs. Molecular BioSystems. 2012; 8(1):268–281.

Dignon GL, Zheng W, Best RB, Kim YC, Mittal J. Relation between single-molecule properties and phase behavior of intrinsically disordered proteins. Proceedings of the National Academy of Sciences. 2018; 115(40):9929–9934.

Dzuricky M, Roberts S, Chilkoti A. Convergence of artificial protein polymers and intrinsically disordered proteins. Biochemistry. 2018; 57(17):2405–2414.

Forman-Kay JD, Mittag T. From sequence and forces to structure, function, and evolution of intrinsically disordered proteins. Structure. 2013; 21(9):1492–1499.

Goverde CA, Wolf B, Khakzad H, Rosset S, Correia BE. De novo protein design by inversion of the AlphaFold structure prediction network. Protein Science. 2023; 32(6):e4653.

Hansen JC, Lu X, Ross ED, Woody RW. Intrinsic protein disorder, amino acid composition, and histone terminal domains. Journal of Biological Chemistry. 2006; 281(4):1853–1856.

Henriques J, Arleth L, Lindorff-Larsen K, Skepö M. On the calculation of SAXS profiles of folded and intrinsically disordered proteins from computer simulations. Journal of molecular biology. 2018; 430(16):2521–2539.

Holehouse AS, Das RK, Ahad JN, Richardson MO, Pappu RV. CIDER: resources to analyze sequence-ensemble relationships of intrinsically disordered proteins. Biophysical journal. 2017; 112(1):16–21.

Holehouse AS, Ginell GM, Griffith D, Böke E. Clustering of Aromatic Residues in Prion-like Domains Can Tune the Formation, State, and Organization of Biomolecular Condensates: Published as part of the Biochemistry virtual special issue “Protein Condensates”. Biochemistry. 2021; 60(47):3566–3581.

Hopkins JB, Gillilan RE, Skou S. BioXTAS RAW: improvements to a free open-source program for small-angle X-ray scattering data reduction and analysis. Journal of applied crystallography. 2017; 50(5):1545–1553.

Jamecna D, Polidori J, Mesmin B, Dezi M, Levy D, Bigay J, Antonny B. An intrinsically disordered region in OSBP acts as an entropic barrier to control protein dynamics and orientation at membrane contact sites. Developmental Cell. 2019; 49(2):220–234.

Joseph JA, Reinhardt A, Aguirre A, Chew PY, Russell KO, Espinosa JR, Garaizar A, Collepardo-Guevara R. Physics-driven coarse-grained model for biomolecular phase separation with near-quantitative accuracy. Nature Computational Science. 2021; 1(11):732–743.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021; 596(7873):583–589.

Kikhney AG, Borges CR, Molodenskiy DS, Jeffries CM, Svergun DI. SASBDB: Towards an automatically curated and validated repository for biological scattering data. Protein Science. 2020; 29(1):66–75.

Kirby N, Cowieson N, Hawley AM, Mudie ST, McGillivray DJ, Kusel M, Samardzic-Boban V, Ryan TM. Improved radiation dose efficiency in solution SAXS using a sheath flow sample environment. Acta Crystallographica Section D: Structural Biology. 2016; 72(12):1254–1266.

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.22.563461; this version posted October 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

677 **Köfinger J**, Hummer G. Empirical optimization of molecular simulation force fields by Bayesian inference. The
678 European Physical Journal B. 2021; 94(12):1–12.

679 **Kuhlman B**, Bradley P. Advances in protein structure prediction and design. Nature Reviews Molecular Cell
680 Biology. 2019; 20(11):681–697.

681 **Lazar T**, Martínez-Pérez E, Quaglia F, Hatos A, Chemes LB, Iserte JA, Méndez NA, Garrone NA, Saldaño TE,
682 Marchetti J, et al. PED in 2021: a major update of the protein ensemble database for intrinsically disordered
683 proteins. Nucleic acids research. 2021; 49(D1):D404–D411.

684 **Leeb S**, Danielsson J. Obtaining Hydrodynamic Radii of Intrinsically Disordered Protein Ensembles by Pulsed
685 Field Gradient NMR Measurements. In: *Intrinsically Disordered Proteins* Springer; 2020.p. 285–302.

686 **Li M**, Cao H, Lai L, Liu Z. Disordered linkers in multidomain allosteric proteins: Entropic effect to favor the open
687 state or enhanced local concentration to favor the closed state? Protein Science. 2018; 27(9):1600–1610.

688 **Lichtinger SM**, Garaizar A, Collepardo-Guevara R, Reinhardt A. Targeted modulation of protein liquid–liquid
689 phase separation by evolution of amino-acid sequence. PLOS Computational Biology. 2021; 17(8):e1009328.

690 **Lin YH**, Chan HS. Phase separation and single-chain compactness of charged disordered proteins are strongly
691 correlated. Biophysical Journal. 2017; 112(10):2043–2046.

692 **Lin Z**, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, et al. Evolutionary-scale
693 prediction of atomic-level protein structure with a language model. Science. 2023; 379(6637):1123–1130.

694 **Lindorff-Larsen K**, Kragelund BB. On the potential of machine learning to examine the relationship between
695 sequence, structure, dynamics and function of intrinsically disordered proteins. Journal of Molecular Biology.
696 2021; 433(20):167196.

697 **Liu J**, Perumal NB, Oldfield CJ, Su EW, Uversky VN, Dunker AK. Intrinsic disorder in transcription factors. Bio-
698 chemistry. 2006; 45(22):6873–6888.

699 **Lotthammer JM**, Ginell GM, Griffith D, Emenecker RJ, Holehouse AS. Direct prediction of intrinsically disordered
700 protein conformational properties from sequence. bioRxiv. 2023; p. 2023–05.

701 **Maristany MJ**, Aguirre Gonzalez A, Collepardo-Guevara R, Joseph JA. Universal predictive scaling laws of phase
702 separation of prion-like low complexity domains. bioRxiv. 2023; p. 2023–06.

703 **Marsh JA**, Forman-Kay JD. Sequence determinants of compaction in intrinsically disordered proteins. Biophys-
704 ical journal. 2010; 98(10):2383–2390.

705 **Martin EW**, Holehouse AS, Peran I, Farag M, Incicco JJ, Bremer A, Grace CR, Soranno A, Pappu RV, Mittag T.
706 Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. Science.
707 2020; 367(6478):694–699.

708 **Martin EW**, Hopkins JB, Mittag T. Small-angle X-ray scattering experiments of monodisperse intrinsically dis-
709 ordered protein samples close to the solubility limit. In: *Methods in Enzymology*, vol. 646 Elsevier; 2021.p.
710 185–222.

711 **Mittag T**, Forman-Kay JD. Atomic-level characterization of disordered protein ensembles. Current opinion in
712 structural biology. 2007; 17(1):3–14.

713 **Mittag T**, Pappu RV. A conceptual framework for understanding phase separation and addressing open ques-
714 tions and challenges. Molecular Cell. 2022; .

715 **Moesa HA**, Wakabayashi S, Nakai K, Patil A. Chemical composition is maintained in poorly conserved in-
716 trinsically disordered regions and suggests a means for their classification. Molecular BioSystems. 2012;
717 8(12):3262–3273.

718 **Müller-Späth S**, Soranno A, Hirschfeld V, Hofmann H, Rüegger S, Reymond L, Nettels D, Schuler B. Charge
719 interactions can dominate the dimensions of intrinsically disordered proteins. Proceedings of the National
720 Academy of Sciences. 2010; 107(33):14609–14614.

721 **Nilsson J**, Grahn M, Wright AP. Proteome-wide evidence for enhanced positive Darwinian selection within
722 intrinsically disordered regions in proteins. Genome biology. 2011; 12(7):1–17.

723 **Norgaard AB**, Ferkinghoff-Borg J, Lindorff-Larsen K. Experimental parameterization of an energy function for
724 the simulation of unfolded proteins. Biophysical journal. 2008; 94(1):182–192.

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.22.563461; this version posted October 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Orioli S, Larsen AH, Bottaro S, Lindorff-Larsen K. How to learn from inconsistencies: Integrating molecular simulations with experimental data. Progress in Molecular Biology and Translational Science. 2020; 170:123–176.

Pajkos M, Mészáros B, Simon I, Dosztányi Z. Is there a biological cost of protein disorder? Analysis of cancer-associated mutations. Molecular BioSystems. 2012; 8(1):296–307.

Pan X, Kortemme T. Recent advances in de novo protein design: Principles, methods, and applications. Journal of Biological Chemistry. 2021; 296.

Pesce F, Lindorff-Larsen K. Refining conformational ensembles of flexible proteins against small-angle x-ray scattering data. Biophysical journal. 2021; 120(22):5124–5135.

Pesce F, Newcombe EA, Seiffert P, Tranchant EE, Olsen JG, Grace CR, Kragelund BB, Lindorff-Larsen K. Assessment of models for calculating the hydrodynamic radius of intrinsically disordered proteins. Biophysical Journal. 2022; .

Prestel A, Bugge K, Staby L, Hendus-Altenburger R, Kragelund BB. Characterization of dynamic IDP complexes by NMR spectroscopy. In: *Methods in enzymology*, vol. 611 Elsevier; 2018.p. 193–226.

Regy RM, Thompson J, Kim YC, Mittal J. Improved coarse-grained model for studying sequence dependent phase separation of disordered proteins. Protein Science. 2021; 30(7):1371–1379.

Riback JA, Bowman MA, Zmyslowski AM, Knoverek CR, Jumper JM, Hinshaw JR, Kaye EB, Freed KF, Clark PL, Sosnick TR. Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water. Science. 2017; 358(6360):238–241.

Santner AA, Croy CH, Vasanwala FH, Uversky VN, Van YYJ, Dunker AK. Sweeping away protein aggregation with entropic bristles: intrinsically disordered protein fusions enhance soluble expression. Biochemistry. 2012; 51(37):7250–7262.

Sawle L, Ghosh K. A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins. The Journal of chemical physics. 2015; 143(8).

Schlessinger A, Schaefer C, Vicedo E, Schmidberger M, Punta M, Rost B. Protein disorder—a breakthrough invention of evolution? Current opinion in structural biology. 2011; 21(3):412–418.

Schuster BS, Dignon GL, Tang WS, Kelley FM, Ranganath AK, Jahnke CN, Simpkins AG, Regy RM, Hammer DA, Good MC, et al. Identifying sequence perturbations to an intrinsically disordered protein that determine its phase-separation behavior. Proceedings of the National Academy of Sciences. 2020; 117(21):11421–11431.

Shea JE, Best RB, Mittal J. Physics-based computational and theoretical approaches to intrinsically disordered proteins. Current opinion in structural biology. 2021; 67:219–225.

Sherry KP, Das RK, Pappu RV, Barrick D. Control of transcriptional activity by design of charge patterning in the intrinsically disordered RAM region of the Notch receptor. Proceedings of the National Academy of Sciences. 2017; 114(44):E9243–E9252.

Shirts MR, Chodera JD. Statistically optimal analysis of samples from multiple equilibrium states. The Journal of chemical physics. 2008; 129(12):124105.

Shoemaker BA, Portman JJ, Wolynes PG. Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. Proceedings of the National Academy of Sciences. 2000; 97(16):8868–8873.

Stejskal EO, Tanner JE. Spin diffusion measurements: spin echoes in the presence of a time-dependent field gradient. The journal of chemical physics. 1965; 42(1):288–292.

Tesei G, Lindorff-Larsen K. Improved predictions of phase behaviour of intrinsically disordered proteins by tuning the interaction range. bioRxiv. 2022; .

Tesei G, Schulze TK, Crehuet R, Lindorff-Larsen K. Accurate model of liquid–liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties. Proceedings of the National Academy of Sciences. 2021; 118(44):e2111696118.

Tesei G, Trolle AI, Jonsson N, Betz J, Pesce F, Johansson KE, Lindorff-Larsen K. Conformational ensembles of the human intrinsically disordered proteome: Bridging chain compaction with function and sequence conservation. bioRxiv. 2023; p. 2023–05.

bioRxiv preprint doi: https://doi.org/10.1101/2023.10.22.563461; this version posted October 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Thomasen FE, Lindorff-Larsen K. Conformational ensembles of intrinsically disordered proteins and flexible multidomain proteins. Biochemical Society Transactions. 2022; 50(1):541–554.

Tompa P, Fuxreiter M. Fuzzy complexes: polymorphism and structural disorder in protein–protein interactions. Trends in biochemical sciences. 2008; 33(1):2–8.

Tranchant EE, Pesce F, Jacobsen NL, Fernandes CB, Kragelund BB, Lindorff-Larsen K. Revisiting the use of dioxane as a reference compound for determination of the hydrodynamic radius of proteins by pulsed field gradient NMR spectroscopy. bioRxiv. 2023; p. 2023–06.

Uversky VN, Dunker AK. Understanding protein non-folding. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics. 2010; 1804(6):1231–1264.

Uversky VN, Gillespie JR, Fink AL. Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins: structure, function, and bioinformatics. 2000; 41(3):415–427.

Van Der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, Fuxreiter M, Gough J, Gsponer J, Jones DT, et al. Classification of intrinsically disordered regions and proteins. Chemical reviews. 2014; 114(13):6589–6631.

Van Rosmalen M, Krom M, Merkx M. Tuning the flexibility of glycine-serine linkers to allow rational design of multidomain proteins. Biochemistry. 2017; 56(50):6565–6574.

Vitalis A, Pappu RV. ABSINTH: a new continuum solvation model for simulations of polypeptides in aqueous solutions. Journal of computational chemistry. 2009; 30(5):673–699.

Wang J, Choi JM, Holehouse AS, Lee HO, Zhang X, Jahnel M, Maharana S, Lemaitre R, Pozniakovsky A, Drechsel D, et al. A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins. Cell. 2018; 174(3):688–699.

Wang Y, Tang H, Huang L, Pan L, Yang L, Yang H, Mu F, Yang M. Self-play reinforcement learning guides protein engineering. Nature Machine Intelligence. 2023; p. 1–16.

Woolfson DN. A brief history of de novo protein design: minimal, rational, and computational. Journal of Molecular Biology. 2021; 433(20):167160.

Wright PE, Dyson HJ. Intrinsically disordered proteins in cellular signalling and regulation. Nature reviews Molecular cell biology. 2015; 16(1):18–29.

Wu D, Chen A, Johnson CS. An improved diffusion-ordered spectroscopy experiment incorporating bipolar-gradient pulses. Journal of magnetic resonance, Series A. 1995; 115(2):260–264.

Yang Z, Milas KA, White AD. Now What Sequence? Pre-trained Ensembles for Bayesian Optimization of Protein Sequences. bioRxiv. 2022; .

Zarin T, Strome B, Peng G, Pritišanac I, Forman-Kay JD, Moses AM. Identifying molecular features that are associated with biological function of intrinsically disordered protein regions. Elife. 2021; 10:e60220.

Zeng X, Liu C, Fossat MJ, Ren P, Chilkoti A, Pappu RV. Design of intrinsically disordered proteins that undergo phase transitions with lower critical solution temperatures. APL Materials. 2021; 9(2):021119.

Zheng W, Dignon G, Brown M, Kim YC, Mittal J. Hydropathy patterning complements charge patterning to describe conformational preferences of disordered proteins. The journal of physical chemistry letters. 2020; 11(9):3408–3415.