nature computational science

Article

https://doi.org/10.1038/s43588-025-00881-y

Generalized design of sequence–ensemble–function relationships for intrinsically disordered proteins

Received: 15 November 2024

Accepted: 28 August 2025

Published online: 06 October 2025

Check for updates

Ryan K. Krueger [1], Michael P. Brenner [1,2] ✉ & Krishna Shrinivas [3,4] ✉

The design of folded proteins has advanced substantially in recent years. However, many proteins and protein regions are intrinsically disordered and lack a stable fold, that is, the sequence of an intrinsically disordered protein (IDP) encodes a vast ensemble of spatial conformations that specify its biological function. This conformational plasticity and heterogeneity makes IDP design challenging. Here we introduce a computational framework for de novo design of IDPs through rational and efficient inversion of molecular simulations that approximate the underlying sequence–ensemble relationship. We highlight the versatility of this approach by designing IDPs with diverse properties and arbitrary sequence constraints. These include IDPs with target ensemble dimensions, loops and linkers, highly sensitive sensors of physicochemical stimuli, and binders to target disordered substrates with distinct conformational biases. Overall, our method provides a general framework for designing sequence–ensemble–function relationships of biological macromolecules.

The basis of biomolecular function is often specified by a sequence that encodes an ensemble of 3D conformations[1]. A prominent example is intrinsically disordered protein regions (IDPs), which are found in most living organisms and play key roles in diverse cellular functions including transcription, cell signaling, cellular immunity and translation[2–4]. Intrinsically disordered protein regions lack a stable 3D structure, they instead dynamically interconvert between a large range of non-random conformations[5–7] whose local and global properties shape cellular functions[2]. Intrinsically disordered protein regions facilitate molecular recognition through embedded short linear motifs[8] and fuzzy interactions with multiple targets[2], and when tethered as intervening linkers or spacers, they modulate interactions between adjacent folded-domains[9]. The conformational plasticity that underlies IDPs is highly sensitive to physicochemical and environmental contexts, and thus they often function as intracellular sensors[10]. Further, IDPs regulate assembly of higher-order biomolecular assemblies and condensates[11–14], often through low-affinity multivalent interactions, which play central roles in cellular signaling and information processing. Finally, dysregulation of IDPs and IDP-dependent interactions is increasingly correlated with multiple pathological states[11,15]. There is thus widespread interest in designing IDPs with tailored functions for a variety of roles in human health and industry.

Despite recent advances in protein structure design enabled by the Protein Data Bank (PDB) and machine learning[16–19], these computational methods remain limited in their ability to design disordered proteins. Structures of IDPs are not characterized by single stable folds, they instead occupy a vast ensemble of dynamic configurations. Recent developments in coarse-grained molecular simulations have successfully predicted ensemble properties of IDPs[20–22]. These simulations produce training data for approximate machine-learning models that predict particular properties[5,23] (for example, the radius of gyration and polymer exponents) and can be subsequently inverted for design[24]. Although each method has found success, using separate algorithms for the forward and inverse

[1]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA. [2]Department of Physics, Harvard University, Cambridge, MA, USA. [3]Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL, USA. [4]Center for Synthetic Biology, Northwestern University, Evanston, IL, USA. ✉e-mail: brenner@seas.harvard.edu; krishna@northwestern.edu

https://doi.org/10.1038/s43588-025-00881-y

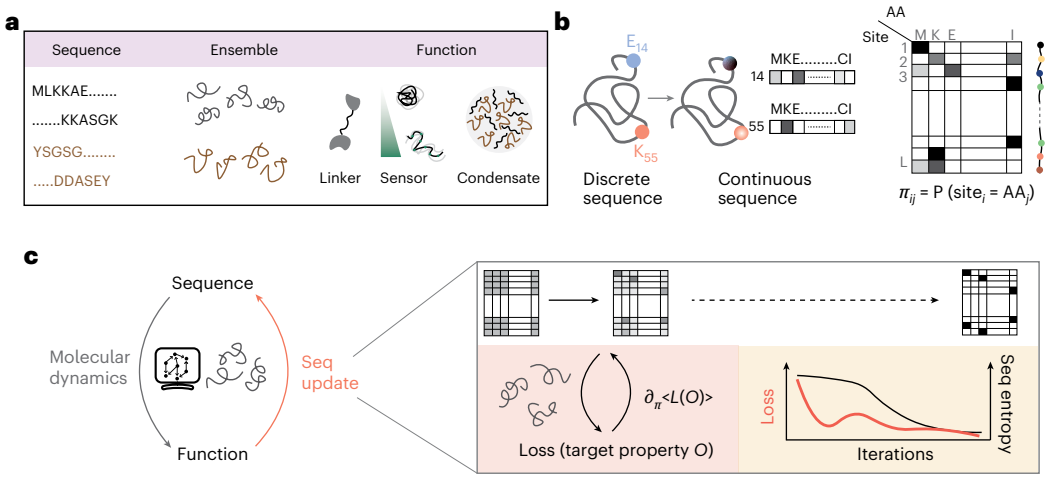

Fig. 1 | Method for inverse design of IDPs. a, The amino acid sequence of an IDP encodes for an ensemble of dynamic 3D conformation structures, which determine properties shaping molecular and cellular functions. **b**, A discrete IDP sequence (a vector of length n, where each position is typically a categorically represented amino acid character) is relaxed to a continuous, probabilistic sequence representation π (a matrix of size $n \times 20$). Here, the (i,j) entry of π is the probability of residue at position i being amino acid j. **c**. To model the forward sequence–ensemble relationship, we simulate the probabilistic sequence through coarse-grained molecular dynamics simulations, defining the Hamiltonian of the system as the expected Hamiltonian over all sequences (refer to the 'General framework for optimizing particle identities' section in the Methods). To invert this relationship for sequence design, we optimize this probabilistic sequence π via gradient descent and anneal to a discrete sequence through the optimization.

problems reduces accuracy and generalizability to different target properties and force-field parameters. It would be much more preferable to directly invert the molecular simulations that model the sequence–ensemble relationship.

In this paper we introduce an algorithmic approach to design IDPs with tailored properties by inverting molecular simulations. Our framework uses gradient-based optimization on molecular simulations to design sequences with arbitrary equilibrium properties—bridging machine-learning technology with ideas from statistical physics. We employ this method to engineer IDP sequences for a wide range of ensemble dimensions of varying complexity, including highly optimized loops and linkers. Our framework naturally accommodates arbitrary sequence constraints, which we highlight through the design of sequence patterning variants with the same composition but distinct ensemble properties. We then construct IDP-based sensors that are sensitive to salt concentrations, temperature and phosphorylation. Finally, we design candidate IDP binders for highly disordered biological and synthetic substrates. Of note, the accuracy of our predictions is limited by the accuracy of simulation parameters that describe IDP sequence–ensemble relationships; our contribution is to show how to find optimal sequences given a potential. Although our proposed method is, in principle, potential-agnostic, it will benefit from the continued iteration between force-field development and experiment. Overall, our paper outlines a flexible strategy for de novo IDP design that can be generalized to engineer sequence–ensemble–function relationships for diverse biopolymers.

Results

Model formulation

Rational de novo design of IDPs requires two key ingredients: (1) a reasonably accurate forward model of the sequence–ensemble–function paradigm (Fig. 1a) and (2) an algorithm to invert this through directed search of sequence space towards a desired functional property. Over the past few years, coarse-grained molecular simulations with custom-pair potentials have made dramatic improvements[5,21,25] in predicting effective ensemble properties of IDPs. Here we focus on molecular dynamics simulations using 1 AA = 1 bead coarse-graining with the Mpipi-GG force-field (refer to the 'Mpipi force-field' section in the Methods and Supplementary Section 1)[21,23].

Our key innovation is the development of a differentiable algorithmic framework to invert the simulation-based sequence–ensemble relationships. To do this, we leverage recent advances in differentiable programming and stochastic gradient estimation[26–29] to compute the gradient of a loss function that depends on any set of ensemble-averaged properties: $\partial_{\text{seq}}\mathcal{L}(\langle P_1^{\text{sim}}\rangle, \langle P_2^{\text{sim}}\rangle, \cdots)$, where P_i^{sim} denotes the ith state-level property. As this quantity is only well-defined for smooth variable changes, we use a continuous representation of the sequence that is amenable to simulation and parallelization on GPUs (see the 'Computational performance and tradeoffs' section in the Methods). For a sequence of L residues, this continuous probabilistic representation (Fig. 1b), $\pi = f(\lambda)$, is defined by logits λ of size $L \times 20$. The residue identity at every site is characterized by a normalized probability vector over the different types of amino acids. A particular discrete sequence corresponds to a one-hot encoding, meaning each position is represented by a vector of length 20 with all entries but one being 0. In general, ensemble-averaged predictions are not identical to predictions from a distribution of discrete sequences sampled from the same distribution (see equation (4) and Supplementary Fig. 3).

Although, in principle, libraries such as JAX-MD[27] enable gradient calculation over unrolled molecular dynamics trajectories, this is slow, scales poorly with system size and is plagued by numerical instability (Supplementary Fig. 1 and Supplementary Section 2). To address this, we expand on a perturbative calculation developed independently by Zhang et al.[29] and Thaler and Zavadlav[28] to calculate the gradient with respect to π from a set of states sampled from the equilibrium Boltzmann distribution. Correspondingly, forward molecular dynamics simulations are set up (see the 'Simulations' section in the Methods) to ensure equilibration and broad sampling of the reference conformational repertoire (Supplementary Fig. 2 and Supplementary Section 2). Employing this calculation provides considerable speed and accuracy increases in gradient estimation and allows re-use of simulation snapshots for multiple sequence updates. Finally, we incorporate an annealing procedure that gradually forces π to become increasingly discrete through the optimization (Fig. 1c; refer also to the 'General framework for optimizing particle identities' section in the Methods). Unless otherwise specified, we initialize all optimizations with a uniform distribution. Once optimized, target properties are evaluated and validated with separate, longer

https://doi.org/10.1038/s43588-025-00881-y

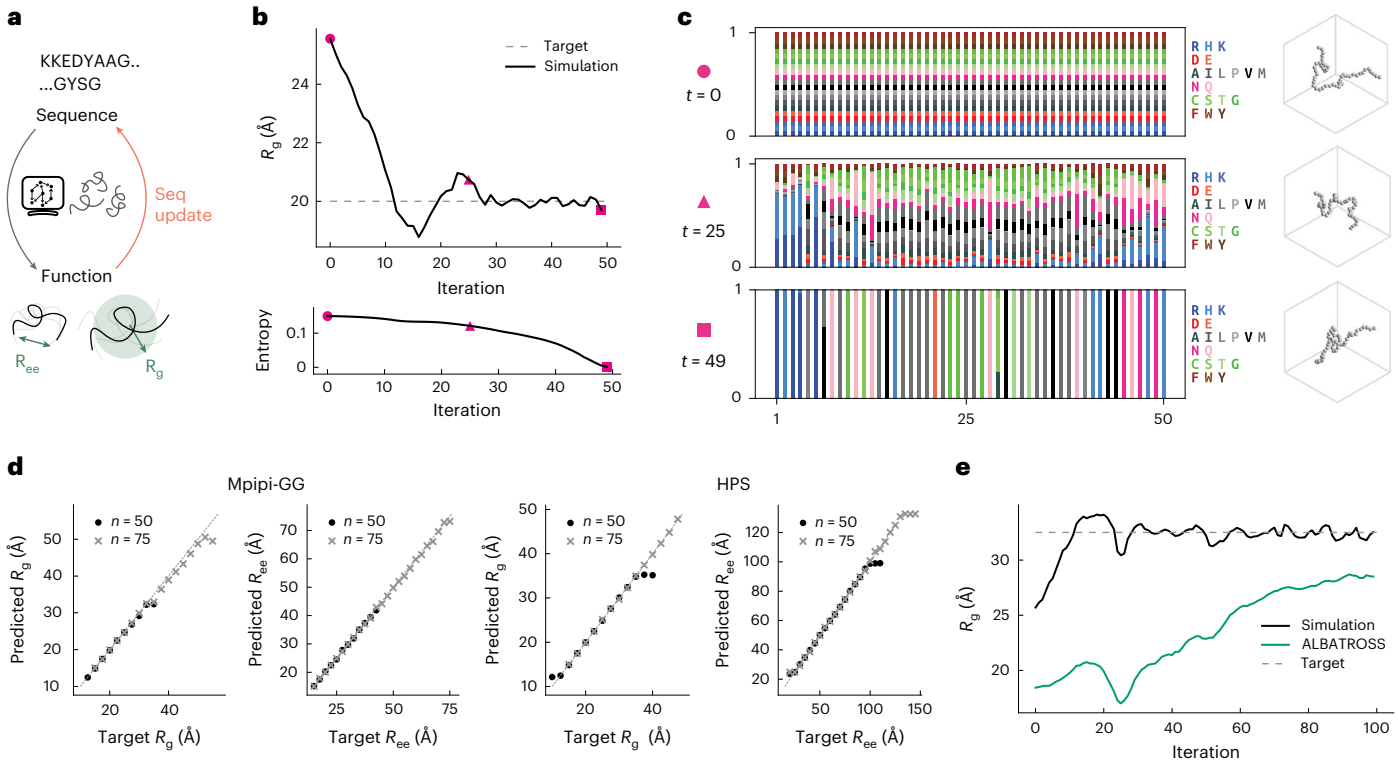

Fig. 2 | Designing IDPs with varying ensemble dimensions. a, We use the framework defined in Fig. 1 for design of IDPs with defined ensemble-averaged physical dimensions, specifically, R_g and R_{ee}. **b**, An example optimization to design an IDP of length $n = 50$ with $R_g = 20$ Å. The top panel represents the R_g from the simulated probabilistic sequence and the bottom panel represents the average sequence entropy at each position. Highlighted points (in pink) represent approximately the start, mid-point and end of the optimization. **c**, The evolution of the probabilistic sequence throughout the optimization depicted in **b** at highlighted points accompanied by a characteristic conformation in a box of side $a = 75$ Å. Each residue is colored differently and the column height corresponding to each residue position is the likelihood of being each residue. The probabilistic sequence is initialized as a uniform distribution of sequences, with each residue having an equal probability at each position, and the final sequence is nearly discrete. **d**, Each panel shows results for a set of optimizations, with each point comparing the predicted versus target ensemble dimension (R_g or R_{ee}) for a particular IDP sequence. The different panels highlight solutions for different sequence lengths ($n = 50, 75$) and for different force-fields (Mpipi-GG, left two panels; HPS, right two panels). **e**, The optimization trajectory for a sequence of length $n = 50$ for target $R_g = 35$ Å in which ALBATROSS under-values the R_g of the final optimized sequence by ~4 Å.

simulations of the entirely discrete designed sequence (Supplementary Section 3).

Designing IDPs with varying ensemble dimensions

Ensemble-averaged dimensions of an IDP, for example, the radius of gyration (R_g) or the end-to-end distance (R_{ee}), are coarse-grained metrics that reveal conformational biases that can correlate with binding and emergent phase behavior[20,30,31]. We therefore first set out to design an IDP of fixed sequence length ($n = 50$) with a target dimension of $\langle R_g \rangle = 20$ Å. We then update π in the direction of the desired $\langle R_g \rangle$ while simultaneously annealing—albeit gradually—towards a discrete sequence (Fig. 2c). Our routine converges (over 50 epochs and 2.5 h on an NVIDIA A100 GPU) to a sequence (Fig. 2b, Supplementary Data and Supplementary Section 3) that explores a range of conformations (Supplementary Fig. 3) with an ensemble-averaged R_g of ~20.1 Å. Re-running the optimization with varying random seeds leads to different sequences with similar values of R_g, highlighting the ability of our approach to identify multiple sequences that exhibit similar ensemble-averaged properties (Supplementary Fig. 3 and Supplementary Data).

With this framework, we are able to generate sequences of multiple lengths ($n = 50, n = 75$) across a wide R_g range (Fig. 2d). When we change the loss to correspond to a different physical property, that is, the end-to-end radius or R_{ee}—a dimension that provides insights into linker function in multi-domain proteins[9]—we are able to design IDPs across a wide R_{ee} range (Fig. 2d and Supplementary Section 4). Predictions of R_g and R_{ee} for designed sequences from the coarse-grained Mpipi model broadly correlate (Supplementary Fig. 4) with more fine-grained all-atom simulations (the GB99dms force-field is taken from ref. 32; see also Supplementary Section 5).

We find that the optima we obtain using this method are more accurate than those obtained with a pure machine-learned predictor derived from Mpipi-GG simulations (ALBATROSS)[23], when compared against the underlying molecular dynamics simulations for the ground truth (Supplementary Table 1). As an example, a sequence we generate ($n = 50$, $\langle R_g \rangle = 32.55$ Å, $\langle R_g \rangle^{\text{target}} = 32.5$ Å) is incorrectly predicted by ALBATROSS to be off by ~4 Å (Fig. 2e). A core strength of our algorithm is that by directly optimizing over simulations, we can explore a wider design space that is not subject to approximations underlying machine-learned descriptors. This means that more generally, our method can be flexibly applied to any force-field without requiring further data generation, architectural engineering, fine tuning or retraining of existing models.

We demonstrate this by designing IDPs of particular ensemble dimensions using the same method but with the HPS force-field[20]—a different commonly used pair potential (Fig. 2d). Although each force-field incorporates distinct biophysical priors, we find a broad correlation in predicted R_g for all sequences between the force-field they were optimized on as well as the one they were not optimized with (Supplementary Section 5 and Supplementary Fig. 4a). Together, our method provides a versatile approach to identify IDPs with specified conformation-averaged single-chain properties.

https://doi.org/10.1038/s43588-025-00881-y

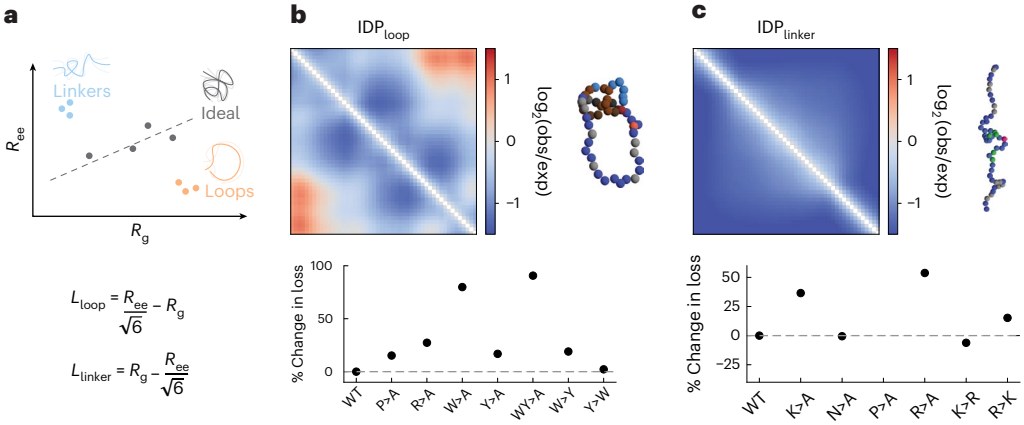

Fig. 3 | Shaping global conformational biases through loops and linker IDPs. **a**, A graphical illustration of ensemble coupling of R_g and R_{ee}, highlighting linear relationships for ideal homopolymer chains ($R_g = R_{ee}/\sqrt{6}$) and decoupled off-diagonal points for loops and linkers. Below, we show the loss we employ for the loop and linker design problems, expressed to maximize the decoupling between R_g and $R_{ee}/\sqrt{6}$. **b**,**c**, For the optimized loop (**b**) and linker (**c**) sequences, we depict the normalized contact frequencies computed over a trajectory. Representative configurations colored by amino acid identity are shown to the right (similar to Fig. 2c); the relative change in loss value for a set of key mutational scans are shown below. For contact frequencies, red (blue) regions represent higher (lower) expected frequencies when contrasted with an ideal polymer of identical length. The generic increase in loss following mutation demonstrates that our solution is highly optimized for the target property, as a higher loss corresponds to decreased agreement with the desired behavior.

De novo design of loops and linkers

We next asked whether we can construct IDPs with more complex descriptors of their conformational ensembles? In particular, we focused on designing sequence variants that maximized decoupling between R_g and R_{ee} as opposed to the linear scaling found in ideal polymers, unfolded proteins, and many naturally occurring IDPs[33,34]. We reasoned that such sequence variants could potentially represent optimally designed loops ($R_g - R_{ee}/\sqrt{6} >> 0$) or linkers ($R_g - R_{ee}/\sqrt{6} << 0$) (Fig. 3a and Supplementary Section 6).

For a sequence of fixed length ($L = 50$), we identify highly optimized loop and linker sequences with finely tuned mechanistic properties. Our loop optimization yields a low-complexity sequence with sticky aromatic patches comprising tryptophans (W) and tyrosines (Y) at either terminus—interspersed by prolines (P) and arginines (A) that kink out the intervening sequence—as highlighted by the normalized contact frequency maps and representative conformations (Fig. 3b). Although the underlying force-field predicts that W–W interactions are stickier and perhaps should thus drive stronger loops, mutating the mixture of Y/Ws in our solution to either all Ys or Ws leads to a less optimal loop (Fig. 3b and Supplementary Table 2). Similarly, mutational scans of each residue type into alanines or choosing less-complex losses lead to suboptimal loops (Supplementary Table 2)—generically reflecting an inability of simple sequence perturbations to decouple reductions in end-to-end distances from concomitant reductions in chain R_g. The optimal loop architecture here hence arises from tradeoffs between overall sequence composition and patterning and emergent many-body interactions. When optimizing for linkers, we find that low-complexity sequences that intersperse prolines amongst a backbone of positively charged arginines, maximally decoupling R_{ee} from R_g (Fig. 3c). This is largely expected as like-charges have short-range repulsive interactions; simple mutation scans are consistent with this intuition (Fig. 3c). Interestingly, we still identify a slightly improved linker variant in which lysines (K) are substituted with arginines (R). Overall, these design problems reinforce the ability of our algorithm to navigate high-dimensional sequence-spaces while balancing tradeoffs in ensemble properties.

Engineering IDPs with arbitrary sequence constraints

An important aspect of protein design is to engineer molecules that are subject to sequence constraints. For IDPs, such constraints could span requirements for highly disordered sequences, particular sequence compositions or motifs, or any other combinatorial sequence features. To incorporate arbitrary constraints, we generically expand our algorithmic framework by building on our previous work[35]. First, constraints are enforced through leaky ReLu functions multiplying the target property loss, resulting in gradients that navigate sequence space while maintaining constraints (Fig. 4a). Second, instead of directly optimizing over the sequence, we optimize over the weights of a pre-trained and fully connected neural network that parametrizes π (Fig. 4a). Together, this presents a modular and generalizable strategy to navigate constrained high-dimensional sequence spaces (Supplementary Section 7).

With this framework, we first set out to identify IDPs that are constrained to high disorder. We leverage a recent machine-learning-based disorder predictor, Metapredict[36], to measure and constrain disorder (Supplementary Section 7). Importantly, as the disorder prediction (and requirement) is only exact for a discrete sequence, the disorder contribution to the loss is gradually made more stringent over the optimization procedure (Fig. 4b). Designing compact proteins (meaning those with a small R_g) without any constraints tends to reveal highly hydrophobic proteins that are typically predicted to be well-folded and not disordered (see Fig. 4c). When we incorporate our disorder constraint, we are able to identify sequences that are simultaneously compact and highly disordered (Fig. 4b) across a range of R_g (Fig. 4c).

We next set out to design IDPs with compositional constraints. Motivated by a past work[37], we explored the effect of sequence patterning (particularly blockiness) on ensemble dimensions while keeping overall composition fixed at 50% positive and negative charges. To perform this multi-constraint optimization (Fig. 4d), we pre-train the overparameterized fully connected neural network to output a set of logits corresponding to the target charge distribution, and then use this in our constrained optimization procedure. Consistent with past predictions, we find an inverse relationship between ensemble dimensions and sequence blockiness (Fig. 4e).

Furthermore, to demonstrate the flexibility of our framework to accommodate experimental constraints, all subsequently designed sensors (Fig. 5 and Supplementary Fig. 7) are constrained to include a start codon amino acid and an N-terminal 6xHis-tag for affinity purification. This constraint is enforced by setting the first seven rows of π to a one-hot array denoting the MHHHHHH subsequence. In principle, our framework could accommodate any such fixed subsequence. Together,

https://doi.org/10.1038/s43588-025-00881-y

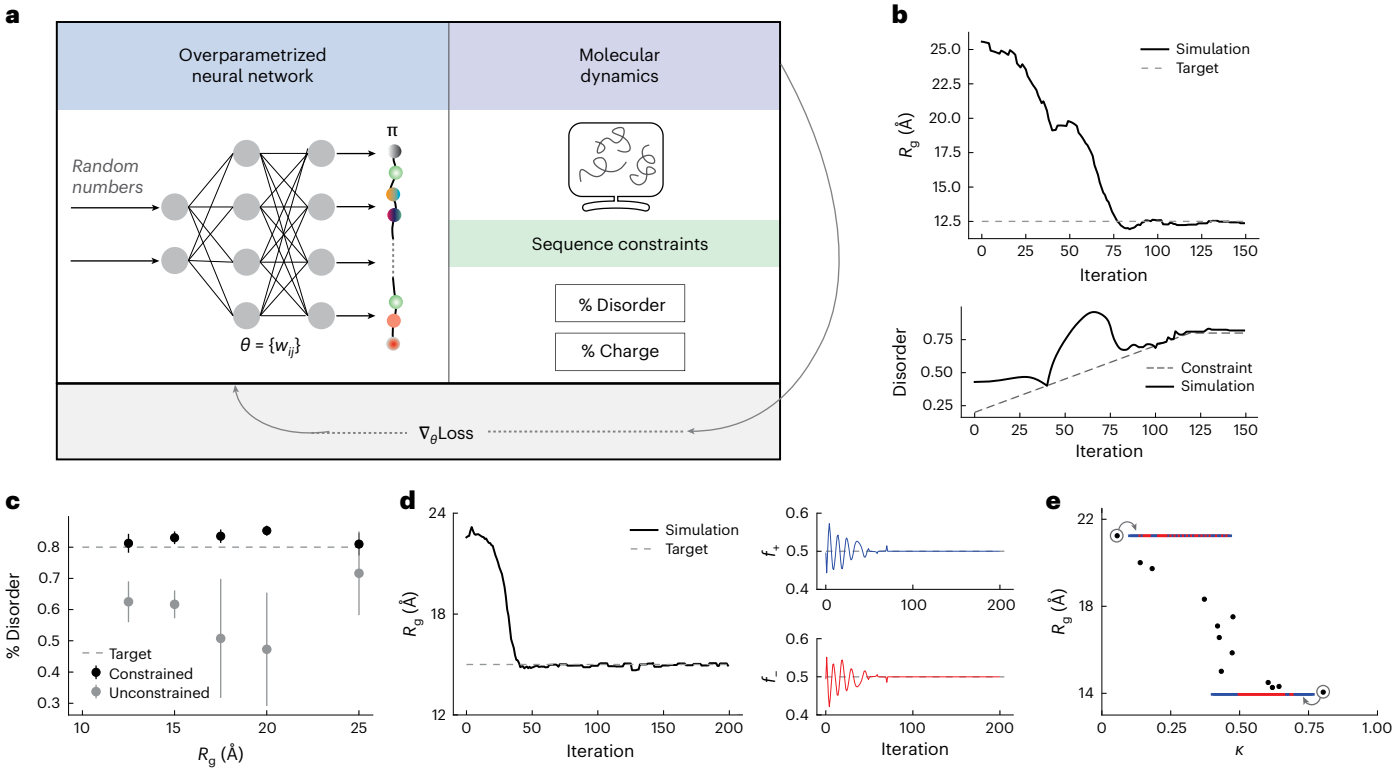

Fig. 4 | Engineering IDPs with arbitrary sequence constraints. a, Our framework for applying sequence constraints. Following ref. 35, we construct a loss function that incorporates arbitrary constraints on the probabilistic sequence and overparameterize the input to the optimization problem (the sequence representation) via a neural network. **b**, An example of IDP design ($n = 50$, $R_g = 12.5$ Å) subject to a constraint requiring a minimum degree of sequence disorder as predicted by Metapredict[36]. The top panel shows the simulation-predicted R_g over training epochs. The bottom panel shows the annealing of the sequence disorder constraint across the optimization. **c**, Average disorder of optimized sequences ($n = 50$, five replicates) versus target R_g value with (black) and without disorder constraints (grey). Dashed lines represent the threshold of enforced disorder constraint. Optimized sequences exhibit a R_g within 5 or 10% of target value for constrained or unconstrained optimizations. Dots represent mean values and whiskers represent the s.d. **d**, An example of IDP design ($n = 50$, $R_g = 17.5$ Å) subject to a constraint that requires 50% positively charged (R/K) and 50% negatively charged (E/D) residues. The constraints (right panels) and the target R_g (left panel) are achieved as the probabilistic sequence is annealed to a discrete sequence. **e**, R_g values of optimized sequences are plotted against κ values—a measure of sequence blockiness introduced in ref. 37—showing the same inverse relationship reported in ref. 37. The insets depict sequence patterning, represented by blue and red lines for positive and negative residues, respectively, and show relatively interspersed (blocky) IDPs for high (low) values of R_g. Each dot represents the most optimized sequence from five trajectories and have an R_g within ~10% of the target and charge ratios within ~5% of target.

our results demonstrate the ability of our model to design IDPs with multiple sequence-based constraints.

Programming stimuli-response in IDPs

A key biological function of many IDPs stems from their ability to sense and respond to cellular and environmental stimuli such as varying salt concentrations, temperature changes, dissolved CO_2 levels and pH[10,38] by changing global- or local-chain conformations. We therefore next decided to create IDP-based sensors, where we defined sensor function as arising from large changes in global conformation (R_g) in response to varying external stimuli (Fig. 5a and Supplementary Section 8). Our algorithm naturally handles such complex design formulations, which require tailored sequence–ensemble–function relationships across multiple conditions: for example, a salt-contractor IDP sensor must have high and low R_g at low and high salt concentrations, respectively. Thus the design optimization must find the sequence that achieves this goal simultaneously over both conditions.

We first began by designing sensors that respond to an increase in salt concentrations from 150 mM to 450 mM. We model the effect of increasing salt concentrations by only changing the screening length in the Coulomb pair potential (Supplementary Section 8) and, for simplicity, ignore other higher-order effects. By optimizing for a salt contractor, we identify a sequence rich in arginines with small clusters of interspersed tyrosine and tryptophan residues (Fig. 5b). The weakening of repulsive interactions between similarly charged arginines with salt leads to an effective and modest compaction, as exemplified by a poly-arginine sequence of identical length (Supplementary Table 4). In our solution, this passive contraction is amplified by the aromatic clusters, whose attraction is salt-insensitive. The periodic spacing and patterning of aromatic solutions in our designed variant only drives compaction under high salt conditions (Fig. 5b). All of the mutants that change composition or patterning show increased compaction, except for one, but they are no longer as salt-sensitive (Supplementary Table 4, Supplementary Fig. 6a and Supplementary Data).

This ability to exploit complex, many-body heteropolymer physics is even more dramatic in our salt-expander variant (Fig. 5c), designed to increase R_g with increasing salt concentration. Our algorithm converges to an expander with three roughly equally sized sequence modules: a positively charged N-terminus, a negatively charged C-terminus and a linker region that is made of proline spacers interspersed with sticky aromatic residues. The weakening of attractive interactions between positive and negative residues with salt only drives modest expansion, as seen in a $K_{25}E_{25}$ variant (Supplementary Table 4; 2.2 Å change). Two linker features, (1) sticky cation–π interactions with aromatic residues and the N-terminus and (2) steric effects from proline residues that reduce contact frequency of N/C termini, work in tandem to drive a

https://doi.org/10.1038/s43588-025-00881-y

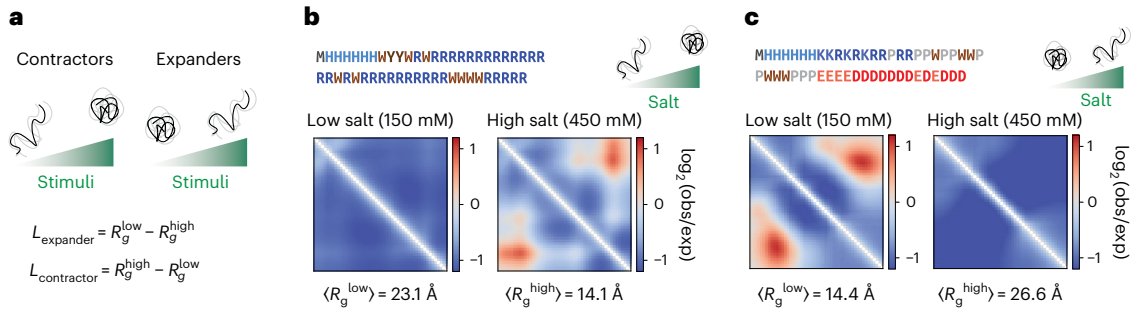

Fig. 5 | Programming stimuli-responsive IDP sensors. a, The stimuli-response is depicted by global change in ensemble dimensions across the stimulus (the green gradient bar could represent salt concentration, temperature or phosphorylation), and contractors or expanders represent the direction of change. Below the illustration we show the specific loss that is expressed as the difference of the two R_g values across the stimuli conditions, with the sign depending on whether we are designing a contractor or expander. **b**,**c**, Optimized contractor (**b**) and expander (**c**) sequences with contact frequency maps at low and high salt concentrations computed from representative trajectories; R_g is reported below. For contact frequencies, red (blue) regions represent higher (lower) expected frequencies when normalized with an ideal polymer of identical length.

salt-sensitive molecular clasp (Fig. 5c) with a nearly 100% change in R_g. Removal of any key features—for instance, through reducing steric hindrance by P → A mutations, or removing sticky residues—leads to a weaker salt response (Supplementary Fig. 6b). Our method therefore identifies a balance between salt-sensitive, salt-independent and steric features whose coupling transforms into a cooperative large-scale stimuli-response. This designed molecular clasp, in turn, sheds light on physical mechanisms that underlie sensitive and plastic conformational ensembles.

Finally, to highlight the generality of our model, we use a similar approach to construct sensors that respond by contraction or expansion to increases in temperature and phosphorylation of serine residues (Supplementary Fig. 7, Supplementary Table 3 and Supplementary Section 8). As the underlying force-fields do not accurately capture temperature-dependent variations in hydrophobic interactions, the effect sizes predicted by our model are rather small (Supplementary Fig. 7d–f). Incorporating temperature-dependent interactions, for instance, in the spirit of ref. 39, into our model framework will improve future sensor design. By contrast, we find an increasing range in sensor dimension change—and thus response size—with more phosphosites (Supplementary Fig. 7a–c). The mechanisms of contraction or expansion rely on interactions between phosphorylatable residues buried in neighborhoods of positively or negatively charged residues (Supplementary Fig. 7a–c). Thus, the addition of negatively charged groups following phosphorylation promotes favorable or repulsive interactions, leading to downstream change in IDP ensemble size. We note that in our simulations, phosphorylation is modeled using phosphomimetic substitutions (S → E) that capture charge effects but may not fully recapitulate the structural nuances of true phosphorylation.

Binders for disordered substrates

The function of many IDPs is driven by binding to disordered substrates, with examples of picomolar-level affinities in highly charged IDPs[30,40,41]. We next asked whether we can design disordered binders for a specific target substrate? To do this, we modify the forward simulation to include both the substrate, whose residue identity is fixed but can still sample a variety of conformations, along with a potential binding ligand whose sequence is learnable (Fig. 6a). Precise calculation of binding constants is computationally expensive, often requiring sophisticated enhanced sampling techniques. To overcome this, we make the following simplifications: (1) strong binders are identified by minimizing the average inter-strand distance and (2) a biasing potential is used to encourage collection of reference samples that are confined to an effective local volume (Supplementary Section 9). These simplifications help identify high-affinity binders but lose the ability to measure precise quantitative rates or constants.

We first seek a binder for a homopolymeric positively charged substrate R_{30}. Our model identifies a predominantly negatively charged ligand (>90% D/E residues, Supplementary Data) as a strong binder. Consistent with poly-electrolyte models, we find that predicted effective interaction coefficients[42,43] (Supplementary Section 9) are highly favorable for unlike-charge mediated substrate–ligand interactions (Supplementary Fig. 8c) and unfavorable, as expected, for like-charge-mediated substrate–substrate interactions. We next identify binders for the low-complexity domain of FUS, a well-studied IDP with prominent roles in human physiology and disease[44,45] and the poly-Q region of Whi3, an IDP with prominent roles in regulating nuclear autonomy and cell cycle in yeast[46]. As shown in Fig. 6b, our optimization leads to an identification of a target binder (Supplementary Data). Although both FUS-LC and Whi3 have strong self-affinity[44,46], effective interaction coefficients predict stronger interactions between our optimized binders and their respective substrates (Supplementary Fig. 8c) over the homotypic substrate–substrate interactions. In unbiased forward simulations, we observe strongly enriched intermolecular interactions for all binder-substrate pairs (Fig. 6d,e and Supplementary Fig. 8b), indicating strong binding at the micromolar concentrations we studied. Across all of the optimizations, we find that a sharp change in the learning dynamics (Fig. 6b,c and Supplementary Fig. 8a) is concomitant with strong binder identification. We expect that future studies will dissect whether this transition represents features of the underlying learning protocol, specifically the annealing schedule or noisy gradient signal due to limited sampling, or reflects the cooperative biophysics of such molecular binding events. Overall, our model lays the framework to generate candidate IDP binders for disordered substrates.

Discussion

Intrinsically disordered proteins and protein regions are biomolecules that are found across the tree of life; play critical roles in molecular recognition, cellular organization and information processing; and, when dysregulated, correlate with pathology. The sequence of an IDP encodes for a vast repertoire of interconverting spatial conformations that shape their emergent function. De novo design of IDPs with diverse and arbitrary properties remains limiting, in large part, due to lack of methods to generally invert the underlying sequence–ensemble–function relationship.

In this paper we introduce a computational framework to discover IDPs for a wide variety of target functions by rationally and efficiently inverting molecular simulations that capture the underlying sequence–ensemble relationship (Fig. 1). Using this framework, we first design IDP sequences with varying and complex coarse-grained ensemble dimension properties. Specifically, we design sequences across a range of R_g

https://doi.org/10.1038/s43588-025-00881-y

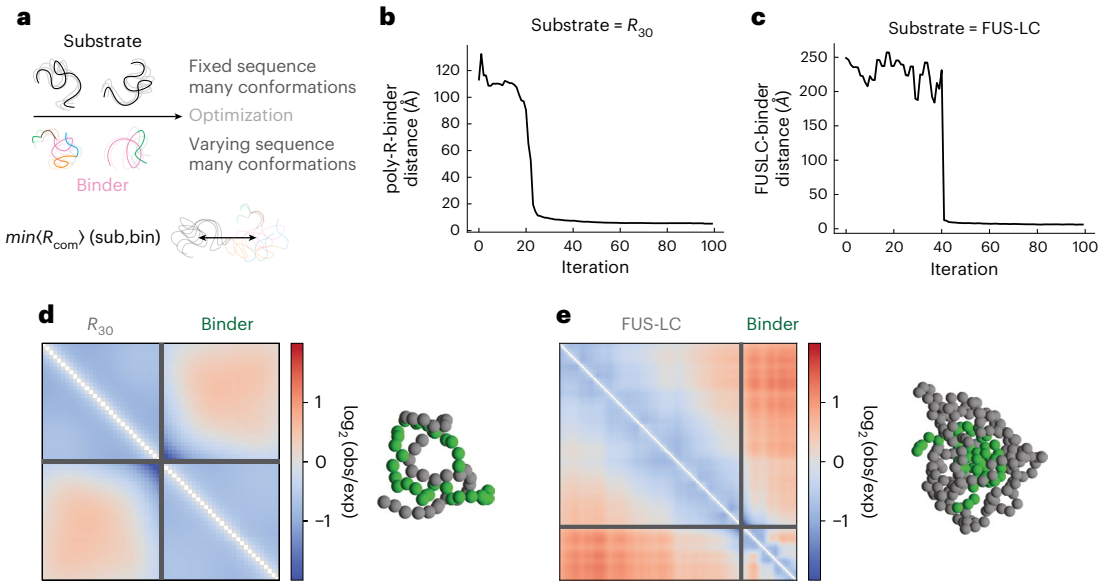

Fig. 6 | Designing IDP binders for disordered substrates. a, For a given substrate, binder design involves finding an IDP sequence that minimizes the interstrand center-of-mass distance between conformationally fluctuating IDPs. **b**, The optimization of a binder of length $n = 30$ for R_{30} is depicted by convergence to low interstrand distance over the trajectory. **c**, The optimization of a binder of length $n = 50$ for FUS-LC ($n = 169$) is depicted by convergence to low interstrand distance over the trajectory. **d**,**e**. Normalized contact frequency maps for R_{30} (**d**) and FUS-LC (**e**) are shown, highlighting both intramolecular and intermolecular interactions. For contact frequencies, red/blue regions represent higher/lower expected frequencies when contrasted with an ideal polymer of total length of binder + substrate. Representative bound snapshots of the binder (green) and substrate (grey) are depicted to the right of the panel; lines within the box separate the substrate and binder.

and R_{ee} (Fig. 2), and with tailored conformational biases such as loops and linkers (Fig. 3), properties that have been shown to shape cellular function[2,30]. We next develop a modular strategy to incorporate any sequence constraints in the design pipeline. With this, we engineer IDPs that are simultaneously compact and disordered, and generate sequence patterning variants with the same overall composition but differing ensemble dimensions (Fig. 4). With this framework, we next design highly sensitive sensors to multiple physicochemical and cellular stimuli such as salt, temperature and phosphorylation (Fig. 5). These designed sensors, in turn, shed light on the physical mechanisms by which the balance of competing intramolecular interactions governs large conformational changes. Finally, we use this method to identify disordered binders for low-complexity substrates (Fig. 6).

Although we focus on the Mpipi and HPS models in this work, a major strength of our framework is the flexibility to use alternative force-fields or study other physicochemical effects. The recently developed model of Rauh et al. for simulating phosphorylated IDPs[47] offers a more realistic force-field to describe phospho-effects. Similarly, combining our method with pH-dependent changes in the effective charge or protonation state of amino acids offers a route to develop pH sensors. Emerging data also indicate that small-molecules can successfully target aberrant condensates[48] and thus exploring methods to marry small-molecule design informed by molecular simulation is a broad area of interest. More generally, there is important potential to apply the outlined framework to distinct biomolecular sequences (proteins, RNA, DNA) with equilibrium sequence–ensemble–property relationships that can be predicted by a wide range of techniques spanning molecular dynamics, Monte Carlo simulations[49], field-theoretical approaches[50] and thermodynamics-informed models[42,43,51].

The framework we propose directly inverts simulation-derived sequence–ensemble relationships to drive de novo IDP design with tailored single-chain, binding and environmentally specific properties. A key aspect of this approach is the integration of continuous and relaxed sequence representations with molecular simulations, inspired by a host of recent efforts that invert analytical calculations or machine-learned approximations for biopolymer design[35,52,53]. Predictions via our framework, which require experimental tests, are fundamentally constrained by the accuracy of the underlying simulations. A key advantage of our approach is the underlying flexibility of gradient-based optimization, which in principle, can be leveraged to calculate gradients and optimize simulation parameters instead of sequence design. Namely, these same methods can be used in combination with experiments to drive an iterative loop to improve simulation accuracy that is benchmarked on multimodal experimental measurements of IDP properties. In a parallel paper we demonstrate how such an approach can improve simulation accuracy by fitting the parameters of a coarse-grained model of DNA to complex experimental data such as melting temperatures and stretch and torsional moduli[54].

Incorporation of emerging machine-learning approaches—for example, simulation-free generative methods to generate conformational ensembles[55,56]; combining alchemical and molecular dynamics simulations for sequence variant design with target single-chain properties[5,57]; and approximate machine-learning models that can rapidly invert pre-trained sequence–single-chain property relationships[23,57]—continues to expand the toolbox for protein engineering. Combining physics-based approaches with recent advances in differentiable programming holds promise for computational design and engineering for a wide variety of biomolecules and their functions.

Limitations of the study

Our paper introduces a framework to design IDPs with tailored equilibrium sequence–ensemble relationships modeled by simulations. First, as we compute gradient estimates via a reweighting scheme that relies on knowledge of unnormalized probabilities, our framework in its current form does not naturally accommodate far-from-equilibrium properties for which state-level probabilities are generally unknown. Opportunities to address such a limitation in future works include exploiting classic results in non-equilibrium statistical mechanics (for example, the Jarzynski equality), jointly learning the parameters of the attractor of a dynamical system (similar to actor-critic methods

in reinforcement learning) and alternative methods of automatic differentiation that sacrifice accuracy for numerical stability and memory overhead. Second, the convergence of this approach to niche sequence-designs has not been stress-tested and may require further algorithmic innovations. A particular challenge for convergence is the inequality between ensemble statistics computed via a continuous representation versus a distribution of discrete sequences sampled from the continuous sequence. Third, we only explored models for which the geometry of each particle identity is identical and probing models with polydisperse and complex geometries may require further methods development. Fourth, although we model increased salt concentrations by adjusting the Debye screening length, this approach neglects nonlinear and ion-specific effects that would require solving the full Poisson–Boltzmann equation. Although this is an often-used approximation, this probably limits accuracy of coarse-grained IDP modeling. Finally, our method is less appealing for the design of properties for which machine-learned approximations are comparable in accuracy (see the 'Computational performance and tradeoffs' section in the Methods).

Methods

General framework for optimizing particle identities

Consider a system of n particles in d dimensions, where each particle is ascribed one of m possible identities. Let $\vec{s} \in \mathbb{R}^n$ denote the identities of each particle where $\vec{s}_i \in \{1, 2, \cdots, m\}$. Given a potential energy function $U : \mathbb{R}^{n\times d} \to \mathbb{R}$ that depends on the particle identities, $\vec{s}$ determines the distribution of states in the canonical ensemble via $p(\vec{x};\vec{s}) \sim \exp(-\beta U(\vec{x};\vec{s}))$ where β is the inverse thermal energy and $\vec{x} \in \mathbb{R}^{n\times d}$. Given some state-level observable $O : \mathbb{R}^{n\times d} \to \mathbb{R}$, one is typically interested in the expected value of O in the entire ensemble, $\mathbb{E}[O(\vec{x})]_{\vec{x}\sim p(\cdot;\vec{s})}$. Consequently, we consider the optimization problem

$$\arg\min_{\vec{s}} \mathbb{E}[O(\vec{x})]_{\vec{x}\sim p(\cdot;\vec{s})} \tag{1}$$

Note that this is equivalent to the maximization or fixed point variants of the optimization problem.

We define an optimization framework for equation (1) that: (1) is general and makes minimal assumptions about the underlying model; (2) operates directly at the level of the model and requires no training; (3) yields an optimized probability distribution of identities from which discrete identities can be sampled; and (4) can be combined naturally with state-of-the-art machine-learning methods. Consider a matrix of particle identities, $\pi \in \mathbb{R}^{n\times m}$, where π_{ij} is the probability of the ith particle having identity j and $\sum_j \pi_{ij} = 1.0$ for all i. Let S denote the set of all possible discrete vectors of particle identities with $|S| = m^n$. We can then define the expected potential energy of a state $\vec{x}$ as

$$\mathbb{E}[U(\vec{x},\pi)] = \sum_{\vec{s}\in S} p(\vec{s}|\pi)U(\vec{x};\vec{s}) \tag{2}$$

where

$$p(\vec{s}|\pi) = \prod_{i=1}^{n} \pi_{i,\vec{s}_i} \tag{3}$$

This yields a corresponding distribution of states in the canonical ensemble,

$$p(\vec{x},\pi) \sim \exp(-\beta\mathbb{E}[U(\vec{x},\pi)]) \tag{4}$$

$$= \exp\left(-\beta \sum_{\vec{s}\in S} p(\vec{s}|\pi)U(\vec{x};\vec{s})\right) \tag{5}$$

$$= \prod_{\vec{s}\in S} \exp\left[-\beta\left(p(\vec{s}|\pi)U(\vec{x};\vec{s})\right)\right] \tag{6}$$

$$= \prod_{\vec{s}\in S} \left(\exp\left[-\beta U(\vec{x};\vec{s})\right]\right)^{p(\vec{s}|\pi)} \tag{7}$$

$$\sim \prod_{\vec{s}\in S} p(\vec{x};\vec{s})^{p(\vec{s}|\pi)} \tag{8}$$

Given this generalized probability distribution, we can generalize equation (1) for the case of probabilistic particle identities:

$$\arg\min_{\pi} \mathbb{E}[O(\vec{x})]_{\vec{x}\approx p(\cdot;\pi)} \tag{9}$$

Note that equation (9) reduces to equation (1) in the case where π is one-hot.

Crucially, π is a continuous variable and can be optimized via gradient descent. Given a stochastic sampler (for example, a Langevin integrator), one can compute $\nabla_\pi \mathbb{E}[O(\vec{x})]_{\vec{x}\sim p(\mathrm{x};\pi)}$ via differentiable trajectory reweighting (DiffTRE)[28]. Consider a set of states $\{\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_T\}$ sampled from the Boltzmann distribution defined by equation (4) for a reference state matrix $\hat{\pi}$. For values of π sufficiently close to $\hat{\pi}$ (see next section), we define a weight

$$w_i = \frac{\exp\left(-\beta\left[U(\vec{x}_i;\pi) - U(\vec{x}_i;\hat{\pi})\right]\right)}{\sum_j \exp\left(-\beta\left[U(\vec{x}_j;\pi) - U(\vec{x}_j;\hat{\pi})\right]\right)} \tag{10}$$

for each $\vec{x}_i$. We can then express our expectation in terms of these weights

$$\mathbb{E}[O(\vec{x})] \approx \sum_i w_i O(\vec{x}_i) \tag{11}$$

This yields an expression for $\mathbb{E}[O(\vec{x})]$ such that $\nabla_\pi \mathbb{E}[O(\vec{x})] \neq 0$. Note that $w_i = \frac{1}{T}$ in the limit where $\pi = \hat{\pi}$. Importantly, gradients are not computed through the unrolled trajectory (as in traditional differentiable molecular dynamics) but only through the energy function, relieving many of the numerical instabilities and memory constraints that typically plague differentiable molecular dynamics. This is equivalent to a low-variance REINFORCE gradient estimator by using knowledge of the unnormalized steady-state probabilities to effectively integrate over all paths yielding the same equilibrium state. Furthermore, the set of reference states must not be computed at every iteration (see the 'Differentiable Monte Carlo' section in the Methods), relaxing the computational cost imposed by running large simulations.

In practice, as the rows of π must be normalized, one optimizes a set of logits $\lambda \in \mathbb{R}^{n\times m}$ that are normalized in the loss function to yield π at each step, that is $\pi_i = \mathrm{softmax}(\lambda_i)$. As equation (9) reduces to equation (1) only when π is one-hot, we anneal π throughout the optimization by introducing a temperature term τ to the normalization procedure, that is $\pi_i = \mathrm{softmax}(\lambda_i/\tau)$. We find that a simple linear annealing scheme using $\tau_{\mathrm{start}} = 1.0$ and $\tau_{\mathrm{end}} = 0.01$ works well in most cases.

In the general case, sampling from the distribution defined by equation (2) is intractable because there are m^n possible permutations of state identities; however, this calculation becomes tractable in the case of an energy function in which the total energy is expressed as the sum of pairwise energies. Consider such an energy function for a fixed set of particle identities $\vec{s}$:

$$U_{\mathrm{tot}}(\vec{x};\vec{s}) = \sum_{i,j} U_{\mathrm{pair}}(\vec{x}_i,\vec{x}_j;\vec{s}_i,\vec{s}_j) \tag{12}$$

This can be generalized to the case of continuous particle identities:

$$\mathbb{E}[U_{\text{tot}}(\vec{x};\pi)] = \sum_{i,j}\mathbb{E}[U_{\text{pair}}(\vec{x}_i,\vec{x}_j;\pi)] \tag{13}$$

where

$$\mathbb{E}\left[U_{\text{pair}}(\vec{x}_i,\vec{x}_j;\pi)\right] = \sum_{\vec{s}_i,\vec{s}_j\in m}\pi(i,\vec{s}_i)\pi(j,\vec{s}_j) \times U_{\text{pair}}(\vec{x}_i,\vec{x}_j;\vec{s}_i,\vec{s}_j) \tag{14}$$

Crucially, all terms in equation (14) are independent and we can therefore rewrite $\mathbb{E}[U_{\text{pair}}(\vec{x}_i,\vec{x}_j;\pi)]$ as

$$\mathbb{E}\left[U_{\text{pair}}(\vec{x}_i,\vec{x}_j;\pi)\right] = \vec{U_{ij}}\cdot\vec{\pi_{ij}} \tag{15}$$

where

$$(\vec{U_{ij}})_{kl} = U_{\text{pair}}(\vec{x}_i,\vec{x}_j;k,l) \tag{16}$$

and

$$\vec{\pi_{ij}} = \pi_i \otimes \pi_j \tag{17}$$

where $\otimes$ denotes the Kronecker product. When performed in serial, the complexity of this calculation reduces to $\mathcal{O}(n^2m^2)$ and the n^2 factor can be further reduced by the use of neighbor lists. Crucially, however, the entire calculation can be highly parallelized on a modern GPU as the terms in equation (13) are independent. Although it is standard for coarse-grained models to be pairwise, this formulation could be extended to models with k-body interactions where the complexity of the expected energy calculation will scale as $\mathcal{O}(n^km^k)$ (before any neighbor list optimizations).

Differentiable Monte Carlo

Unlike a general reinforcement learning environment, we often know things about a physical system under study. Importantly, for example, we often know the probability distribution of the microstates of a given dynamical system. In this section we focus on the simple case of an equilibrium system in the canonical ensemble where the probability of state $\vec{x}_i$ is $\frac{e^{-\beta U(\vec{x}_i)}}{Z}$, where β is the inverse thermal energy, $U(\vec{x}_i)$ is the potential energy of $\vec{x}_i$ and $Z = \sum_j e^{-\beta U(\vec{x}_j)}$ is the partition function.

Consider a set of states sampled from this distribution via some control parameters θ, $X_\theta = \{\vec{x}_1,\vec{x}_2,\cdots\vec{x}_N\}$. Note that there are many schemes for efficiently sampling from the Boltzmann distribution such as standard molecular dynamics and Monte Carlo algorithms, and even generative deep learning methods. Examples of θ are parameters of the potential energy or parameters of the initial conditions. Via ergodicity, we can compute the expectation of some state-level observable $O(\vec{x},\theta)$ as

$$\langle O(\vec{x},\theta)\rangle_{\vec{x}_i\in X} = \frac{1}{N}\sum_i O(\vec{x}_i,\theta) \tag{18}$$

This time, our expectation is defined with respect to a set of sampled states (whose probability distribution we know) rather than with respect to a set of trajectories (or equivalently, random seeds). When formulated in this fashion, our calculation of the expectation has no history dependence; we do not care how the states are sampled, only that they are sampled from the underlying distribution.

However, we cannot immediately compute an accurate gradient of equation (18). Although we know that the relative probabilities of each microstate will change as we change θ, we lose this dependence in our gradient signal by only considering the final set of sampled states as $\nabla_\theta\frac{1}{N} = 0$. To recover this signal, Zhang et al.[29] and Thaler and Zavadlav[28] independently introduced a simple reweighting scheme (termed differentiable trajectory reweighting, or DiffTRE by the latter publication) in which we rewrite equation (18) as

$$\langle O(\vec{x},\theta)\rangle_{\vec{x}_i\in X} = \sum_i w_i O(\vec{x}_i,\theta) \tag{19}$$

where

$$w_i = \frac{p_\theta(\vec{x}_i)/p_{\hat{\theta}}(\vec{x}_i)}{\sum_j p_\theta(\vec{x}_j)/p_{\hat{\theta}}(\vec{x}_j)} \tag{20}$$

and $\hat{\theta}$ is the reference potential via which X_θ was sampled. Equation (20) only requires unnormalized probabilities as the normalizing factors cancel. For example, in the case of the canonical ensemble, equation (20) does not require knowledge of the partition function:

$$w_i = \frac{e^{-\beta(U_\theta(\vec{x}_i)-U_{\hat{\theta}}(\vec{x}_i))}}{\sum_j e^{-\beta(U_\theta(\vec{x}_j)-U_{\hat{\theta}}(\vec{x}_j))}} \tag{21}$$

Crucially, in the case in which $\theta = \hat{\theta}$, $w_i = \frac{1}{N}$ but $\nabla_\theta \log(p(\vec{x}_i)) \neq 0$, Thaler and Zavadlav introduced the notion that reference states collected via $\hat{\theta}$ can be reused for small differences between θ and $\hat{\theta}$, but as this difference grows few states dominate the average and the reference states should be resampled. This is captured via an expression for effective sample size:

$$N_{\text{eff}} = e^{-\sum_{i=1}^{N} w_i \ln(w_i)} \tag{22}$$

Refer to ref. 29 and ref. 28 for a complete introduction to this method.

This reweighting scheme solves three major problems in differentiable programming for dynamical systems. Foremost, it resolves both problems related to memory, and numerical instability as gradients are no longer computed with respect to unrolled trajectories. However, there is a third benefit: the entire sampling procedure does not have to be rewritten in an automatic differentiation framework. Instead, one only must write the energy function in such a framework. Furthermore, objective functions that do not explicitly depend on θ also do not have be differentiable, permitting the immediate use of the rich ecosystem of libraries that already exist for the analysis of molecular dynamics trajectories. This reduces a massive barrier to entry for differentiable programming in cases where the unnormalized probability of sampled states is known, particularly as it relates to larger and more complex code bases.

In the language of stochastic gradient estimators, DiffTRE can be regarded as a low-variance REINFORCE estimator. A traditional REINFORCE estimator would regard the probability of each state as the probability of its corresponding trajectory, drastically inflating the variance of the estimator as many trajectories can yield the same equilibrium state. DiffTRE permits us to use our knowledge about the distribution from which we are sampling in our estimate of the gradient, effectively integrating over all trajectories for a given state.

Mpipi force-field

Mpipi is a coarse-grained model of protein–protein and protein–RNA interactions for studying biomolecular liquid–liquid phase separation[21]. Introduced in 2021, Mpipi has gained widespread popularity for the computational study of liquid–liquid phase separation and the underlying biophysics[58–62]. Recent machine-learning methods use Mpipi to generate ground-truth training data, with which neural networks are trained to either predict ensemble properties or generate sequences with target characteristics[23,63]. Note that such methods for inverse design are limited not only because they generate sequences with respect to a learned approximation of Mpipi rather than Mpipi itself,

but also because in principle designing sequences for a different target ensemble property demands an entirely new deep learning model.

In Mpipi, each amino acid monomer is represented a single isotropic sphere. Each amino acid type is assigned a mass, diameter, charge and energy scale. Like oxDNA, all interactions are pairwise and the potential energy is given by

$$V_{\text{Mpipi}} = \sum_{\text{nn}} V_{\text{bond}} + \sum_{\text{other pairs}} (V_{\text{elec}} + V_{\text{pair}}) \quad (23)$$

where nn denotes a fixed set of consecutive bonded pairs. V_{bond} is computed as a harmonic bond potential, V_{elec} as a Coulomb term with Debye–Hückel electrostatic screening and V_{pair} as a Wang–Frenkel interaction[64]. The parameters of this potential were fit to reproduce both the atomistic potential-of-mean-force calculations, and the bioinformatics data. We modulate salt concentration effects through changing ionic strength and thus adjusting the Debye screening length in the Coulomb term. Although only an approximation, accurately modeling high salt concentrations would in principle require solving the nonlinear Poisson–Boltzmann equation and more explicit treatments of ion distributions that are beyond the reach of coarse-grained approaches (refer to ref. 21 for complete details of the model and its parameterization, and ref. 23 for a description of the modified parameters used in this work).

Simulations

All simulations were performed in JAX-MD[27] on an NVIDIA A100 80 GB GPU. We used a Langevin thermostat with a timestep of 10 fs at standard conditions of 300 K and 150 mM salt concentration unless specified otherwise. Forces are computed via automatic differentiation, circumventing the need to manually derive forces for the expected Hamiltonian over all discrete sequences. Specific simulation parameters (for instance, equilibration time, simulation length, sample frequency) are provided for each optimization in the Supplementary Information. Importantly, the parameters above are designed to ensure that simulated trajectories are uncorrelated to initial conformation and run long enough to sufficiently sample the equilibrium reference ensemble (Supplementary Fig. 2 and Supplementary Section 3). Note that although we use molecular dynamics simulations in this work, our design framework is agnostic to the method of obtaining reference states (refer to the 'Code availability' section for details on the code used).

Computational performance and tradeoffs

Forward simulations of probabilistic sequences scales near-linearly with sequence length and only incurs a modest cost (Supplementary Fig. 1a,b) over the discrete counterpart on GPUs (see Supplementary Section 2). Performance on CPUs is overall much less efficient due to lack of parallelism.

More generally, in the following we report key computational tradeoffs and considerations in using this method. Each optimization for R_g at typical conditions (300 K, 150 mM salt concentration, $L_{\text{seq}} = 50$) in this work requires several hours of compute on a single GPU. By contrast, using ALBATROSS—that is, the machine-learned predictor used for comparison in Fig. 2—for design, as in ref. 24, only takes tens of seconds. We next compare with a method that directly operates at the level of molecular simulation that however does not employ a continuous sequence space representation. In Pesce and colleagues' design framework, which performs a Monte Carlo search over a discrete sequence space, and applies alchemical calculations to minimize the need to re-simulate, a representative optimization requires 4,500 iterations and 20 days[57].

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All optimized sequences are provided in the Supplementary Data 1. Source data are provided with this paper.

Code availability

The complete codebase is available at the following GitHub repository: https://github.com/rkruegs123/idp-design. This repository includes a notebook containing a scaffold of a simple optimization for a custom state-level property. A snapshot of this repository, including the full source code and corresponding documentation, has also been archived on Zenodo at https://doi.org/10.5281/zenodo.15311353 (ref. 65).

References

1. Boehr, D. D., Nussinov, R. & Wright, P. E. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* **5**, 789–796 (2009).
2. Holehouse, A. S. & Kragelund, B. B. The molecular basis for cellular function of intrinsically disordered protein regions. *Nat. Rev. Mol. Cell Biol.* **25**, 187–211 (2024).
3. van der Lee, R. et al. Classification of intrinsically disordered regions and proteins. *Chem. Rev.* **114**, 6589–6631 (2014).
4. Uversky, V. N. Recent developments in the field of intrinsically disordered proteins: intrinsic disorder–based emergence in cellular biology in light of the physiological and pathological liquid–liquid phase transitions. *Ann. Rev. Biophys.* **50**, 135–156 (2021).
5. Tesei, G. et al. Conformational ensembles of the human intrinsically disordered proteome. *Nature* **626**, 897–904 (2024).
6. Thomasen, F. E. & Lindorff-Larsen, K. Conformational ensembles of intrinsically disordered proteins and flexible multidomain proteins. *Biochem. Soc. Trans.* **50**, 541–554 (2022).
7. Mittag, T. & Forman-Kay, J. D. Atomic-level characterization of disordered protein ensembles. *Cur. Opin. Struct. Biol.* **17**, 3–14 (2007).
8. Davey, N. E., Simonetti, L. & Ivarsson, Y. The next wave of interactomics: mapping the SLiM-based interactions of the intrinsically disordered proteome. *Curr. Opin. Struct. Biol.* **80**, 102593 (2023).
9. Huang, Q., Li, M., Lai, L. & Liu, Z. Allostery of multidomain proteins with disordered linkers. *Curr. Opin. Struct. Bio.* **62**, 175–182 (2020).
10. Moses, D., Ginell, G. M., Holehouse, A. S. & Sukenik, S. Intrinsically disordered regions are poised to act as sensors of cellular chemistry. *Trends Biochem. Sci.* **48**, 1019–1034 (2023).
11. Banani, S. F. et al. Genetic variation associated with condensate dysregulation in disease. *Develop. Cell* **57**, 1776–1788.e8 (2022).
12. Shrinivas, K. et al. Enhancer features that drive formation of transcriptional condensates. *Mol. Cell* **75**, 549–561 (2019).
13. Sabari, B. R. Biomolecular condensates and gene activation in development and disease. *Develop. Cell* **55**, 84–96 (2020).
14. Shi, M., Zhang, P., Vora, S. M. & Wu, H. Higher-order assemblies in innate immune and inflammatory signaling: a general principle in cell biology. *Cur. Opin. Cell Biol.* **63**, 194–203 (2020).
15. Tsang, B., Pritišanac, I., Scherer, S. W., Moses, A. M. & Forman-Kay, J. D. Phase separation as a missing mechanism for interpretation of disease mutations. *Cell* **183**, 1742–1756 (2020).
16. Dauparas, J. et al. Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).
17. Mirdita, M. et al. ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
18. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *nature* **596**, 583–589 (2021).
19. Krishna, R. et al. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science* **384**, eadl2528 (2024).

20. Dignon, G. L., Zheng, W., Best, R. B., Kim, Y. C. & Mittal, J. Relation between single-molecule properties and phase behavior of intrinsically disordered proteins. *Proc. Natl Acad. Sci. USA* **115**, 9929–9934 (2018).
21. Joseph, J. A. et al. Physics-driven coarse-grained model for biomolecular phase separation with near-quantitative accuracy. *Nat. Comput. Sci.* **1**, 732–743 (2021).
22. Tesei, G., Schulze, T. K., Crehuet, R. & Lindorff-Larsen, K. Accurate model of liquid–liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties. *Proc. Natl Acad. Sci. USA* **118**, e2111696118 (2021).
23. Lotthammer, J. M., Ginell, G. M., Griffith, D., Emenecker, R. J. & Holehouse, A. S. Direct prediction of intrinsically disordered protein conformational properties from sequence. *Nat. Methods* **21**, 465–476 (2024).
24. Emenecker, R. J., Guadalupe, K., Shamoon, N. M., Sukenik, S. & Holehouse, A. S. Sequence–ensemble–function relationships for disordered proteins in live cells. Preprint at *bioRxiv* https://doi.org/10.1101/2023.10.29.564547 (2023).
25. Regy, RoshanMammen, Thompson, J., Kim, Y. C. & Mittal, J. Improved coarse-grained model for studying sequence dependent phase separation of disordered proteins. *Protein Sci.* **30**, 1371–1379 (2021).
26. Bradbury, J. et al. JAX: composable transformations of Python+ Numpy programs. *GitHub* https://github.com/jax-ml/jax (2018).
27. Schoenholz, S. & Cubuk, E. D. JAX MD: a framework for differentiable physics. In *Proc. 34th International Conference on Neural Information Processing System* Vol. 33, 11428–11441 (2020).
28. Thaler, S. & Zavadlav, J. Learning neural network potentials from experimental data via differentiable trajectory reweighting. *Nat. Commun.* **12**, 6884 (2021).
29. Zhang, Shi-Xin, Wan, Zhou-Quan & Yao, H. Automatic differentiable Monte Carlo: theory and application. *Phys. Rev. Res.* **5**, 033041 (2023).
30. González-Foutel, NicolásS. et al. Conformational buffering underlies functional selection in intrinsically disordered protein regions. *Nat. Struct. Mol. Biol.* **29**, 781–790 (2022).
31. Lin, Yi-Hsuan & Chan, HueSun Phase separation and single-chain compactness of charged disordered proteins are strongly correlated. *Biophys. J.* **112**, 2043–2046 (2017).
32. G. Greener, J. Differentiable simulation to develop molecular dynamics force fields for disordered proteins. *Chem. Sci.* **15**, 4897–4909 (2024).
33. Mugnai, M. L. et al. Sizes, conformational fluctuations, and SAXS profiles for intrinsically disordered proteins. *Protein Sci.* **34**, e70067 (2025).
34. Riback, J. A. et al. Commonly used FRET fluorophores promote collapse of an otherwise disordered protein. *Proc. Natl Acad. Sci. USA* **116**, 8889–8894 (2019).
35. Krueger, R. K. & Ward, M. JAX-RNAfold: scalable differentiable folding. *Bioinformatics* **41**, btaf203 (2025).
36. Emenecker, R. J., Griffith, D. & Holehouse, A. S. Metapredict V2: an update to metapredict, a fast, accurate, and easy-to-use predictor of consensus disorder and structure. Preprint at *bioRxiv* https://www.biorxiv.org/content/10.1101/2022.06.06.494887v1 (2022).
37. Das, R. K. & Pappu, R. V. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc. Natl Acad. Sci. USA* **110**, 13392–13397 (2013).
38. Zhang, M. et al. The intrinsically disordered region from PP2C phosphatases functions as a conserved CO_2 sensor. *Nat. Cell Biol.* **24**, 1029–1037 (2022).
39. Dignon, G. L., Zheng, W., Kim, Y. C. & Mittal, J. Temperature-controlled liquid–liquid phase separation of disordered proteins. *ACS Central Sci.* **5**, 821–830 (2019).
40. Borgia, A. et al. Extreme disorder in an ultrahigh-affinity protein complex. *Nature* **555**, 61–66 (2018).
41. Schuler, B. et al. Binding without folding—the biomolecular function of disordered polyelectrolyte complexes. *Curr. Opin. Struc. Biol.* **60**, 66–76 (2020).
42. Adachi, K. & Kawaguchi, K. Predicting heteropolymer interactions: demixing and hypermixing of disordered protein sequences. *Phys. Rev. X* **14**, 031011 (2024).
43. Ginell, G. M. et al. Sequence-based prediction of intermolecular interactions driven by disordered regions. *Science* **388**, eadq8381 (2025).
44. Portz, B., Lee, Bo. Lim & Shorter, J. FUS and TDP-43 phases in health and disease. *Trends Biochem. Sci.* **46**, 550–563 (2021).
45. Wang, J. et al. A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins. *Cell* **174**, 688–699.e16 (2018).
46. Roden, C. & Gladfelter, A. S. RNA contributions to the form and function of biomolecular condensates. *Nat. Rev. Mol. Cell Biol.* **22**, 183–195 (2021).
47. Rauh, A. S., Hedemark, G. S., Tesei, G. & Lindorff-Larsen, K. A coarse-grained model for simulations of phosphorylated disordered proteins. *Biophys J.* https://doi.org/10.1016/j.bpj.2025.07.001 (2025).
48. Kilgore, H. R. et al. Distinct chemical environments in biomolecular condensates. *Nat. Chem. Biol.* **20**, 291–301 (2024).
49. Choi, Jeong-Mo & Pappu, R. V. Improvements to the ABSINTH force field for proteins based on experimentally derived amino acid specific backbone conformational statistics. *J. Chem. Theory Comput.* **15**, 1367–1382 (2019).
50. Wessén, J., Das, S., Pal, T. & Chan, HueSun Analytical formulation and field-theoretic simulation of sequence-specific phase separation of protein-like heteropolymers with short- and long-spatial-range interactions. *J. Phys. Chem. B* **126**, 9222–9245 (2022).
51. Shrinivas, K. & Brenner, M. P. Phase separation in fluids with many interacting components. *Proc. Natl Acad. Sci. USA* **118**, e2108551118 (2021).
52. Frank, C. et al. Scalable protein design using optimization in a relaxed sequence space. *Science* **386**, 439–445 (2024).
53. Matthies, M. C., Krueger, R., Torda, A. E. & Ward, M. Differentiable partition function calculation for RNA. *Nucleic Acids Res.* **52**, e14 (2024).
54. Krueger, R. K., Engel, M. C., Hausen, R. & Brenner, M. P. Fitting coarse-grained models to macroscopic experimental data via automatic differentiation. Preprint at https://arxiv.org/abs/2411.09216 (2025).
55. Janson, G. & Feig, M. Transferable deep generative modeling of intrinsically disordered protein conformations. *PLoS Comput. Biol.* **20**, e1012144 (2024).
56. Liu, C. et al. Diffusing protein binders to intrinsically disordered proteins. *Nature* **644**, 809–817 (2025)
57. Pesce, F. et al. Design of intrinsically disordered protein variants with diverse structural properties. *Sci. Adv.* **10**, eadm9926 (2024).
58. Sanchez-Burgos, I., Espinosa, J. R., Joseph, J. A. & Collepardo-Guevara, R. RNA length has a non-trivial effect in the stability of biomolecular condensates formed by RNA-binding proteins. *PLoS Comput. Biol.* **18**, e1009810 (2022).
59. Zhu, H. et al. The chromatin regulator HMGA1a undergoes phase separation in the nucleus. *ChemBioChem* **24**, e202200450 (2023).

60. Alston, J. J., Soranno, A. & Holehouse, A. S. Conserved molecular recognition by an intrinsically disordered region in the absence of sequence conservation. *Biophys. J.* **123**, 26a (2024).
61. Wessén, J., Das, S., Pal, T. & Chan, HueSun Analytical formulation and field-theoretic simulation of sequence-specific phase separation of protein-like heteropolymers with short-and long-spatial-range interactions. *J. Phys. Chem. B* **126**, 9222–9245 (2022).
62. Garaizar, A. et al. Aging can transform single-component protein condensates into multiphase architectures. *Proc. Natl Acad. Sci. USA* **119**, e2119800119 (2022).
63. Taneja, I. & Lasker, K. Machine learning based methods to generate conformational ensembles of disordered proteins. *Biophys. J.* **123**, 101–113 (2023).
64. Wang, X., Ramírez-Hinestrosa, Simón, Dobnikar, J. & Frenkel, D. The Lennard–Jones potential: when (not) to use it. *Phys. Chem. Chem. Phys.* **22**, 10624–10633 (2020).
65. Krueger, R. K. & Shrinivas, K. rkruegs123/idp-design: file format change for figures. *Zenodo* https://doi.org/10.5281/zenodo.15311353 (2025).

Acknowledgements

We thank M. Ward for his collaboration on sequence design via overparameterization in the context of differentiable RNA folding, which inspired this work, J. Smith for helpful discussions relating to stochastic gradient estimators, W. Snead for helpful discussions on IDP design, and J. Boodry, N. Tyagi and members of the Shrinivas laboratory, for discussions and experimentation with the codebase. We acknowledge support from the Simons Foundation through the Simons Foundation Investigator award (R.K.K., M.P.B. and K.S). We acknowledge support from the NSF AI Institute of Dynamic Systems (grant no. 2112085), the Office of Naval Research (grant no. N00014-17-1-3029) and the Harvard Materials Research Science and Engineering Center (DMR no. 20-11754 to R.K.K and M.P.B.). We acknowledge support from NSF-Simons Center for Mathematical and Statistical Analysis of Biology at Harvard (grant no. 1764269) and Northwestern University for startup funding (K.S.). The computations in this paper were in part run on the FASRC Cannon cluster supported by the FAS Division of Science Research Computing Group at Harvard University. This research was supported in part through the computational resources and staff contributions provided for the Quest high performance computing facility at Northwestern University.

Author contributions

R.K.K., M.P.B. and K.S. designed the study. R.K.K. and K.S. performed the research. All authors contributed to the writing and revision of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43588-025-00881-y.

Correspondence and requests for materials should be addressed to Michael P. Brenner or Krishna Shrinivas.

Peer review information *Nature Computational Science* thanks Mikael Lund and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Kaitlin McCardle, in collaboration with the *Nature Computational Science* team. Peer reviewer reports are available.