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

Ryan Krueger [1], **Michael P. Brenner** [1,2,*], **and Krishna Shrinivas** [3,4,*]

[1]School of Engineering and Applied Sciences, Harvard University, 29 Oxford St, Cambridge, MA 02138
[2]Department of Physics, Harvard University, 17 Oxford St, Cambridge, MA 02138
[3]Department of Chemical and Biological Engineering, Northwestern University, 2145 Sheridan Road, Evanston, IL, 60208
[4]Center for Synthetic Biology, Northwestern University, 2145 Sheridan Rd, Evanston, IL 60208
[*]correspondence to: krishna@northwestern.edu, brenner@seas.harvard.edu

The design of folded proteins has advanced significantly in recent years. However, many proteins and protein regions are intrinsically disordered (IDPs) and lack a stable fold i.e., the sequence of an 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 to 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.

Introduction

The basis of biomolecular function is often specified by a sequence which 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). IDPs lack a stable 3D structure, rather, they dynamically interconvert between a large range of non-random conformations (5–7) whose local and global properties shape cellular functions (2). IDPs 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, that play central roles in cellular signaling and information processing. Finally, dysregulation of IDPs and IDP-dependent interactions are increasingly correlated with multiple pathological states (11, 15). Thus, there is widespread interest to design 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 have had limited ability for designing disordered proteins. Structures of IDPs are not characterized by single stable folds, rather, they 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) (e.g. radius of gyration and polymer exponents) and can be subsequently inverted for design (24). While each method has found success, using separate algorithms for the forward and inverse problems reduces accuracy and generalizability to different target properties and force field parameters. It would be far 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 through inverting molecular simulations. Our framework uses gradient-based optimization on molecular simulations for designing 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 and complexity of ensemble dimensions, 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 concentrations of modification-driving enzymes. 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*. Our proposed method, while generically potential-agnostic, 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.

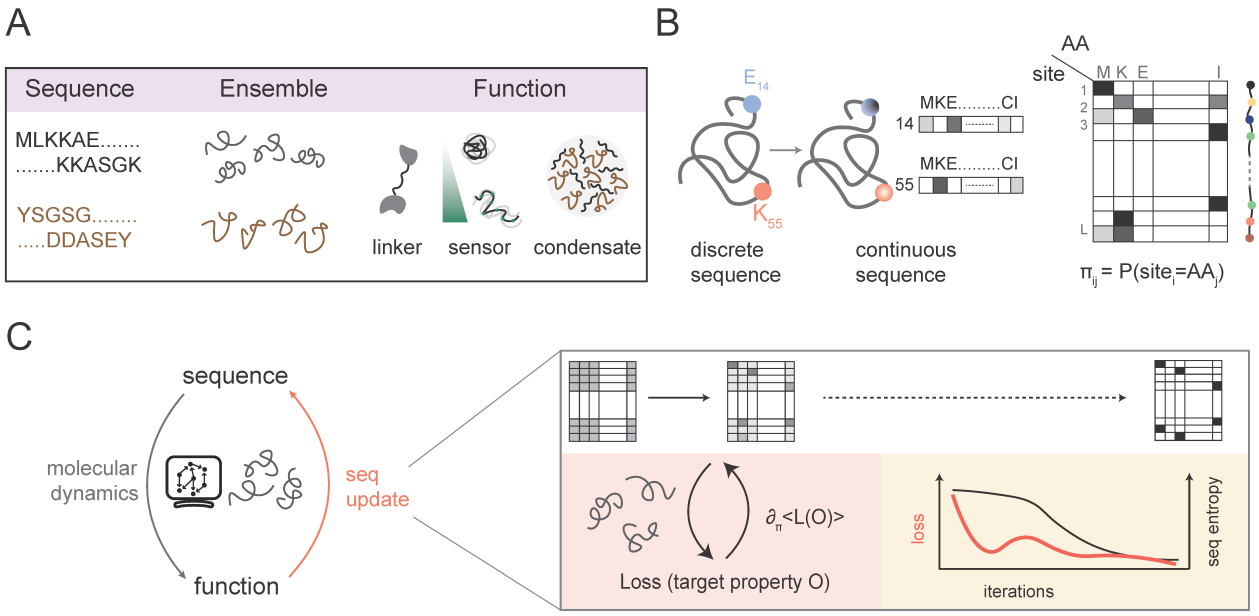

Fig. 1. Method for inverse design of IDPs
A. The amino acid sequence of an IDP encodes for an ensemble of dynamic 3D conformations structures that determines properties shaping molecular and cellular functions.
B. A discrete IDP sequence, a vector of length n where each position is typically a categorically 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 (see 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.

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 (Figure 1A) and (2) an algorithm to "invert" this through directed search of sequence space towards a desired functional property. Over the last few years, coarse-grained molecular simulations with custom pair potentials have made (5, 21, 25) dramatic improvements in predicting effective ensemble properties of IDPs. In this paper, we focus on molecular dynamics simulations using 1 AA=1 bead coarse-graining with the Mpipi-GG (see Methods and SI Note 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: $\partial_{\text{seq}}\mathcal{L}(\langle P_1^{\text{sim}}\rangle, \langle P_2^{\text{sim}}\rangle, \cdots)$. Since 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 Methods). For a sequence of L residues, this continuous probabilistic representation (Figure 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 i.e., 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 SI for derivation, Figure S2).

While in principle libraries like JAX-MD enable gradient calculation over unrolled MD trajectories, this is slow, scales poorly with system size, and is plagued by numerical instability (Figure S1, SI Note 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. This calculation provides significant speedup and accuracy in gradient estimation and allows reuse of simulation snapshots for multiple sequence updates. Finally, we incorporate an annealing procedure that gradually forces π to become increasingly discrete through the optimization (Figure 1C, see Methods). Unless otherwise specified, we initialize all optimizations with a uniform distribution.

Designing IDPs with varying ensemble dimensions

Ensemble-averaged dimensions of an IDP, for e.g., the radius of gyration (R_g) or the end-to-end radius (R_{ee}), are coarse-grained metrics that reveal conformational biases which can correlate with binding and emergent phase behavior (20, 30, 31). Therefore, we first set out to design an IDP of fixed sequence length ($n = 50$) with a target di-

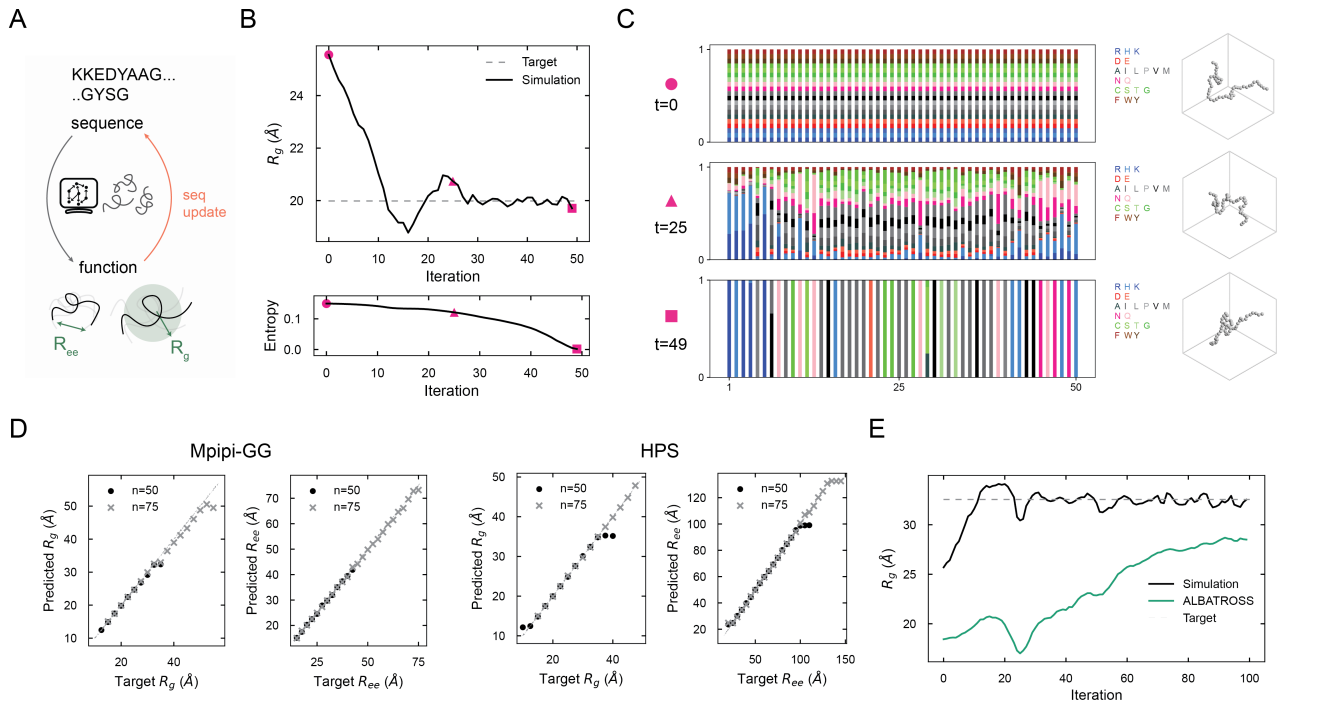

Fig. 2. Designing IDPs with varying ensemble dimensions

A. Employing the framework defined in Figure 1 for design of IDPs with defined ensemble-averaged physical dimensions, specifically the R_g, 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 underpredicts R_g of the final optimized sequence by ~ 4 Å.

mension of $\langle R_g \rangle = 20$ Å. We then update π in the direction of desired $\langle R_g \rangle$ while simultaneously annealing, albeit gradually, towards a discrete sequence (Figure 2C). Our routine converges (over 50 epochs and 2.5 hours on an NVIDIA A100 GPU) to a sequence (Figure 2B, Supplementary Data, SI Note 3) which explores a range of conformations (Figure S2) with an ensemble-averaged R_g of $\sim$ 20.1 Å. Rerunning the optimization with varying random seeds leads to different sequences with similar R_g – highlighting the ability of our approach to identify multiple sequences that exhibit similar ensemble-averaged properties (Figure S2, Supplementary Data).

With this framework, we are able to generate sequences of multiple lengths ($n = 50$, $n = 75$) across a wide span of R_g (Figure 2D). When we change the loss to correspond to a different physical property, the end-to-end radius or R_{ee} – a dimension which provides insights into linker function in multi-domain proteins (9) – we are able to design IDPs across a wide range of R_{ee} (Figure 2D, SI Note 4). 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), when compared against the underlying molecular dynamics simulations for ground truth (Table S1). 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 (Figure 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, architecture engineering, fine tuning, or retraining of existing models. We demonstrate this by designing IDPs of particular ensemble dimensions using the same method but with a different commonly used pair potential (Figure 2D). Together, our method provides a versatile approach to identify IDPs with specified conformation-averaged single-chain properties.

De novo design of loops and linkers

We next asked, can we 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 (32, 33). We reasoned that such sequence variants could potentially represent optimally designed loops ($R_g - R_{ee}/\sqrt{6} >> 0$) or linkers ($R_g - R_{ee}/\sqrt{6} << 0$) (Figure 3A, SI Note 5).

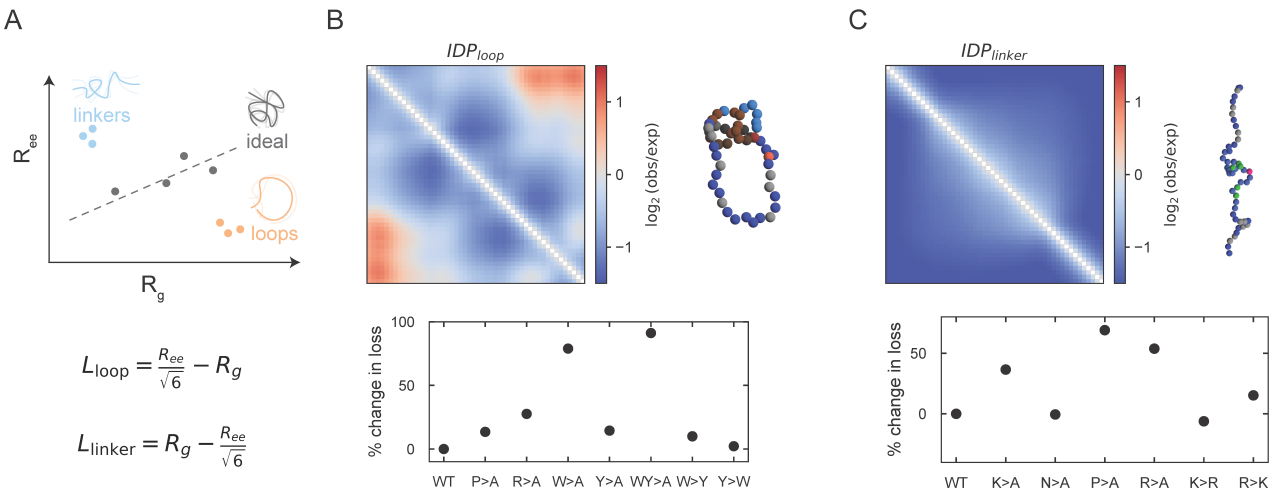

Fig. 3. Shaping global conformational biases through loops and linker IDPs
A. A graphical illustration of ensemble coupling of R_g and R_{ee}, highlighting linear relationships for ideal homopolymer chains ($R_g = R_{ee}/\sqrt{6}$) and decoupled off-diagonal points for loops and linkers. Below, we show 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. To the right of it, a representative configuration that is colored by amino acid identity (similar to Figure 2C), and below it, the relative change in loss value for a set of key mutational scans. 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 upon mutation represents that our solution is highly optimized for the target property.

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 and tyrosines at either termini, interspersed by prolines and arginines that kink out the intervening sequence – highlighted by the normalized contact frequency maps and representative conformations (Figure 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 (Figure 3B, Supplementary Table S2). Similarly, mutational scans of each residue type into alanines or choosing less-complex losses lead to suboptimal loops (Supplementary Table S2) – generically reflecting an inability of simple sequence perturbations to decouple reductions in end-to-end distances from concomitant reductions in chain R_g. Hence, the optimal loop architecture here 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 (Figure 3C). This is largely expected since like-charges have short-range repulsive interactions and simple mutation scans (Figure 3C) are consistent with this intuition. Interestingly, we still identify a variant (R $\rightarrow$ K) that leads to slightly more optimal linkers. 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 (34). First, constraints are enforced through leaky ReLu functions multiplying the target property loss, resulting in gradients that navigate sequence space while maintaining constraints (Figure 4A). Second, instead of directly optimizing over the sequence, we optimize over the weights of a pre-trained and fully connected NN that parametrizes π (Figure 4A). Together, this presents a modular and generalizable strategy to navigate constrained high-dimensional sequence spaces (SI Note 6).

With this framework, we first set out to identify IDPs that are constrained to be highly disordered. We leverage a recent ML-based disorder predictor, Metapredict (35), to measure and constrain disorder (SI Note 6). Importantly, since 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 (Figure 4B). Designing compact proteins i.e., those with small R_g, without any constraints tends to discover highly hydrophobic proteins that are typically predicted to be well-folded and not disordered (see Figure 4C). When we incorporate our disorder constraint, we are able to identify sequences that are simultaneously compact and highly disordered (Figure 4B) across a range of R_g (Figure 4C).

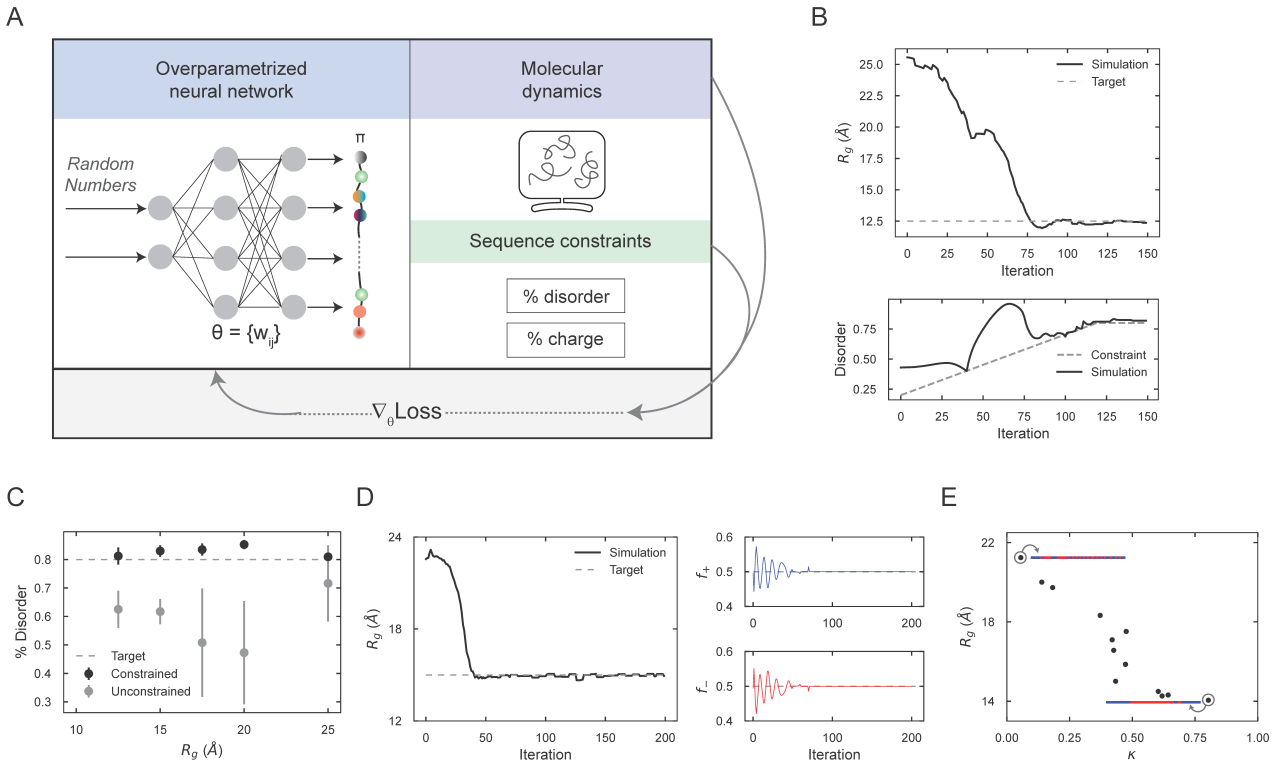

Fig. 4. Engineering IDPs with arbitrary sequence constraints
A. Our framework for applying sequence constraints. Following Ref. (34), we construct a loss function that incorporates arbitrary constraints on the probabilistic sequence and overparameterize the input to the optimization problem i.e., the sequence representation, via a neural network.
B. An example of IDP design ($n = 50$, $Rg = 12.5$ Å) subject to a constraint requiring a minimum degree of sequence disorder as predicted by Metapredict (35). 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 ($n = 50$, 5 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% / 10% of target value for constrained/unconstrained optimizations.
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 panel) and the target Rg (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 (36), recapitulate the inverse relationship shown in (36). The inset depicts sequence patterning represented by blue/red lines for positive/negative residues and shows relatively interspersed/blocky IDPs for high/low R_g values. Each dot represents the most optimized sequence from 5 trajectories and have an R_g within $\sim 10\%$ of the target and charge ratios within $\sim 5\%$ of target.

We next set out to design IDPs with *compositional* constraints. Motivated by previous work (36), 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 (Figure 4D), we pretrain the overparameterized fully-connected NN to output a set of logits corresponding to the target charge distribution, and then use this in our constrained optimization procedure. Consistent with previous predictions, we find an inverse relationship between ensemble dimensions and sequence blockiness (Figure 4E). Together, 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, 37) through changing global or local chain conformations. Thus, we 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 (Figure 5A, SI Note 7). 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 R_g at low salt and low R_g at high salt concentrations. 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, where the salt concentration affects electrostatic screening lengths (SI Note 7). By optimizing for a salt-contractor, we identify a sequence (Figure 5B) rich in arginines with small clusters of interspersed H/Y/W residues. The weakening of repulsive interactions between like-charge R's with salt leads to an effective and modest compaction – as exemplified by a poly-R sequence of identical length (Table S4). In our solution, this passive contraction is am-

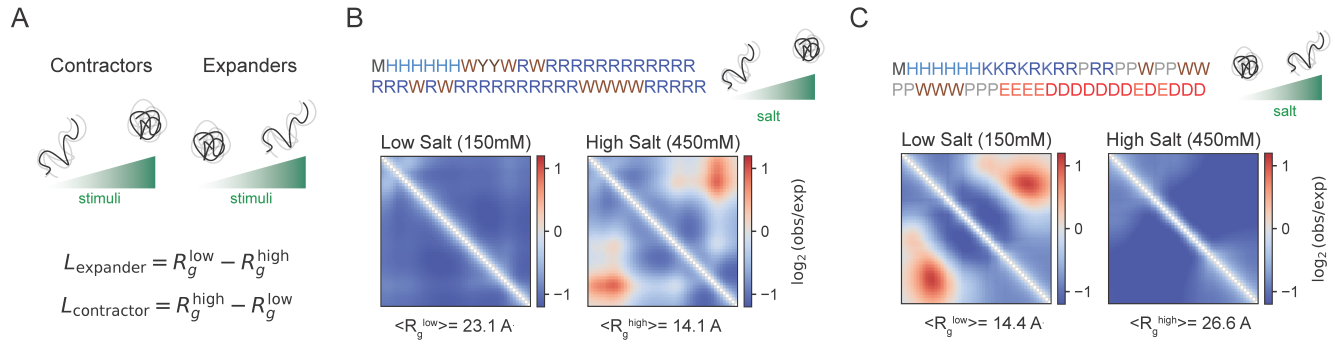

Fig. 5. Programming stimuli-responsive IDP sensors
A. The stimuli-response is depicted by global change in ensemble dimensions across the stimulus (green gradient bar could mean salt, temperature, or phosphorylation) and contractors/expanders represent the direction of change. Below the illustration, 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 and expander sequences with contact frequency maps at low and high salt computed from representative trajectories and 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.

plified by the aromatic clusters whose attractive bonding is salt-insensitive. The periodic spacing and patterning of aromatic solutions in our designed variant only drives compaction under high salt conditions (Figure 5B). All but one mutants that change composition or patterning lead to more compaction but are no longer as salt-sensitive (Table S4, Figure S4A, Supplementary Data).

This ability to exploit complex, many-body heteropolymer physics is even more dramatic in our salt-expander variant (Figure 5C), designed to *increase* R_g with increasing salt concentration. Our algorithm converges to an expander with 3 roughly equal-size sequence modules: positively charged N-terminus, 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 (Table S4, 2.2 Å change). Two linker features, (a) a sticky pi-cation interactions with aromatic residues and the N-terminus and (b) steric effects from proline residues that reduce contact frequency of N/C termini, work in tandem to drive a salt-sensitive "molecular-clasp" (Figure 5C) with a nearly 100% change in R_g. Removal of any key features, e.g. through reducing steric hindrance by P $\rightarrow$ A mutations, or removing sticky residues, leads to a weaker salt-response (Figure S4B). Thus, our method 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 to phosphorylation of serine residues (Figure S5, Supplementary Table S3, SI Note 7). Since the underlying force-fields do not accurately capture temperature dependent variations in hydrophobic interactions, the effect sizes predicted by our model are rather small (Figure S5D-F). Incorporating temperature dependent interactions e.g., in the spirit of Ref. (38), 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 (Figure S5A-C). The mechanisms of contraction/expansion rely on interactions between phosphorylatable residues buried in neighborhoods of positively/negatively charged residues (Figure S5A-C). Thus, the addition of negatively charged groups upon phosphorylation promotes favorable or repulsive interactions, leading to downstream change in IDP ensemble size.

Binders for disordered substrates

The function of many IDPs is driven by binding to disordered substrates, with examples of pico-molar level affinities in highly charged IDPs (30, 39, 40). We next asked, can we 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 (Figure 6A). Precise calculation of binding constants is computationally expensive, often requiring sophisticated enhanced sampling techniques. To overcome this, we make the following simplifications: (a) strong binders are identified by minimizing average interstrand distance and (b) a biasing potential is employed to encourage collection of reference samples that are confined to an effective local volume (SI Note 8). 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 (41, 42) (SI Note 8) are highly fa-

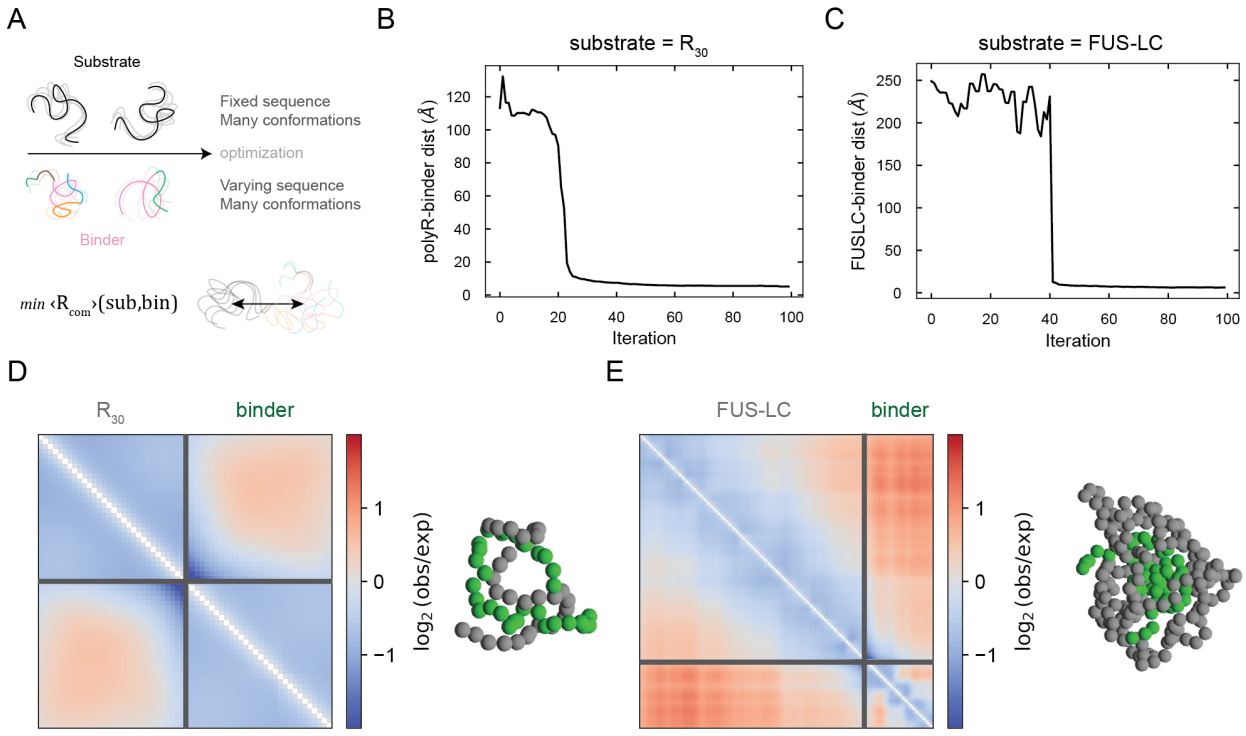

Fig. 6. Designing IDP binders for disordered substrates

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 binder (green) and substrate (grey) are depicted to the right of the panel and lines within box separate out substrate and binder.

vorable for unlike-charge mediated substrate-ligand interactions (Figure S6C) 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 (43, 44) and the poly-Q region of Whi3, an IDP with prominent roles in regulating nuclear autonomy and cell cycle in budding yeast (45). As shown in Figure 6B, our optimization leads to an identification of a target binder (Supplementary Data). While both FUS-LC and Whi3 have strong self-affinity (43, 45), effective interaction coefficients predict stronger interactions between our optimized binders and their respective substrates (Figure S6C) over the homotypic substrate-substrate interactions. In unbiased forward simulations, we observe strongly enriched intermolecular interactions for all binder-substrate pairs (Figures 6D-E, S6B), indicating strong binding at the μM concentrations we studied. Across all the optimizations, we find that a sharp change in the learning dynamics (Figures 6B-C, S6A) is concomitant with strong binder identification. We expect that future studies will dissect whether this transition represents features of the underlying learning protocol i.e., 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 (IDPs) 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 (Figure 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 and R_{ee} (Figure 2), and with tailored conformational biases i.e., loops and linkers (Figure 3), properties that have been shown to shape cellular function (2, 30). We next develop a modular strat-

egy 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 (Figure 4). With this framework, we next design highly sensitive sensors to multiple physicochemical and cellular stimuli such as salt, temperature, and phosphorylation (Figure 5). These designed sensors, in turn, shed light on the physical mechanisms by which the balance of competing intramolecular interactions encodes for large conformational changes. Finally, we use this method to identify disordered binders for low-complexity substrates (Figure 6). More generally, there is significant 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 (46), field-theoretical approaches (47), and thermodynamics-informed models (41, 42, 48).

The framework we propose directly inverts simulation-derived sequence-ensemble relationships to drive *de novo* IDP design with tailored single-chain, binding, and environmental-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 (34, 49, 50). 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 (51).

Incorporation of emerging machine-learning approaches – for e.g., simulation-free generative methods to generate conformational ensembles (52, 53), combining alchemical and molecular dynamics simulations for sequence variant design with target single-chain properties (5, 54), and approximate ML models that can rapidly invert pre-trained sequence-single-chain property relationships (23, 54), 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, since we compute gradient estimates via a reweighting scheme that relies on knowledge of unnormalized probabilities, our framework in its present form does not naturally accommodate far-from-equilibrium properties for which state-level probabilities are generally not known. Opportunities to address this limitation in future work include exploiting classic results in non-equilibrium statistical mechanics (e.g. Jarczynski 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. Finally, directly inverting molecular simulations has a direct tradeoff contrasting increased accuracy with additional speed and compute requirements, thus making it less appealing for design of properties for which machine-learned approximations are comparable in accuracy.

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 $\overrightarrow{s} \in \mathbb{R}^n$ denote the identities of each particle where $\overrightarrow{s}_i \in \{1, 2, \cdots, m\}$. Given a potential energy function $U : \mathbb{R}^{n\times d} \to \mathbb{R}$ that depends on the particle identities, $\overrightarrow{s}$ determines the distribution of states in the canonical ensemble via $p(\overrightarrow{x}; \overrightarrow{s}) \sim \exp(-\beta U(\overrightarrow{x}; \overrightarrow{s}))$ where β is the inverse temperature and $\overrightarrow{x} \in \mathbb{R}^{n\times d}$. 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, $\mathbb{E}[O(\overrightarrow{x})]_{\overrightarrow{x}\sim p(\cdot;\overrightarrow{s})}$. Consequently, we consider the optimization problem

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

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 (i) is general and makes minimal assumptions about the underlying model, (ii) operates directly at the level of the model and requires no training, (iii) yields an optimized probability distribution of identities from which discrete identities can be sampled, and (iv) 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 i^{th} 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 $\overrightarrow{x}$ as

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

where

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

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

$$\begin{aligned} p(\overrightarrow{x}, \pi) &\sim \exp(-\beta \mathbb{E}[U(\overrightarrow{x}, \pi)]) && (4)\\ &= \exp(-\beta \sum_{\overrightarrow{s}\in S} p(\overrightarrow{s}|\pi) U(\overrightarrow{x}; \overrightarrow{s})) && (5)\\ &= \prod_{\overrightarrow{s}\in S} \exp\left[-\beta \left(p(\overrightarrow{s}|\pi) U(\overrightarrow{x}; \overrightarrow{s})\right)\right] && (6)\\ &= \prod_{\overrightarrow{s}\in S} \left(\exp\left[-\beta U(\overrightarrow{x}; \overrightarrow{s})\right]\right)^{p(\overrightarrow{s}|\pi)} && (7)\\ &\sim \prod_{\overrightarrow{s}\in S} p(\overrightarrow{x}; \overrightarrow{s})^{p(\overrightarrow{s}|\pi)} && (8) \end{aligned}$$

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

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

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 (e.g. a Langevin integrator), one can compute $\nabla_\pi \mathbb{E}[O(\overrightarrow{x})]_{\overrightarrow{x}\sim p(\cdot;\pi)}$ via Differentiable Trajectory Reweighting (DiffTRE) (28). Consider a set of states $\{\overrightarrow{x}_1, \overrightarrow{x}_2, \cdots, \overrightarrow{x}_T\}$ sampled from the Boltzmann distribution defined by Equation 5 for a reference state matrix $\hat{\pi}$. For values of π sufficiently close to $\hat{\pi}$ (see Methods), we define a weight

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

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

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

This yields an expression for $\mathbb{E}[O(\overrightarrow{x})]$ such that $\nabla_\pi \mathbb{E}[O(\overrightarrow{x})] \neq 0$. 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 MD) but only through the energy function, relieving many of the numerical instabilities and memory constraints that typically plague differentiable MD. 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. Additionally, the set of reference states must not be computed at every iteration (see Methods), relaxing the computational cost imposed by running large simulations.

In practice, since 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, i.e. $\pi_i = \text{softmax}(\lambda_i)$. Since 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, i.e. $\pi_i = \text{softmax}(\lambda_i/\tau)$. We find that a simple linear annealing scheme using $\tau_{\text{start}} = 1.0$ and $\tau_{\text{end}} = 0.01$ 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 $\overrightarrow{s}$:

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

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

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

where

$$\mathbb{E}[U_{pair}(\overrightarrow{x}_i,\overrightarrow{x}_j;\pi)] = \sum_{\overrightarrow{s}_i,\overrightarrow{s}_j\in m}\pi(i,\overrightarrow{s}_i)\pi(j,\overrightarrow{s}_j) \times U_{pair}(\overrightarrow{x}_i,\overrightarrow{x}_j;\overrightarrow{s}_i,\overrightarrow{s}_j) \quad \textbf{(14)}$$

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

$$\mathbb{E}[U_{pair}(\overrightarrow{x}_i,\overrightarrow{x}_j;\pi)] = \overrightarrow{U_{ij}}\cdot\overrightarrow{\pi_{ij}} \quad \textbf{(15)}$$

where

$$(\overrightarrow{U_{ij}})_{kl} = U_{pair}(\overrightarrow{x}_i,\overrightarrow{x}_j;k,l) \quad \textbf{(16)}$$

and

$$\overrightarrow{\pi_{ij}} = \pi_i\otimes\pi_j \quad \textbf{(17)}$$

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. While 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 $\mathcal{O}(n^km^k)$ (prior to any neighbor list optimizations).

Differentiable Monte Carlo (DiffTRE)

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 the following, we focus on the simple case of an equilibrium system in the canonical ensemble where the probability of state $\overrightarrow{x}_i$ is $\frac{e^{-\beta U(\overrightarrow{x}_i)}}{Z}$ where β is the inverse temperature, $U(\overrightarrow{x}_i)$ is the potential energy of $\overrightarrow{x}_i$, and $Z=\sum_j e^{-\beta U(\overrightarrow{x}_j)}$ is the partition function.

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

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

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

$$\langle O(\overrightarrow{x},\theta)\rangle_{\overrightarrow{x}_i\in X} = \sum_i w_i O(\overrightarrow{x}_i,\theta) \quad \textbf{(19)}$$

where

$$w_i = \frac{p_\theta(\overrightarrow{x}_i)/p_{\hat{\theta}}(\overrightarrow{x}_i)}{\sum_j p_\theta(\overrightarrow{x}_j)/p_{\hat{\theta}}(\overrightarrow{x}_j)} \quad \textbf{(20)}$$

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:

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

Crucially, in the case where $\theta=\hat{\theta}$, $w_i=\frac{1}{N}$ but $\nabla_\theta\log(p(\overrightarrow{x}_i))\neq 0$. 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:

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

See Refs. (29) and (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 (e.g. simulation code) does not have to be rewritten in an automatic differentiation framework. Instead, one only must write the energy function in such a framework. In addition, 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 MD 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 (LLPS) (21). Introduced in 2021, Mpipi has gained widespread popularity for the computational study of LLPS and the underlying biophysics (55–59). 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, 60). 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 monomer (i.e. amino acid or nucleic acid) is represented a single isotropic sphere. Each monomer type (i.e. amino acid or nucleotide identity) is assigned a mass, diameter, charge, and energy scale. Like oxDNA, all interactions are pairwise and the potential energy is given by

$$V_{\text{mpipi}} = \sum_{\text{nn}} V_{\text{bond}} + \sum_{\text{other pairs}} (V_{\text{elec}} + V_{\text{pair}}) \qquad \textbf{(23)}$$

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 (61). The parameters of this potential were fit to reproduce both atomistic potential-of-mean-force calculations and bioinformatics data. See (21) for complete details of the model and its parameterization, and see (23) for a description of the modified parameters used in this work.

Simulations

All simulations with 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 300K and 150 mM salt concentration unless specified otherwise. We make all code available via the following GitHub repository: https://github.com/rkruegs123/idp-design.

Acknowledgments

We thank Max Ward for his collaboration on sequence design via overparameterization in the context of differentiable RNA folding, which inspired this work, Jamie Smith for helpful discussions relating to stochastic gradient estimators, and Wilton Snead for helpful discussions on IDP design. R.K.K., M.P.B., and K.S. acknowledge support from the Simons Foundation through the Simons Foundation Investigator award. R.K.K and M.P.B. acknowledge support from the NSF AI Institute of Dynamic Systems (#2112085), Office of Naval Research (N00014-17-1-3029), and the Harvard Materials Research Science and Engineering Center (DMR 20-11754). K.S. acknowledges support from NSF–Simons Center for Mathematical and Statistical Analysis of Biology at Harvard (Award #1764269) and Northwestern University for startup funding.

Bibliography

1. David D. Boehr, Ruth Nussinov, and Peter E. Wright. The role of dynamic conformational ensembles in biomolecular recognition. *Nature Chemical Biology*, 5(11):789–796, November 2009. ISSN 1552-4469. doi: 10.1038/nchembio.232.
2. Alex S. Holehouse and Birthe B. Kragelund. The molecular basis for cellular function of intrinsically disordered protein regions. *Nature Reviews Molecular Cell Biology*, 25(3):187–211, March 2024. ISSN 1471-0080. doi: 10.1038/s41580-023-00673-0.
3. Robin van der Lee, Marija Buljan, Benjamin Lang, Robert J. Weatheritt, Gary W. Daughdrill, A. Keith Dunker, Monika Fuxreiter, Julian Gough, Joerg Gsponer, David T. Jones, Philip M. Kim, Richard W. Kriwacki, Christopher J. Oldfield, Rohit V. Pappu, Peter Tompa, Vladimir N. Uversky, Peter E. Wright, and M. Madan Babu. Classification of Intrinsically Disordered Regions and Proteins. *Chemical Reviews*, 114(13):6589–6631, July 2014. ISSN 0009-2665. doi: 10.1021/cr400525m.
4. Vladimir N Uversky. 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. *Annual Review of Biophysics*, 50(1):135–156, 2021.
5. Giulio Tesei, Anna Ida Trolle, Nicolas Jonsson, Johannes Betz, Frederik E. Knudsen, Francesco Pesce, Kristoffer E. Johansson, and Kresten Lindorff-Larsen. Conformational ensembles of the human intrinsically disordered proteome. *Nature*, 626(8000):897–904, February 2024. ISSN 1476-4687. doi: 10.1038/s41586-023-07004-5.
6. F. Emil Thomasen and Kresten Lindorff-Larsen. Conformational ensembles of intrinsically disordered proteins and flexible multidomain proteins. *Biochemical Society Transactions*, 50(1):541–554, February 2022. ISSN 0300-5127. doi: 10.1042/BST20210499.
7. Tanja Mittag and Julie D Forman-Kay. Atomic-level characterization of disordered protein ensembles. *Current opinion in structural biology*, 17(1):3–14, 2007.
8. Norman E. Davey, Leandro Simonetti, and Ylva Ivarsson. The next wave of interactomics: Mapping the SLiM-based interactions of the intrinsically disordered proteome. *Current Opinion in Structural Biology*, 80:102593, June 2023. ISSN 0959-440X. doi: 10.1016/j.sbi.2023.102593.
9. Qiaojing Huang, Maodong Li, Luhua Lai, and Zhirong Liu. Allostery of multidomain proteins with disordered linkers. *Current Opinion in Structural Biology*, 62:175–182, June 2020. ISSN 0959-440X. doi: 10.1016/j.sbi.2020.01.017.
10. David Moses, Garrett M. Ginell, Alex S. Holehouse, and Shahar Sukenik. Intrinsically disordered regions are poised to act as sensors of cellular chemistry. *Trends in Biochemical Sciences*, 0(0), August 2023. ISSN 0968-0004. doi: 10.1016/j.tibs.2023.08.001.
11. Salman F. Banani, Lena K. Afeyan, Susana W. Hawken, Jonathan E. Henninger, Alessandra Dall'Agnese, Victoria E. Clark, Jesse M. Platt, Ozgur Oksuz, Nancy M. Hannett, Ido Sagi, Tong Ihn Lee, and Richard A. Young. Genetic variation associated with condensate dysregulation in disease. *Developmental Cell*, 57(14):1776–1788.e8, July 2022. ISSN 1534-5807. doi: 10.1016/j.devcel.2022.06.010.
12. Krishna Shrinivas, Benjamin R Sabari, Eliot L Coffey, Isaac A Klein, Ann Boija, Alicia V Zamudio, Jurian Schuijers, Nancy M Hannett, Phillip A Sharp, Richard A Young, et al. Enhancer features that drive formation of transcriptional condensates. *Molecular cell*, 75(3):549–561, 2019.
13. Benjamin R Sabari. Biomolecular condensates and gene activation in development and disease. *Developmental cell*, 55(1):84–96, 2020.
14. Ming Shi, Pengfei Zhang, Setu M Vora, and Hao Wu. Higher-order assemblies in innate immune and inflammatory signaling: A general principle in cell biology. *Current opinion in cell biology*, 63:194–203, 2020.
15. Brian Tsang, Iva Pritišanac, Stephen W. Scherer, Alan M. Moses, and Julie D. Forman-Kay. Phase Separation as a Missing Mechanism for Interpretation of Disease Mutations. *Cell*, 183(7):1742–1756, December 2020. ISSN 0092-8674. doi: 10.1016/j.cell.2020.11.050.
16. J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker. Robust deep learning–based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, October 2022. doi: 10.1126/science.add2187.
17. Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. ColabFold: Making protein folding accessible to all. *Nature Methods*, 19(6):679–682, June 2022. ISSN 1548-7105. doi: 10.1038/s41592-022-01488-1.
18. John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.
19. Rohith Krishna, Jue Wang, Woody Ahern, Pascal Sturmfels, Preetham Venkatesh, Indrek Kalvet, Gyu Rie Lee, Felix S. Morey-Burrows, Ivan Anishchenko, Ian R. Humphreys, Ryan McHugh, Dionne Vafeados, Xinting Li, George A. Sutherland, Andrew Hitchcock, C. Neil Hunter, Alex Kang, Evans Brackenbrough, Asim K. Bera, Minkyung Baek, Frank DiMaio, and David Baker. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science*, 384(6693):eadl2528, March 2024. doi: 10.1126/science.adl2528.
20. Gregory L. Dignon, Wenwei Zheng, Robert B. Best, Young C. Kim, and Jeetain Mittal. Relation between single-molecule properties and phase behavior of intrinsically disordered proteins. *Proceedings of the National Academy of Sciences*, 115(40):9929–9934, October 2018. doi: 10.1073/pnas.1804177115.
21. Jerelle A. Joseph, Aleks Reinhardt, Anne Aguirre, Pin Yu Chew, Kieran O. Russell, Jorge R. Espinosa, Adiran Garaizar, and Rosana Collepardo-Guevara. Physics-driven coarse-grained model for biomolecular phase separation with near-quantitative accuracy. *Nature Computational Science*, 1(11):732–743, November 2021. ISSN 2662-8457. doi: 10.1038/s43588-021-00155-3.
22. Giulio Tesei, Thea K. Schulze, Ramon Crehuet, and Kresten Lindorff-Larsen. Accurate model of liquid–liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties. *Proceedings of the National Academy of Sciences*, 118(44):e2111696118, November 2021. doi: 10.1073/pnas.2111696118.
23. Jeffrey M. Lotthammer, Garrett M. Ginell, Daniel Griffith, Ryan J. Emenecker, and Alex S. Holehouse. Direct prediction of intrinsically disordered protein conformational properties from sequence. *Nature Methods*, 21(3):465–476, March 2024. ISSN 1548-7105. doi: 10.1038/s41592-023-02159-5.
24. Ryan J. Emenecker, Karina Guadalupe, Nora M. Shamoon, Shahar Sukenik, and Alex S. Holehouse. Sequence-ensemble-function relationships for disordered proteins in live cells. *bioRxiv*, page 2023.10.29.564547, November 2023. doi: 10.1101/2023.10.29.564547.
25. Roshan Mammen Regy, Jacob Thompson, Young C Kim, and Jeetain Mittal. Improved coarse-grained model for studying sequence dependent phase separation of disordered proteins. *Protein Science*, 30(7):1371–1379, 2021.
26. James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, et al. Jax: composable transformations of python+ numpy programs. 2018.
27. Samuel Schoenholz and Ekin Dogus Cubuk. Jax md: A framework for differentiable physics. *Advances in Neural Information Processing Systems*, 33: 11428–11441, 2020.
28. Stephan Thaler and Julija Zavadlav. Learning neural network potentials from experimental data via Differentiable Trajectory Reweighting. *Nature communications*, 12(1):6884, 2021.
29. Shi-Xin Zhang, Zhou-Quan Wan, and Hong Yao. Automatic differentiable Monte Carlo: Theory and application. *Physical Review Research*, 5(3): 033041, July 2023. doi: 10.1103/PhysRevResearch.5.033041.
30. Nicolás S. González-Foutel, Juliana Glavina, Wade M. Borcherds, Matías Safranchik, Susana Barrera-Vilarmau, Amin Sagar, Alejandro Estaña, Amelie Barozet, Nicolás A. Garrone, Gregorio Fernandez-Ballester, Clara Blanes-Mira, Ignacio E. Sánchez, Gonzalo de Prat-Gay, Juan Cortés, Pau Bernadó, Rohit V. Pappu, Alex S. Holehouse, Gary W. Daughdrill, and Lucía B. Chemes. Conformational buffering underlies functional selection in intrinsically disordered protein regions. *Nature Structural & Molecular Biology*, 29(8):781–790, August 2022. ISSN 1545-9985. doi: 10.1038/s41594-022-00811-w.
31. Yi-Hsuan Lin and Hue Sun Chan. Phase Separation and Single-Chain Compactness of Charged Disordered Proteins Are Strongly Correlated. *Biophysical Journal*, 112(10):2043–2046, May 2017. ISSN 0006-3495. doi: 10.1016/j.bpj.2017.04.021.
32. Mauro L. Mugnai, Debayan Chakraborty, Abhinaw Kumar, Hung T. Nguyen, Wade Zeno, Jeanne C. Stachowiak, John E. Straub, and D. Thirumalai. Sizes, conformational fluctuations, and SAXS profiles for Intrinsically Disordered Proteins. page 2023.04.24.538147, July 2024. doi: 10.1101/2023.04.24.538147.
33. Joshua A. Riback, Micayla A. Bowman, Adam M. Zmyslowski, Kevin W. Plaxco, Patricia L. Clark, and Tobin R. Sosnick. Commonly used FRET fluorophores promote collapse of an otherwise disordered protein. *Proceedings of the National Academy of Sciences*, 116(18):8889–8894, April 2019. doi: 10.1073/pnas.1813038116.
34. Ryan Krueger and Max Ward. Scalable Differentiable Folding for mRNA Design. page 2024.05.29.594436, June 2024. doi: 10.1101/2024.05.29.594436.
35. Ryan J Emenecker, Daniel Griffith, and Alex S Holehouse. Metapredict V2: An update to metapredict, a fast, accurate, and easy-to-use predictor of consensus disorder and structure. *BioRxiv*, pages 2022–06, 2022.
36. Rahul K Das and Rohit V Pappu. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged

residues. *Proceedings of the National Academy of Sciences*, 110(33):13392–13397, 2013.

37. Mao Zhang, Cheng Zhu, Yuanyuan Duan, Tongbao Liu, Haoping Liu, Chang Su, and Yang Lu. The intrinsically disordered region from pp2c phosphatases functions as a conserved co2 sensor. *Nature cell biology*, 24(7):1029–1037, 2022.
38. Gregory L. Dignon, Wenwei Zheng, Young C. Kim, and Jeetain Mittal. Temperature-Controlled Liquid–Liquid Phase Separation of Disordered Proteins. *ACS Central Science*, 5(5):821–830, May 2019. ISSN 2374-7943. doi: 10.1021/acscentsci.9b00102.
39. Alessandro Borgia, Madeleine B. Borgia, Katrine Bugge, Vera M. Kissling, Pétur O. Heidarsson, Catarina B. Fernandes, Andrea Sottini, Andrea Soranno, Karin J. Buholzer, Daniel Nettels, Birthe B. Kragelund, Robert B. Best, and Benjamin Schuler. Extreme disorder in an ultrahigh-affinity protein complex. *Nature*, 555(7694):61–66, March 2018. ISSN 1476-4687. doi: 10.1038/nature25762.
40. Benjamin Schuler, Alessandro Borgia, Madeleine B Borgia, Pétur O Heidarsson, Erik D Holmstrom, Daniel Nettels, and Andrea Sottini. Binding without folding – the biomolecular function of disordered polyelectrolyte complexes. *Current Opinion in Structural Biology*, 60:66–76, February 2020. ISSN 0959-440X. doi: 10.1016/j.sbi.2019.12.006.
41. Kyosuke Adachi and Kyogo Kawaguchi. Predicting Heteropolymer Interactions: Demixing and Hypermixing of Disordered Protein Sequences. *Physical Review X*, 14(3):031011, July 2024. doi: 10.1103/PhysRevX.14.031011.
42. Garrett M. Ginell, Ryan J. Emenecker, Jeffrey M. Lotthammer, Emery T. Usher, and Alex S. Holehouse. Direct prediction of intermolecular interactions driven by disordered regions. page 2024.06.03.597104, June 2024. doi: 10.1101/2024.06.03.597104.
43. Bede Portz, Bo Lim Lee, and James Shorter. FUS and TDP-43 Phases in Health and Disease. *Trends in Biochemical Sciences*, 46(7):550–563, July 2021. ISSN 0968-0004. doi: 10.1016/j.tibs.2020.12.005.
44. Jie Wang, Jeong-Mo Choi, Alex S. Holehouse, Hyun O. Lee, Xiaojie Zhang, Marcus Jahnel, Shovamayee Maharana, Régis Lemaitre, Andrei Pozniakovsky, David Drechsel, Ina Poser, Rohit V. Pappu, Simon Alberti, and Anthony A. Hyman. A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins. *Cell*, 174(3):688–699.e16, July 2018. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2018.06.006.
45. Christine Roden and Amy S Gladfelter. Rna contributions to the form and function of biomolecular condensates. *Nature Reviews Molecular Cell Biology*, 22(3):183–195, 2021.
46. Jeong-Mo Choi and Rohit V. Pappu. Improvements to the ABSINTH Force Field for Proteins Based on Experimentally Derived Amino Acid Specific Backbone Conformational Statistics. *Journal of Chemical Theory and Computation*, 15(2):1367–1382, February 2019. ISSN 1549-9618. doi: 10.1021/acs.jctc.8b00573.
47. Jonas Wessén, Suman Das, Tanmoy Pal, and Hue Sun Chan. Analytical Formulation and Field-Theoretic Simulation of Sequence-Specific Phase Separation of Protein-Like Heteropolymers with Short- and Long-Spatial-Range Interactions. *The Journal of Physical Chemistry B*, 126(45):9222–9245, November 2022. ISSN 1520-6106. doi: 10.1021/acs.jpcb.2c06181.
48. Krishna Shrinivas and Michael P. Brenner. Phase separation in fluids with many interacting components. *Proceedings of the National Academy of Sciences*, 118(45), November 2021. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2108551118.
49. Christopher Frank, Ali Khoshouei, Yosta de Stigter, Dominik Schiewitz, Shihao Feng, Sergey Ovchinnikov, and Hendrik Dietz. Efficient and scalable de novo protein design using a relaxed sequence space. *bioRxiv*, page 2023.02.24.529906, February 2023. doi: 10.1101/2023.02.24.529906.
50. Marco C Matthies, Ryan Krueger, Andrew E Torda, and Max Ward. Differentiable partition function calculation for RNA. *Nucleic Acids Research*, 52(3):e14, February 2024. ISSN 0305-1048. doi: 10.1093/nar/gkad1168.
51. Ryan Krueger, Megan C Engel, Ryan Hausen, and Michael P Brenner. A differentiable model of nucleic acid dynamics. *in prep*, 2024.
52. Giacomo Janson and Michael Feig. Transferable deep generative modeling of intrinsically disordered protein conformations. *PLOS Computational Biology*, 20(5):e1012144, May 2024. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1012144.
53. Caixuan Liu, Kejia Wu, Hojun Choi, Hannah Han, Xulie Zhang, Joseph L. Watson, Sara Shijo, Asim K. Bera, Alex Kang, Evans Brackenbrough, Brian Coventry, Derrick R. Hick, Andrew N. Hoofnagle, Ping Zhu, Xingting Li, Justin Decarreau, Stacey R. Gerben, Wei Yang, Xinru Wang, Mila Lamp, Analisa Murray, Magnus Bauer, and David Baker. Diffusing protein binders to intrinsically disordered proteins. page 2024.07.16.603789, July 2024. doi: 10.1101/2024.07.16.603789.
54. Francesco Pesce, Anne Bremer, Giulio Tesei, Jesse B. Hopkins, Christy R. Grace, Tanja Mittag, and Kresten Lindorff-Larsen. Design of intrinsically disordered protein variants with diverse structural properties. *Science Advances*, 10(35):eadm9926, August 2024. doi: 10.1126/sciadv.adm9926.
55. Ignacio Sanchez-Burgos, Jorge R Espinosa, Jerelle A Joseph, and Rosana Collepardo-Guevara. RNA length has a non-trivial effect in the stability of biomolecular condensates formed by RNA-binding proteins. *PLoS computational biology*, 18(2):e1009810, 2022.
56. Hongjia Zhu, Masako Narita, Jerelle A Joseph, Georg Krainer, William E Arter, Ioana Olan, Kadi L Saar, Niklas Ermann, Jorge R Espinosa, Yi Shen, et al. The chromatin regulator HMGA1a undergoes phase separation in the nucleus. *ChemBioChem*, 24(1):e202200450, 2023.
57. Jhullian Alston, Andrea Soranno, and Alex S Holehouse. Conserved molecular recognition by an intrinsically disordered region in the absence of sequence conservation. *bioRxiv*, pages 2023–08, 2023.
58. Jonas Wessén, Suman Das, Tanmoy Pal, and Hue Sun Chan. Analytical formulation and field-Theoretic simulation of sequence-specific phase separation of protein-like Heteropolymers with Short-and long-spatial-range interactions. *The Journal of Physical Chemistry B*, 126(45):9222–9245, 2022.
59. Adiran Garaizar, Jorge R Espinosa, Jerelle A Joseph, Georg Krainer, Yi Shen, Tuomas PJ Knowles, and Rosana Collepardo-Guevara. Aging can transform single-component protein condensates into multiphase architectures. *Proceedings of the National Academy of Sciences*, 119(26):e2119800119, 2022.
60. Ishan Taneja and Keren Lasker. Machine learning based methods to generate conformational ensembles of disordered proteins. *Biophysical Journal*, 2023.
61. Xipeng Wang, Simón Ramírez-Hinestrosa, Jure Dobnikar, and Daan Frenkel. The Lennard-Jones potential: When (not) to use it. *Physical Chemistry Chemical Physics*, 22(19):10624–10633, 2020.
62. Shakir Mohamed, Mihaela Rosca, Michael Figurnov, and Andriy Mnih. Monte carlo gradient estimation in machine learning. *The Journal of Machine Learning Research*, 21(1):5183–5244, 2020.
63. Megan C Engel, Jamie A Smith, and Michael P Brenner. Optimal control of nonequilibrium systems through automatic differentiation. *Physical Review X*, 13(4):041032, 2023.
64. Luke Metz, C Daniel Freeman, Samuel S Schoenholz, and Tal Kachman. Gradients are not all you need. *arXiv preprint arXiv:2111.05803*, 2021.