Active learning of the thermodynamics–dynamics tradeoff in protein condensates

Yaxin An,[1,2,*] Michael A. Webb,[1,†] and William M. Jacobs[2,‡]

[1] *Department of Chemical and Biological Engineering, Princeton University, Princeton, NJ 08544 USA*

[2] *Department of Chemistry, Princeton University, Princeton, NJ 08544 USA*

(Dated: June 6, 2023)

Phase-separated biomolecular condensates exhibit a wide range of dynamical properties, which depend on the sequences of the constituent proteins and RNAs. However, it is unclear to what extent condensate dynamics can be tuned without also changing the thermodynamic properties that govern phase separation. Using coarse-grained simulations of intrinsically disordered proteins, we show that the dynamics and thermodynamics of homopolymer condensates are strongly correlated, with increased condensate stability being coincident with low mobilities and high viscosities. We then apply an "active learning" strategy to identify heteropolymer sequences that break this correlation. This data-driven approach and accompanying analysis reveal how heterogeneous amino-acid compositions and non-uniform sequence patterning map to a range of independently tunable dynamical and thermodynamic properties of biomolecular condensates.

I. INTRODUCTION

Biomolecular condensates—also known as "membraneless organelles"—are self-organized structures within the cytoplasm and nucleoplasm of living cells [1–3]. Condensates play diverse roles in a wide variety of biological processes [4–6], in large part because of their ability to concentrate proteins, RNAs, and small molecules in chemically specific environments [5–8]. Nonetheless, the ability of condensates to tune the rates of biochemical reactions is not only determined by their biochemical compositions. Dynamical properties, such as viscosities and molecular mobilities, also affect how condensates exchange biomolecules with the rest of the cell [9–11] and interact with other intracellular structures, such as chromatin [12] and cytoskeletal components [13]. Dynamical properties are also directly relevant to condensate (dis)assembly kinetics [14] and irreversible aging processes [15]. Recent experiments have demonstrated that dynamical properties vary widely across different condensates, both *in vitro* and *in vivo* [11, 16].

Many condensates are believed to form spontaneously as a result of phase separation taking place close to a local thermodynamic equilibrium [1, 17]. The thermodynamic stability of a condensate, which can be quantified by the critical temperature of a biopolymer solution [18] or by the saturation concentration required for phase separation [9, 19], is an equilibrium property related to the multiplicity and strengths of the "multivalent" interactions among the biomolecules comprising a condensate [9, 20, 21]. Intuitively, one might expect that the thermodynamic stability of a condensate is correlated with its dynamical properties, which also emerge from the collective effects of biomolecular interactions in the condensed phase. This anticipated relationship represents a "tradeoff" between stability and dynamics, since stronger thermodynamic driving forces are expected to correlate with lower molecular mobilities [22]. Experimental evidence suggests that such a correlation exists for certain ribonucleic condensates, such as Arg/Gly-rich repeat protein/RNA condensates [11] and LAF-1/RNA condensates [16]. However, understanding the general relationship between thermodynamic stability and the internal dynamics of biomolecular condensates requires additional and systematic exploration [23, 24].

Determining whether a stability–dynamics tradeoff extends to phase-separating biopolymers is challenging due to the enormous diversity of biopolymers that can form condensates. Here, we focus on condensates composed of intrinsically disordered proteins (IDPs), a class of proteins that tend not to adopt stable secondary and tertiary structures and are known to be essential components of many naturally occurring condensates [25]. To study the properties of IDP sequences, we employ a coarse-grained (CG) model that represents IDPs at residue-level resolution [26]. Models of this type have been shown to reproduce experimental measurements of both single-molecule and condensed-phase properties [18, 27–29], suggesting that CG IDP models capture essential physics of disordered polypeptides and retain sufficient chemical specificity to predict differences among condensates composed of various IDP sequences. Thus, CG IDP models seem well-suited to systematically explore the relationship between IDP stability and condensate dynamics.

In this Article, we combine molecular dynamics simulations and machine learning to navigate the sequence space [30, 31] of IDPs and elucidate the relationship between the phase behavior and internal dynamics of single-component IDP condensates. We first establish a strong correlation between proxies for condensate stability and dynamics in homomeric polypeptides, which is understood using simple physical arguments. We then deploy a computationally efficient machine-learning strategy based on "active learning"[32–34] to identify IDP

* Present address: Department of Chemical Engineering, Louisiana State University, Baton Rouge, LA 70803 USA

† mawebb@princeton.edu

‡ wjacobs@princeton.edu

sequences that break this correlation, exhibiting faster internal dynamics than homomeric polypeptides that assemble into equally stable condensates. In this way, we identify "Pareto-optimal" IDP sequences, meaning that sequence perturbations cannot enhance the dynamics further without reducing the stability of the condensate. Finally, we examine sequence features of Pareto-optimal sequences and perform a counterfactual analysis[35, 36] to identify the sequence determinants of the limiting thermodynamics–dynamics tradeoff. Taken together, our results demonstrate how sequence design can be used to tune thermodynamic and dynamical properties independently in the context of biomolecular condensates.

II. RESULTS

A. Thermodynamic and dynamic properties of homomeric polypeptides are strongly correlated

To establish a baseline expectation for the thermodynamics–dynamics tradeoff in IDP condensates, we begin by studying a collection of homopolymeric sequences. Specifically, we consider homopolymers consisting of each amino-acid type with chain lengths varying between $N = 20$ and 50 monomers. Molecular dynamics simulations are performed using the CG model developed by Mittal and coworkers [26, 28], which treats IDPs at the single-residue level in implicit solvent. Amino-acid residues interact via a combination of bonded and nonbonded pair potentials, the latter of which account for both electrostatic interactions and short-ranged hydrophobic forces. The details of this model are provided in Sec. IV A.

For each homopolymer, we compute the second-virial coefficient, B_2, which serves as a proxy for the thermodynamic stability of the condensed phase, and the self-diffusion coefficient in the condensed phase, D, which serves as a measure of the internal condensate dynamics. B_2 quantifies the net attractive or net repulsive intermolecular interactions of a pair of molecules in dilute solution [37]. In both colloidal [38] and homopolymer solutions [39], B_2 is strongly correlated with the critical temperature for phase separation, with a negative B_2 value required for phase separation to occur. We therefore take B_2 as a simple proxy for the thermodynamic stability of a phase-separated condensate throughout this work. In practice, we compute B_2 by calculating the potential of mean force, $u(r)$, between the centers of mass of two polymer chains at 300 K using adaptive biasing force simulations [40] (Fig. 1a; see Sec. IV B). B_2 is then obtained from the equation

$$B_2 = 2\pi \int_0^\infty dr\, r^2 \left[1 - e^{-\beta u(r)}\right], \qquad (1)$$

where r represents the distance between the center of mass of each chain and $\beta \equiv 1/k_\mathrm{B}T$. For convenience, we report B_2 relative to a reference volume, $V_0 = 5529\,\text{Å}^3$,

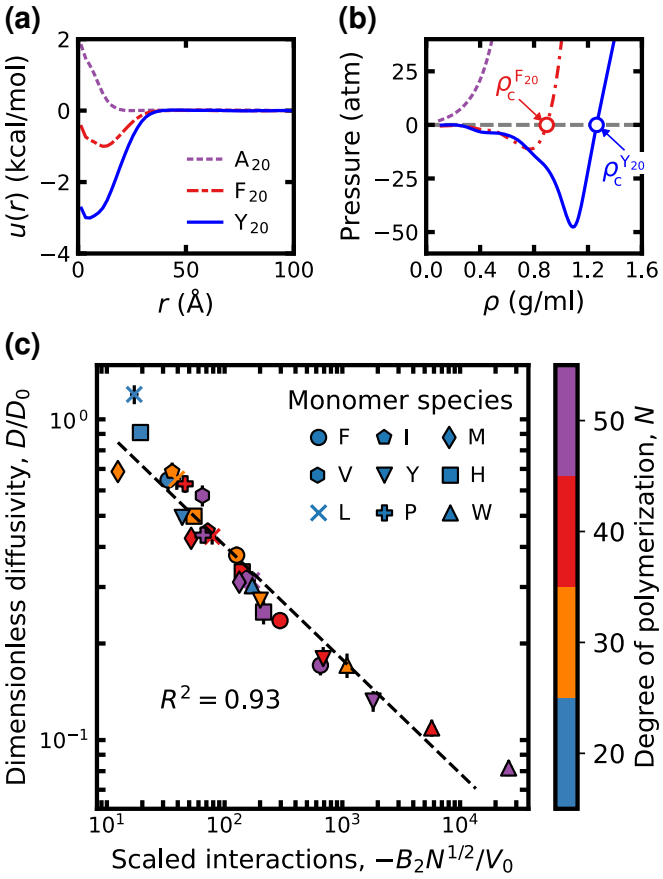

FIG. 1. **Coarse-grained simulations of homomeric polypeptides predict a strong relationship between condensate stability and internal dynamics.** **(a)** The potential of mean force, $u(r)$, between the centers of mass of two polymer chains at 300 K. Three example curves are shown for net repulsive, weakly attractive, and strongly attractive 20-mers, A_{20}, F_{20}, and Y_{20}, respectively. **(b)** Empirical equation-of-state curves at 300 K for the three homomeric polypeptides in panel **a**. Only two of these 20-mers, F_{20} and Y_{20}, are predicted to phase separate to form condensed phases with the indicated densities $\rho_c^{F_{20}}$ and $\rho_c^{Y_{20}}$, respectively. **(c)** The dimensionless self-diffusion coefficients, D/D_0, of homopolymers in the condensed phase are anticorrelated with their scaled dimensionless second-virial coefficients, $-B_2N^{1/2}/V_0$. Points are shown on a log–log plot for the 29 homopolymers (out of 80 simulated) that phase separate at 300 K. Self-diffusion coefficients are obtained from canonical-ensemble simulations conducted at the condensed-phase density, ρ_c, determined from the equation-of-state analysis. Statistical errors reflect the standard error of the mean and are comparable to the symbol size.

equal to the pervaded volume of an ideal polymer chain, $V_0 = (4\pi/3)b_0^3(N/6)^{3/2}$ [39], with the CG equilibrium bond length, $b_0 = 3.8\,\text{Å}$, and a chain length of $N = 50$.

To calculate D, we first need to verify whether phase separation occurs for a particular polymer sequence and, if so, to establish the condensed-phase polymer density, ρ_c. To this end, we apply an empirical equation-of-state (EOS) analysis to approximate ρ_c for polymers with neg-

ative second-virial coefficients (Fig. 1b; see Sec. IV B). We use short simulations of 100 polymer chains at 300 K to compute the pressure as a function of density (i.e., the EOS) and then look for a non-monotonic "van der Waals loop" in the EOS [41], which is indicative of a phase-separated region. Making the approximation that the coexistence pressure is near zero, corresponding to a dilute low-density phase, we approximate ρ_c as the highest-density zero of the EOS. By contrast, we consider strictly nonnegative EOS as evidence that phase separation does not occur at 300 K. In Sec. II D, we show that this method, which is amenable to high-throughput simulations, is accurate to within a few percent of ρ_c values determined via direct-coexistence slab simulations [28]. Finally, we compute D from the mean-squared displacement of a tagged chain in a canonical-ensemble simulation conducted at the condensed-phase density ρ_c (see Sec. IV B). We report self-diffusion coefficients relative to a reference value, $D_0 = 0.42 \times 10^{-9}\,\mathrm{m}^2/\mathrm{s}$, which corresponds to an ideal Rouse chain with chain length $N = 50$, $D_0 = \tau k_B T / N\bar{M}$ [39], where $\tau = 1\,\mathrm{ps}$ is the damping time in our Langevin simulations and $\bar{M} = 118\,\mathrm{g/mol}$ is the average molecular weight of the 20 amino acid types. Due to strong negative correlation between the viscosities of the homopolymer condensed phases and the condensed-phase self-diffusion coefficients but the lower statistical uncertainties in the latter (SI Fig. 5), we focus on self-diffusion coefficients as a simple measure of condensate dynamics in this study.

Plotting D/D_0 versus B_2/V_0, we observe a strong negative correlation between these quantities across all homopolymeric sequences that are determined to phase separate (Fig. 1c). Surprisingly, this correlation is stronger than the direct correlation between the condensed-phase density, ρ_c, with either of these quantities alone (SI Fig. 6). This indicates that even though the condensed-phase density at 300 K cannot be predicted solely on the basis of the second-virial coefficient, this quantity nonetheless captures the key physics necessary to predict the internal dynamics of a homopolymer condensate *at its equilibrium density*.

To rationalize this empirical relationship, we consider a Rouse model [39] of an unentangled polymer melt. This model can describe the condensed phases in these simulations, which have densities on the order of 1 g/ml. In this model, the self-diffusion coefficient is inversely proportional to the total frictional force experienced by the chain, which typically follows an Arrhenius law for a thermally activated process [24, 39]. It is thus reasonable to relate the logarithm of the total friction to the reversible work required to separate a pair of chains, which is proportional to $\log(-B_2)$ in the limit of strong attractive interactions. Accounting for the number of interacting chains within the pervaded volume of a tagged polymer [39], we therefore propose that the condensed-phase self-diffusion coefficient for phase-separating homopolymers should scale as $D \sim -N^{1/2} B_2$. Empirically, we find that including the prefactor $N^{1/2}$ in the scaling relationship indeed improves the correlation across the 29 phase-separating homopolymers that we simulated (Fig. 1c and SI Fig. 7). We thus conclude that this simple model captures the essential relationship between the second-virial coefficients and condensed-phase self-diffusion coefficients of homomeric polypeptide sequences.

B. Short polypeptides from known intrinsically disordered regions do not form condensed phases

Having established a correlation between D/D_0 and B_2/V_0 for homomeric polypeptides, we next examine the behavior of heteromeric polypeptides based on known IDP sequences. To do so, we perform CG simulations of 1,266 short ($20 \leq N \leq 50$) polypeptide sequences from DisProt [42], a manually curated database of disordered regions of proteins. Although these polypeptides feature diverse sequence characteristics (SI Fig. 8), none of these polypeptide sequences form condensed phases according to the criteria set forth in Sec. II A. This observation is surprising, considering the important role that many IDP sequences are known to play in intracellular phase separation [42, 43], but likely stems from the short sequence lengths considered in these calculations. The vast majority (c.a. 80%) of the sequences selected from DisProt have positive B_2 (SI Fig. S9), while the remaining 255 sequences exhibit monotonically increasing EOS curves, indicating that phase separation does not occur at 300 K when simulated using this CG IDP model. Nonetheless, we note that some of these B_2 values are more negative than some of the phase-separating homomeric polypeptides considered in Fig. 1. This observation highlights the limitations of using B_2 alone to predict phase separation [18, 44] and the importance of performing the independent EOS analysis.

C. Active learning identifies limits of tunability for condensate thermodynamic-dynamic properties

Since the lack of phase separation precludes analysis of condensed-phase dynamical properties using short DisProt sequences, we set out to generate novel heteromeric sequences and explore the thermodynamics–dynamics tradeoff beyond homopolymers. To this end, we adopt an active learning scheme [32] that iteratively selects and simulates new polypeptides to probe this relationship methodically and efficiently (Fig. 2a). The selection of new polypeptides is guided by Bayesian optimization, which has been usefully deployed in materials design, including the design of heteropolymer sequences, by informing the next "best" sequence(s) to characterize [30, 45–49]. Within the context of the thermodynamics–dynamics tradeoff, the "best" sequences are those considered as *Pareto optimal* [50, 51], meaning that no other polypeptide sequence can be found that simultaneously exhibits an increased D and a decreased B_2.

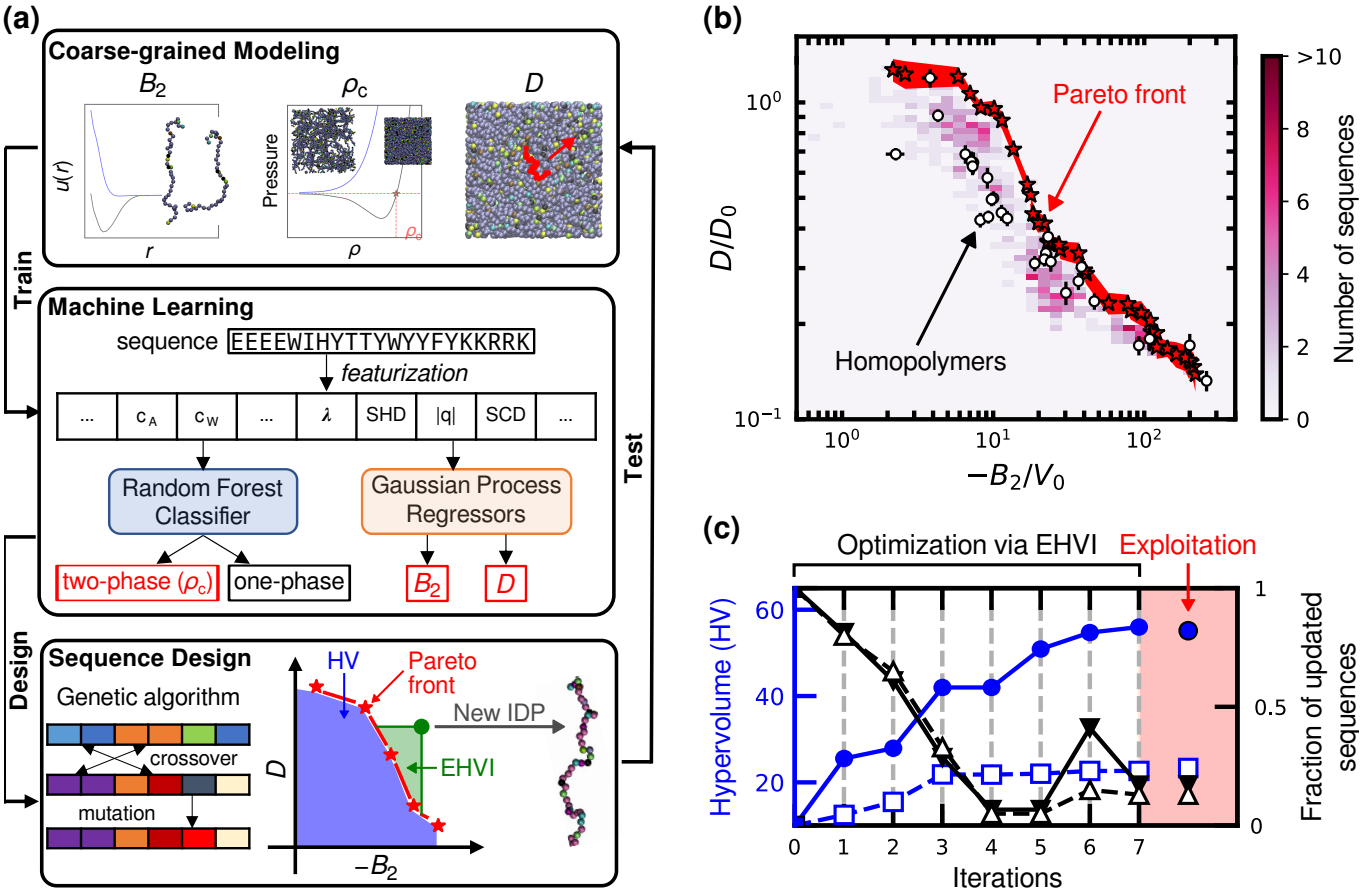

FIG. 2. **Active-learning approach to heteromeric sequence design. (a)** The workflow for designing new IDP sequences by integrating coarse-grained modeling, machine learning (ML), and genetic-algorithm-based optimization. Active learning is an iterative process consisting of (i) performing molecular dynamics simulations to compute three physical properties of IDPs (B_2, phase behavior, and D); (ii) learning the predictive features of IDP sequences by training ML models to estimate these three physical properties; (iii) designing new sequences by maximizing a fitness function (see Sec. IV C); and (iv) returning to step (i) to test the physical properties of the proposed IDP sequences. The data set and ML models are subsequently updated via successive iterations. **(b)** The relationship between the scaled condensed-phase self-diffusion coefficients, D/D_0, and scaled second-virial coefficients, $-B_2/V_0$, of all the designed IDP sequences that undergo phase separation. The color of each grid element indicates the number of distinct sequences within that region. The red highlighted area represents a one-standard-deviation region spanned by the Pareto-optimal heteromeric sequences (red stars). Homopolymer sequences (white circles) are shown for comparison. **(c)** Convergence analysis of the active-learning approach. The hypervolume under the Pareto front (blue symbols, left axis) increases with active-learning iterations, while the fraction of newly updated data points at the Pareto front (black symbols, right axis) tends to zero as the Pareto front converges. Filled symbols/solid lines show the convergence of the entire Pareto front, while open symbols/dashed lines describe the convergence of the Pareto front in the high-diffusivity regime, where $-B_2/V_0 \gtrsim 50$. A final "exploitation" round (shaded region) of sequence design confirms convergence of the hypervolume, introducing only a few new sequences into the Pareto front representation.

Polypeptides are chosen for characterization based on a policy of expected hypervolume improvement (EHVI) [52], which has previously been demonstrated to converge towards a true set of Pareto-optimal sequences [53]. These Pareto-optimal sequences are referred to as the *Pareto front*. Optimizing EHVI over sequence space is facilitated by a genetic algorithm that leverages machine-learning (ML) models, trained using the results of prior CG simulations, to predict physical properties (i.e., phase behavior, B_2, and D) from sequence features of prospective polypeptides. In particular, the ML models employ a 30-dimensional *feature vector* that incorporates the amino-acid composition [54, 55] and ten additional sequence-level descriptors, including the sequence length, N; the average hydrophobicity per residue, $\bar{\lambda}$; the sequence hydropathy decoration parameter, SHD [56]; the fraction of positively and negatively charged residues, $\bar{q}_+$ and $\bar{q}_-$; the net charge per residue, $|q|$; and the sequence charge decoration, SCD; the mean-field second virial coefficient [56], B_2^{MF}; the Shannon entropy, S; and the average molecular weight, $\bar{M}$. All the ten sequence-level feature values are normalized and standardized as described in Sec. IV C. Each iteration of active learning involves training ML models with the

current simulation data, generating 96 new polypeptide sequences via the genetic algorithm, and subsequently simulating these sequences according to the protocols introduced in Sec. II A. Additional technical details are reported in Sec. IV C; complete descriptions of ML model generation and validation are provided in SI Sec. V and SI Fig. S10.

We find that this active-learning approach rapidly identifies a representative set of Pareto-optimal polypeptides (Fig. 2b). First, using a ML model for B_2 trained using DisProt sequence data, we perform an initial round of simulations involving IDPs chosen to minimize B_2 only (Iteration 0). These simulations establish relevant data for training a ML model for D, enabling subsequent multiobjective optimization via EHVI (see Sec. IV C). We quantitatively assess convergence of the active-learning approach by monitoring both the hypervolume (HV) under the current approximation of the Pareto front and the fraction of newly added Pareto-optimal sequences relative to the current number of Pareto-optimal points (Fig. 2c). Active learning leads to a rapid expansion of the Pareto front (Iterations 1–3). Subsequent iterations (Iterations 4–7) result in substantially diminished expansion and mostly inconsequential additions to the Pareto front, especially along the high-diffusivity portion of the front (dashed lines in Fig. 2c). In these latter rounds, the expansion of the front occurs primarily via the addition of sequences with extremely negative B_2 and low D. We therefore halt optimization based on EHVI at this point and perform one last iteration of sequence design based on pure exploitation of the ML models, which confirms convergence of the Pareto front (Fig. 2c). The final data set includes 2,114 polypeptide sequences, including 35 Pareto-optimal sequences (denoted P1–P35) that provide a good representation of the true Pareto front of heteromeric polypeptides.

By contrast with homomeric polypeptides, active learning identifies "designed" heteropolymer sequences spanning a range of D values at fixed B_2 and vice versa (Fig. 2b). Although heteromeric polypeptides follow the same trend of decreasing D with decreasing B_2 as found for as homomeric polypeptides, the data cannot be neatly collapsed by any obvious scaling. For strongly attracting polypeptides ($-B_2/V_0 \gtrsim 50$), all sequences possess limited condensed-phase mobility. Most of these, including the homomeric sequences Y_{40}, W_{40}, and W_{50}, are within the statistical error region spanned by Pareto-optimal heteromeric sequences (Fig. 2b). However, more weakly interacting polypeptides at the Pareto front ($-B_2/V_0 \lesssim 50$) exhibit D values that are roughly twice those of homopolymer sequences with the same B_2. (We note that, because the Bayesian optimization only targets expansion of the Pareto front, it is likely that the maximum dynamic range of D values at fixed B_2 can exceed a factor of two.) Thus, Fig. 2b illustrates the overall prospects and limitations of tuning the thermodynamic–dynamic tradeoff via sequence design, as we discuss in Secs. II E and II F. Specifically, when IDPs interact relatively weakly, sequence design can be used to fine-tune the dynamical properties of single-component IDP condensates that possess similar phase-separation propensities, or, alternatively, to design IDP condensates that undergo phase separation below disparate critical temperatures while exhibiting similar dynamical properties.

D. Coexistence simulations validate Pareto-optimal sequences discovered via active learning

We validate the IDP sequences designed via active learning by performing large-scale molecular dynamics simulations of phase coexistence. We carry out direct coexistence simulations of three representative sequences from the thermodynamics–dynamics Pareto front (P1, P8, and P35) using 1000 chains in a slab geometry [18, 27] at 300 K. Simulation durations of at least 2 µs are required to reach equilibrium, resulting in the formation of an interface between equilibrium condensed and dilute phases (Fig. 3a). As anticipated, the differences between the condensed and dilute-phase densities are anticorrelated with the second-virial coefficients, which decrease from P1 to P35 (Fig. 3b). The interfaces between the coexisting dilute phases also become sharper as we move along the Pareto front from P1 to P8 to P35 (Fig. 3b). Both observations are consistent with the previously reported correlation between critical temperatures and second-virial coefficients in single-component IDP solutions [18].

Importantly, these simulations validate the condensed-phase dynamical properties of Pareto-optimal sequences predicted by our active-learning approach. From equilibrated simulations of phase coexistence, we determine the mean condensed-phase density and compute the self-diffusion coefficient of chains within the bulk of the condensed phase (i.e, away from the interface with the dilute phase). The physical properties determined in this way compare favorably with the values determined from the EOS analysis (Fig. 3c) and single-phase simulations of condensed-phase dynamical properties (Fig. 3d). In all cases, quantities computed via these different methods differ by less than 10% and typically agree within the statistical error. This agreement demonstrates that the high-throughput simulations utilized by our active-learning approach are sufficiently accurate compared to more computationally expensive direct-coexistence simulations. Furthermore, these results suggest that our active-learning protocol (Fig. 2c) indeed converges to a representative sequences on the thermodynamics–dynamics Pareto front for this CG IDP model.

E. Sequence features change drastically along the thermodynamics–dynamics Pareto front

To better understand what governs the heteropolymer thermodynamics–dynamics tradeoff (Fig. 2b), we ana-

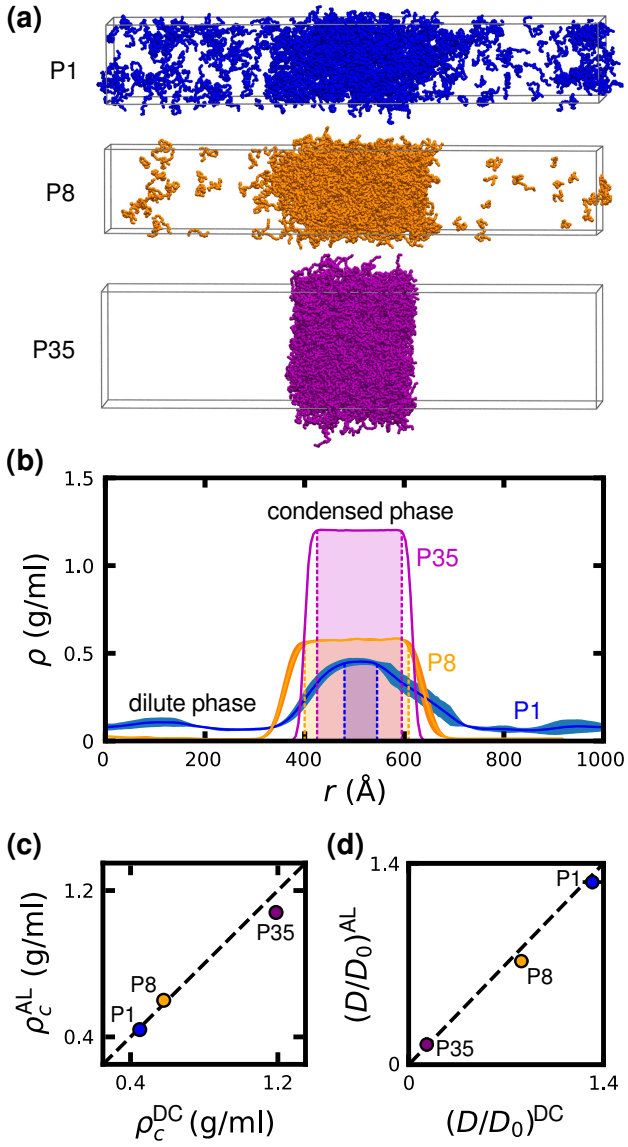

FIG. 3. **Validation of Pareto-optimal IDP sequences using direct-coexistence simulations.** **(a)** Snapshots of slab-geometry direct-coexistence simulations at 300 K for three representative IDP sequences from the thermodynamics–dynamics Pareto front (see Fig. 2b): P1 (blue), P8 (orange), and P35 (purple). Slab simulations of 1000 chains are performed in the canonical ensemble for 4.5 µs (P1) and 2 µs (P8 and P35). **(b)** Average density profiles, measured orthogonally to the interface between the dilute and condensed phases, obtained from the simulations in panel **a**. The final 3 µs of the P1 trajectory and the final 0.5 µs of the P8 and P35 trajectories are analyzed to compute these density profiles. Shaded areas indicate each bulk condensed phase, defined as the region where variations in the average density are less than 0.02 g/ml. **(c)** Comparison of the condensed-phase densities, ρ_c, and **(d)** comparison of the dimensionless condensed-phase self-diffusion coefficients, D/D_0, computed via the equation-of-state method within the active-learning framework (AL) or via the direct-coexistence (DC) simulations shown in panel **a**.

lyze the sequence characteristics of IDPs on the Pareto front (Fig. 4a, stars). An aggregate comparison among the Pareto-optimal polypeptides (P1–P35) is based on the pairwise cosine similarity of their feature vectors, where higher scores indicate greater similarity. For clarity, we decompose sequence similarity into the contributions from *(i)* the amino-acid composition (Fig. 4b, lower triangle) and *(ii)* the various sequence-derived descriptors (Fig. 4b, upper triangle). The average cosine similarities among pairs of Pareto-optimal sequences are *(i)* 0.60 and *(ii)* 0.56, respectively. For reference, the average cosine similarities among pairs of DisProt sequences ($20 \leq N \leq 50$) are *(i)* 0.57 and *(ii)* 0.22; these reference values are indicated by the vanishing points of the color bars in Fig. 4b.

Based on relative similarities, we propose that the Pareto-optimal sequences roughly divide into three groups: P1–P8, P9–P20, and P21–P35, outlining a high-B_2/high-D regime, a transition regime, and a low-B_2/low-D regime, respectively (top-left, middle, and bottom-right in Fig. 4a,b). The polypeptides P1–P8 on-average bear stronger composition and sequence-descriptor similarity than randomly selected pairs of disordered regions from DisProt but present stark differences from the other Pareto-optimal sequences. Collectively, P9–P35 generally exhibit strong composition and sequence-descriptor similarities with one another, but the transition group (P9–P20) possesses lower intra-group similarity than the low-B_2/low-D group (P21–P35).

This aggregate-level analysis reveals how sequence (dis)similarity does not necessarily dictate differences in macroscopic properties in obvious ways. On the one hand, we observe greater variability in sequence characteristics, and thus greater intra-group dissimilarity, among Pareto-optimal sequences in the transition regime (e.g., see dashed lines in SI Fig. S11). For example, P13 and P17 are quite distinct in terms of sequence descriptors relative to all other Pareto-optimal sequences, including adjacent Pareto-optimal sequences in the transition regime that exhibit similar macroscopic properties. This variability among sequence characteristics suggests enhanced degeneracy of near-Pareto-optimal polypeptides (i.e., multiple distinct sequences that display similar D and B_2 values) in the transition regime relative to the low-B_2/low-D regime, where the sequence characteristics of Pareto-optimal polypeptides are conserved to a much greater extent. On the other hand, sequences with high overall composition and sequence-descriptor similarity can present very different macroscopic properties. For example, the extremal sequences P1 and P35 exhibit greater sequence-descriptor similarity than P1 does with several other sequences in the high-B_2/high-D regime. Even within the higher-degeneracy transition regime, P9 and P20 have overall high composition and sequence-descriptor similarity despite the 2.4-fold change in B_2 and 1.9-fold change in D between these points on the Pareto front. Taken together, these observations indicate that there may be substantial flexibility in achieving Pareto-

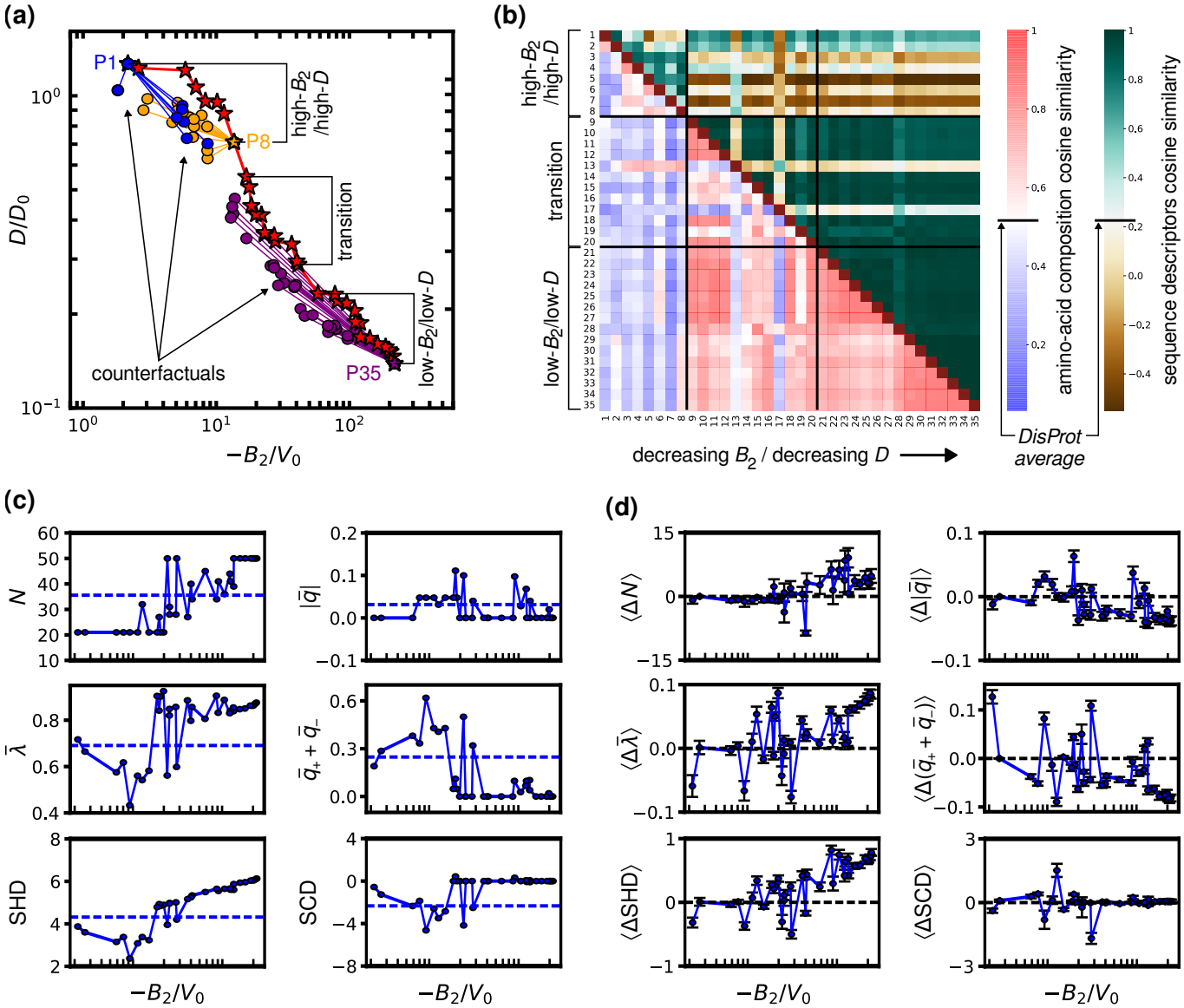

FIG. 4. **Physicochemical features of sequences at the thermodynamics–dynamics Pareto front. (a)** Representative IDP sequences along the Pareto front (stars) divide into three regimes. Also shown are counterfactual sequences (circles) of three representative Pareto-optimal sequences, P1 (blue), P8 (orange), and P35 (purple); see text for the criteria used to select counterfactual sequences. **(b)** Cosine similarities between the amino-acid composition component of the feature vectors (lower triangle, left color bar) and between the sequence-descriptor component of the feature vectors (upper triangle, right color bar) of the 35 representative Pareto-optimal sequences. **(c)** Variations of selected sequence descriptors (degree of polymerization, N; average hydrophobicity, $\bar{\lambda}$; hydropathy distribution, SHD; absolute net charge per residue, $|\bar{q}|$; the total charge, $\bar{q}_+ + \bar{q}_-$; and charge distribution, SCD) along the Pareto front. Dashed lines indicate the mean value of each feature across all phase-separating sequences studied via active learning. **(d)** Average feature differences between each Pareto-optimal sequence and its counterfactuals, Eq. (2). Error bars represent standard errors of the mean, and dashed lines indicate zero feature difference.

optimal condensate properties using diverse sequences in some cases, while in other cases selectively changing only a few sequence features may yield substantially different macroscopic condensate properties.

At a more granular level, we find that specific sequence characteristics are more prominent in different regimes. Based on previously cited importance in IDP condensates [9, 56–58], we highlight trends in sequence length, N; average sequence hydropathy, $\bar{\lambda}$; sequence hydropathy decoration, SHD; the absolute net charge per residue, $|\bar{q}|$; the fraction of charged residues (or the total charge), $\bar{q}_+ + \bar{q}_-$; and sequence charge decoration, SCD (Fig. 4c). Data for all features are shown in SI Fig. S11. These descriptors display an array of behaviors when moving across the Pareto front. For example, N is roughly sigmoidal with $-B_2$, starting at small N

in the high-B_2/high-D regime (P1–P8) before gradually but noisily increasing to plateau at large N in the low-B_2/low-D regime (P27–P35). On the other hand, $\bar{\lambda}$ and SHD exhibit U-shaped behavior, first generally decreasing in the high-B_2/high-D regime and then increasing once entering the transition region. Interestingly, the net charge, $|\bar{q}|$, of Pareto-optimal sequences is always close to zero, but there are qualitative differences in total charge, $\bar{q}_+ + \bar{q}_-$, and SCD in different regimes. In particular, $\bar{q}_+ + \bar{q}_-$ starts at finite fractional charge at high-B_2/high-D and initially increases before diminishing to zero in the transition and low-B_2/low-D regimes. SCD displays similar behavior, but with the opposite sign.

These trends suggest that different polypeptide characteristics and the physicochemical driving forces dictate the limits of the thermodynamics–dynamics tradeoff. The trends in charge-related features indicate that Pareto-optimal polypeptides are polyampholytic only in the high-B_2/high-D regime. Polypeptides in this regime also tend to be shorter and less hydrophobic than other phase-separating IDPs generated during active learning. Therefore, polyampholicity in combination with nonuniform charge distribution and weak hydrophobicity is key to achieving phase separation with short (and therefore fast-diffusing) chains. By contrast, Pareto-optimal polypeptides are largely composed of uncharged, hydrophobic amino acids in the low-B_2/low-D regime. In particular, sequences P27–P35 are predominantly composed of tyrosine, tryptophan, and phenylalanine, which have the highest hydrophobic character in the CG model and are also among the more massive amino acids. Greater hydrophobicity and longer chain lengths therefore result in both stronger inter-chain attraction and more sluggish diffusion, consistent with the observations of Ref. [9], in the low-B_2/low-D regime.

F. Pareto-optimality is characterized by regime-dependent sequence determinants

We perform a counterfactual analysis [36, 59, 60] of the Pareto-optimal IDPs to gain further insight into the sequence determinants of the heteropolymer thermodynamics–dynamics tradeoff. In this context, a "counterfactual" is an IDP sequence that bears strong sequence similarity to a Pareto-optimal sequence but exhibits distinctly different physical properties (Fig. 4a). Such counterfactuals can therefore be used to attribute Pareto-optimality to small variations in IDP sequences. Here, we consider a polypeptide to be a counterfactual of a specific sequence on the Pareto front if *(i)* the cosine similarity of their feature vectors is greater than 0.9 and *(ii)* the counterfactual candidate is within a dimensionless distance of 0.15–0.3 away from the Pareto front; for this calculation, we use a Euclidean distance in the in the standard-normalized B_2–D plane (see SI Sec. V E for details). Because each Pareto-optimal polypeptide may have several counterfactuals satisfying the aforementioned criteria, we examine the average feature differences,

$$\langle \Delta x_k \rangle \equiv \frac{1}{n_{\mathcal{C}_{ij}}} \sum_{j=1}^{n_{\mathcal{C}_{ij}}} (x_k^{(\mathcal{P}_i)} - x_k^{(\mathcal{C}_{ij})}), \tag{2}$$

where k denotes a particular feature of a feature vector $\vec{x}$, $\mathcal{P}_i$ is the ith Pareto-optimal sequence with its value of x_k indicated as $x_k^{(\mathcal{P}_i)}$, $\mathcal{C}_{ij}$ is the jth counterfactual of the ith Pareto-optimal sequence with its value of x_k indicated as $x_k^{(\mathcal{C}_{ij})}$, and $n_{\mathcal{C}_{ij}}$ is the total number of counterfactuals of the ith Pareto-optimal sequence. We then analyze statistically resolvable differences between the feature vectors of Pareto-optimal/counterfactual sequence pairs to provide explanations for disparities among physical properties.

Plotting $\langle \Delta x_k \rangle$ for all features versus $-B_2/V_0$ (Fig. 4d and SI Fig. S12), we find that the distinguishing characteristics of Pareto-optimal polypeptides are regime-dependent. In the low-B_2/low-D regime, Pareto-optimal sequences are generally longer, feature more hydrophobic residues, and have larger values of the sequence hydropathy decoration parameter, SHD, relative to counterfactual sequences. Pareto-optimal sequences also tend to have fewer charged residues and a lower net charge than counterfactuals in this regime. Since the Pareto-optimal sequences in this regime are neutral (Fig. 4c), this latter observation implies that the counterfactual sequences possess a net charge despite having near zero values of the sequence charge decoration parameter, SCD (Fig. 4c,d). This analysis therefore indicates that hydrophobic interactions are the dominant determinants of Pareto-optimal sequences in the low-B_2/low-D regime.

Outside the low-B_2/low-D regime, differences between counterfactuals and Pareto-optimal sequences are more difficult to resolve. In the high-B_2/high-D regime, no single sequence descriptor differs consistently between Pareto-optimal and counterfactual sequences. Instead, variations in B_2 and D are manifest by subtle manipulations, such as exchanging residues with comparable hydrophobicity but different masses (SI Fig. S12). Within the transition regime, there are statistically significant differences between features of Pareto-optimal polypeptides and counterfactuals, but the directions of such differences are not necessarily preserved across the Pareto front. In particular, the hydrophobicity differences, $\langle \Delta \bar{\lambda} \rangle$, and total charge differences, $\langle \Delta(\bar{q}_+ + \bar{q}_-) \rangle$, fluctuate between positive and negative values within this regime (P9–P20 in Fig. 4d). This consistently enhanced variability across multiple features (see also SI Fig. S12) can be attributed to the greater degeneracy of sequence space near the Pareto front in the transition regime, as discussed in Sec. II E. Overall, this counterfactual analysis indicates that the relationship between sequence features and Pareto optimality is nuanced, since relatively subtle changes from counterfactuals dictate behavior in the high-B_2/high-D regime and myriad feature combina-

tions exist for achieving near-optimality in the transition regime.

III. DISCUSSION

Using an active-learning approach, we have investigated the extent to which thermodynamic and dynamical properties are coupled in IDP condensates. Our approach combines high-throughput simulations of a coarse-grained model of IDPs, Bayesian active learning strategies for predicting physical properties from IDP sequences, and numerical optimization techniques to efficiently explore sequence space. In this way, we demonstrate that the anticipated tradeoff between thermodynamic stability and internal condensate dynamics is a general feature of dense protein-based condensates. The correlation between proxies for thermodynamic stability and condensed-phase dynamics is strongest in the case of *homomeric* polypeptides. In particular, both the thermodynamic stability and the condensed-phase dynamics of homomeric polypeptide condensates can be predicted with high accuracy on the basis of the second-virial coefficient and the chain length alone. However, the correlation between the thermodynamic and dynamic properties weakens when we consider *heteromeric* IDP sequences, indicating that we can design sequences to tune the thermodynamic and dynamic properties of IDP condensates independently. Consistent with this notion, we observe that the second-virial coefficient is not always sufficient for predicting phase separation of heteromeric IDP sequences, and the condensed-phase self-diffusion coefficient can be increased by approximately a factor of two or more relative to homomeric polypeptides with the same second-virial coefficient.

Fundamentally, deviations from the behavior of homomeric polypeptides arise due to the effects of sequence patterning. By identifying a representative set of Pareto-optimal polypeptides—heteromeric sequences that outline a limiting tradeoff boundary with respect to the second-virial and diffusion coefficients—we find that the sequence determinants of the thermodynamics–dynamics tradeoff are surprisingly nuanced. Whereas the extremal regions (high-B_2/high-D and low-B_2/low-D) display consistent trends with respect to the sequence features, the transition regime is typified by high sequence variability, which likely reflects greater degeneracy with respect to sequence space within this regime. Overall, we find that chain length, hydrophobicity, and hydrophobic patterning exhibit consistent trends along the entirety of the Pareto front, irrespective of regime. On the other hand, total charge and charge patterning descriptors are primarily relevant in the high-B_2/high-D regime that is dominated by polyampholytes. Lastly, our sequence and counterfactual analysis indicates that relatively minor sequence perturbations can result in large changes to both B_2 and D, particularly in the transition regime. These findings provide insights into the design principles governing different regimes of physicochemical parameters and highlight the importance of considering multiple sequence features when optimizing polypeptide sequences to achieve specific physical properties.

We emphasize that our numerical results apply to single-component IDP condensates and rely on the accuracy of the underlying coarse-grained IDP model, which lacks physicochemical features such as directional sidechain interactions, hydrogen bonds, and secondary structure. Coarse-grained models are also known to predict dynamical properties with relatively low fidelity compared to all-atom models with many more degrees of freedom [61–63]. It is likely that a greater range of variability of dynamical properties would be predicted by a model that incorporates such features, especially if such modifications were to allow for more accurate descriptions of low-density condensed phases [16]. Nonetheless, we expect that our findings regarding the relative importance of chain length, charge patterning, and hydrophobicity along the thermodynamics–dynamics Pareto front are robust with respect to alternative coarse-grained IDP models [19, 29, 64]. Experiments [16] and higher-resolution simulations [61] will be needed to test such predictions.

A major strength of our approach is its ability to explore IDP sequence space efficiently and characterize relationships among disparate biophysical properties. Crucially, the strategy of combining high-throughput simulations and Bayesian-guided optimization is independent of any limitations of the underlying coarse-grained model. Moreover, we have demonstrated that this approach can efficiently converge the Pareto front for a pair of biophysical properties that depend on a complex sequence of computations. We anticipate that this framework could be applied to other combinations of biophysical traits and may also be useful for studying multicomponent and multiphasic systems [65–67]. Our work therefore establishes a promising approach for engineering custom biomolecular condensates with tunable thermodynamic and dynamic properties.

IV. METHODS

A. Model of intrinsically disordered proteins

The coarse-grained (CG) model developed by Regy et al. [26] represents IDP sequences, comprising 20 amino-acid types, in implicit solvent. The force field consists of a harmonic bonded potential as well as short-range van der Waals (vdW) and long-range electrostatic (el) nonbonded interactions,

$$U_{\mathrm{tot}} = \sum_i k_{\mathrm{b}}(r_{i,i+1} - b_0)^2 + \sum_{i,j} \phi^{\mathrm{vdw}}(r_{ij}) + \sum_{i,j} \phi^{\mathrm{el}}(r_{ij}), \tag{3}$$

where $r_{i,i+1}$ is the distance between bonded residues i and $i+1$, the force constant is $k_{\mathrm{b}} = 10\,\mathrm{kcal}/(\mathrm{mol}\cdot\mathrm{\AA}^2)$,

and the equilibrium bond length is $b_0 = 3.82$ Å. The vdW interaction takes the Ashbaugh–Hatch functional form,

$$\phi^{\mathrm{vdW}}(r_{ij}) = \begin{cases} \phi^{\mathrm{LJ}}(r_{ij}) + (1-\lambda_{ij})\epsilon & r_{ij} \leq 2^{1/6}\sigma_{ij} \\ \lambda_{ij}\phi^{\mathrm{LJ}}(r_{ij}) & r_{ij} > 2^{1/6}\sigma_{ij} \end{cases} \tag{4}$$

where r_{ij} is the distance between nonbonded residues i and j, and $\phi^{\mathrm{LJ}}(r_{ij}) = 4\epsilon[(\sigma_{ij}/r_{ij})^{12} - (\sigma_{ij}/r_{ij})^{6}]$ is the Lennard–Jones potential with $\epsilon = 0.2$ kcal/mol. The interaction strength, $\lambda_{ij} = (\lambda_i + \lambda_j)/2$, and distance, $\sigma_{ij} = (\sigma_i + \sigma_j)/2$, parameters are determined from the hydropathy scaling factors, λ_i, and vdW diameters, σ_i, of the residues, respectively [26]. Lastly, screened electrostatic interactions take the form

$$\phi^{\mathrm{el}}(r_{ij}) = \frac{q_i q_j}{4\pi D r_{ij}} e^{-\kappa r_{ij}}, \tag{5}$$

where q_i is the charge of residue i, $D = 80$ is the dielectric constant of the solvent, and $\kappa = 10$ Å is the Debye screening length.

B. Physical property calculations

Second-virial coefficient. Second-virial coefficients are determined by calculating the potential of mean force, $u(r)$, as a function of the center-of-mass (COM) distance, r, between an isolated pair of chains. These calculations are carried out using the adaptive biasing force (ABF) method [40] implemented in the COLVARS package in LAMMPS [68]. A Langevin thermostat is applied to maintain a constant temperature of 300 K. Two molecules are placed in a cubic simulation box with dimensions $300 \times 300 \times 300$ Å. ABF simulations are performed as a function of r, which varies from 1 Å to 100 Å. Force statistics are stored in bins of width 0.2 Å. The biasing force is applied after 1000 samples are collected in each bin, after which a final production run is performed for 5 µs with a timestep of 10 fs. The second-virial coefficient is then obtained from Eq. (1). Thirty independent simulations are performed for each free energy calculation to determine the statistical error.

Condensed-phase density. Considering the high computation costs of slab simulations, we developed an alternative method to determine whether a CG sequence undergoes phase separation at 300 K. To this end, we look for evidence of a van der Waal's loop in the equation of state (EOS) of a small system. Specifically, we compute the pressure using canonical-ensemble simulations of 100 chains with periodic boundary conditions at densities ranging from 0.2 g/ml–1.2 g/ml. A Langevin thermostat is applied to maintain a constant temperature of 300 K. Simulations are carried out for 100 ns with a timestep of 10 fs. We assume a coexistence pressure of zero for simplicity, so that phase separation can be identified by the existence of a negative pressure at a finite density. In cases where negative pressures were observed, we fit a cubic spline to the EOS to determine the condensed-phase density, ρ_{c}, defined as the highest density at which the pressure crossed zero. The pressure values are bootstrapped with replacement 50 times to determine the average condensate densities and their standard errors.

Self-diffusion coefficient. To determine the self-diffusion coefficient of a chain within the condensed phase, canonical-ensemble simulations of 100 chains are performed at the condensed-phase density, ρ_{c}, with periodic boundary conditions. The long-time behavior of the mean-squared displacement is then used to compute the self-diffusion coefficient,

$$D = \lim_{\Delta t \to \infty} \frac{1}{6\Delta t} \langle |\vec{r}_{t+\Delta t} - \vec{r}_t|^2 \rangle, \tag{6}$$

where $\vec{r}$ is the position of the COM of a tagged chain. The temperature is maintained at 300 K using a Langevin thermostat with a damping frequency of 1 ps^{-1}. Simulations are carried out for 100 ns with a timestep of 10 fs. For each diffusion-coefficient calculation, the statistical error is quantified using 30 independent simulations.

C. Design of novel intrinsically disordered proteins

Overview of framework. We deploy an active learning [32, 51] framework based on Bayesian optimization [34] to identify IDP sequences that define a Pareto front of the properties B_2 and D. The active learning framework utilizes coarse-grained simulations to generate data for B_2 as well as pressure-density data and D, as appropriate. These data are used to train three different machine-learning (ML) models: two Gaussian process regressors [69] (GPR) and one random forest (RF) binary classifier [70]. These ML models are used to make surrogate predictions during the optimization step of active learning, during which new IDP sequences are selected for simulation. Sequence optimization is carried out using a genetic algorithm, as described below. Following optimization, simulations are performed to generate new data and begin the next iteration of active learning.

Machine-learning models. GPR models are trained to predict B_2 and D, while a RF model is trained to predict whether a given IDP sequence will undergo phase separation. For a given sequence i, the input for all ML models is $\vec{x}^{(i)}$, size-explicit augmented fingerprint as a feature vector [54]. In particular, the feature vector is 30-dimensional, consisting of 20 features reflecting the composition of the amino acids and ten sequence-level chain descriptors. The elements of the feature vector are all normalized via linear transformation methods to be unit-scale. Additional details regarding the featurization, training, and hyperparameter optimization for all ML models are provided in SI Sec. V. Although model accuracy is not a primary objective, the three trained ML models show high accuracy in predicting the aforementioned properties (SI Fig. S10) at the conclusion of

active learning. The coefficients of determination R^2 of the two GPR models are as high as 0.975 ± 0.015 and 0.970 ± 0.015, and the accuracy of the RF classifier is 0.96.

Sequence optimization. Different fitness functions were used for sequence design by a genetic algorithm depending on the active-learning iteration. Denoting the iteration number by the index k, the fitness function $f(\vec{x}^{(i)})$ is

$$f(\vec{x}^{(i)}) = \begin{cases} -B_2(\vec{x}^{(i)}) & \text{if } k = 0 \\ \mathrm{EHVI}(\vec{x}^{(i)}) \times \mathrm{RF}(\vec{x}^{(i)}) \times \alpha(\vec{x}^{(i)}) & \text{if } 0 < k < 8 \\ -\tilde{B}_2(\vec{x}^{(i)}) + \tilde{D}(\vec{x}^{(i)}) & \text{if } k = 8, \end{cases} \tag{7}$$

where $\mathrm{EHVI}(\vec{x}^{(i)})$ is the expected hypervolume improvement acquisition function, $\mathrm{RF}(\vec{x}^{(i)})$ is the output label of the random forest classifier (evaluating to 1 if phase separation is predicted and to 0 otherwise), $\alpha(\vec{x}^{(i)})$ is a scaling function that biases against generation of sequences that have high similarity to previously simulated or proposed sequences, and $\tilde{A}$ indicates a standard-normalized transformation of the property A based on the data acquired up to the given iteration. For each iteration, 96 unique sequences are generated. In Eq. (7), we compute the EHVI [52] for a given sequence $\vec{x}^*$,

$$\begin{aligned} \mathrm{EHVI}(\vec{x}^*, \mathcal{P}') = \sum_{i=1}^{n+1} & \Psi(\tilde{D}^{(i)}, \tilde{D}^{(i)}, \tilde{D}^*, \sigma^*_{\tilde{D}}) \\ & \times \Bigg[(\tilde{B}_2^{(i)} - \tilde{B}_2^{(i-1)}) \Phi\Big(\frac{\tilde{B}_2^* - \tilde{B}_2^{(i)}}{\sigma^*_{\tilde{B}_2}} \Big) \\ & + \Big(\Psi(-\tilde{B}_2^{(i-1)}, -\tilde{B}_2^{(i-1)} - \tilde{B}_2^*, \sigma^*_{\tilde{B}_2}) \\ & - \Psi(-\tilde{B}_2^{(i-1)}, -\tilde{B}_2^{(i)}, -\tilde{B}_2^*, \sigma^*_{\tilde{B}_2}) \Big) \Bigg], \end{aligned} \tag{8}$$

where $\tilde{B}_2^*$ and $\tilde{D}^*$ are predicted values of the second-virial and diffusion coefficients, respectively, obtained from the GPR models; $\sigma^*_{\tilde{B}_2}$ and $\sigma^*_{\tilde{D}}$ are corresponding uncertainty estimates from the GPR models; $\mathcal{P}' = \{\vec{x}^{(0)}, \vec{x}^{(1)}, \ldots, \vec{x}^{(n)}, \vec{x}^{(n+1)}\}$ represents the current n-point approximation of the Pareto front, augmented with additional reference points $\vec{x}^{(0)} = (X, -\infty)^\top$ and $\vec{x}^{(0)} = (-\infty, X)^\top$ to facilitate the hypervolume calculation; $\Phi(s)$ denotes the cumulative probability distribution function for a standard normal distribution of s; and Ψ is a function defined as

$$\Psi(a, b, \mu, \sigma) = \int_{-\infty}^{b} \frac{(a - z)}{\sigma} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2\sigma^2}} dz. \tag{9}$$

The similarity penalty $\alpha(\vec{x^{(j)}})$ is defined to be

$$\alpha(\vec{x}^{(j)}) = \frac{j - 1}{\sum_{k=1}^{j-1} \left\{ \frac{1}{2} [1 - \vec{x}^{(j)} \cdot \vec{x}^{(k)} / (|\vec{x}^{(j)}||\vec{x}^{(k)}|)] \right\}^{-1}}, \tag{10}$$

where $j = 2, 3, \ldots, 96$ are indices of candidate sequences generated in a given iteration. When a sequence is highly similar to one or more previously generated sequences, α tends to zero. For $j = 1$, we take $\alpha = 1$ (i.e., no penalty is applied).

The active learning process is seeded with data obtained from simulations of 1,266 disordered sequences reported in DisProt [71]. We select and simulate all sequences with sequence lengths between 20 and 50 residues (inclusive), taking care to remove duplicated samples. Since none of these sequences are determined to phase separate, we generate new sequences by maximizing a fitness function that involves only B_2 in iteration 0. The resulting data enable the training of ML models for B_2, D, and phase separation. Then in subsequent iterations, we maximize a fitness function based principally on the expected hypervolume improvement (EHVI) to identify polypeptides that outline a Pareto front of $-B_2$ versus D. To confirm the convergence of the Pareto front, a final exploitation-only round of optimization is performed by maximizing a simple linear function of the standard-normalized properties $-\tilde{B}_2$ and $\tilde{D}$.

For each iteration, 96 "child" sequences are successively generated that aim to maximize the relevant fitness function in Eq. (7). To facilitate convergence towards selecting each child sequence, 96 independent trials of sequence optimization are executed in parallel, and only the sequence with the best fitness score is chosen for explicit simulation; overall, this means that 96×96 sequence optimizations are performed in each iteration.

In iteration 0, the initial "parent" sequences are DisProt sequences with negative B_2. In all subsequent iterations, the parent sequences are Pareto-optimal sequences identified from the previous iteration. Candidate child sequences are derived from the parent sequences via a combination of crossover and mutation moves. In addition, "deletion" and "growth" moves are performed, which either remove a portion of the current sequence or add a portion of the current sequence within itself. These moves are executed *independently* in the order of crossover, mutation, deletion and growth with probabilities of 0.5, 0.8, 0.2, and 0.5, respectively. These moves are performed for 100 steps, which is sufficient for the genetic algorithm to converge to a local maximum of the fitness function (SI Fig. S13). Additional details can be found in SI Sec. V D.

ACKNOWLEDGMENTS

This work was supported by a grant from the Princeton Biomolecular Condensate Program to M.A.W. and W.M.J. The authors also acknowledge assistance from Princeton Research Computing at Princeton University, which is a consortium led by the Princeton Institute for Computational Science and Engineering (PICSciE) and Office of Information Technology's Research Computing.

[1] Brangwynne CP, Eckmann CR, Courson DS, Rybarska A, Hoege C, Gharakhani J, Jülicher F, Hyman AA. Germline P granules are liquid droplets that localize by controlled dissolution/condensation. Science. 2009;324(5935):1729–1732.

[2] Gomes E, Shorter J. The molecular language of membraneless organelles. J Biol Chem. 2019;294(18):7115–7127.

[3] Mittag T, Bremer A, Farag M, Borcherds WM, Pappu RV. Phase behavior of intrinsically disordered prion-like domains. The FASEB J. 2022;36.

[4] Kang JY, Wen Z, Pan D, Zhang Y, Li Q, Zhong A, Yu X, Wu YC, Chen Y, et al. LLPS of FXR1 drives spermiogenesis by activating translation of stored mRNAs. Science. 2022;377(6607):eabj6647.

[5] He G, GrandPre T, Wilson H, Zhang Y, Jonikas MC, Wingreen NS, Wang Q. Phase-separating pyrenoid proteins form complexes in the dilute phase. Commun Biol. 2023;6(1):19.

[6] Shin Y, Brangwynne CP. Liquid phase condensation in cell physiology and disease. Science. 2017;357(6357):eaaf4382.

[7] Deng NN, Huck WT. Microfluidic formation of monodisperse coacervate organelles in liposomes. Angew Chem. 2017;129(33):9868–9872.

[8] Kojima T, Takayama S. Membraneless compartmentalization facilitates enzymatic cascade reactions and reduces substrate inhibition. ACS Appl Mater & Interfaces. 2018;10(38):32782–32791.

[9] Hazra MK, Levy Y. Biophysics of phase separation of disordered proteins is governed by balance between short-and long-range interactions. J Phys Chem B. 2021;125(9):2202–2211.

[10] Zhao YG, Zhang H. Phase separation in membrane biology: the interplay between membrane-bound organelles and membraneless condensates. Dev Cell. 2020;55(1):30–44.

[11] Alshareedah I, Moosa MM, Pham M, Potoyan DA, Banerjee PR. Programmable viscoelasticity in protein-RNA condensates with disordered sticker-spacer polypeptides. Nat Commun. 2021;12(1):1–14.

[12] Cai L, Wang GG. Through the lens of phase separation: intrinsically unstructured protein and chromatin looping. Nucleus. 2023;14(1):2179766.

[13] Price KL, Presler M, Uyehara CM, Shakes DC. The intrinsically disordered protein SPE-18 promotes localized assembly of MSP in Caenorhabditis elegans spermatocytes. Development. 2021;148(5):dev195875.

[14] Mediani L, Antoniani F, Galli V, Vinet J, Carra AD, Bigi I, Tripathy V, Tiago T, Cimino M, et al. Hsp90-mediated regulation of DYRK3 couples stress granule disassembly and growth via mTORC1 signaling. EMBO Rep. 2021;22(5):e51740.

[15] Agarwal A, Rai SK, Avni A, Mukhopadhyay S. An intrinsically disordered pathological prion variant Y145Stop converts into self-seeding amyloids via liquid–liquid phase separation. Proc Natl Acad Sci. 2021;118(45):e2100968118.

[16] Wei MT, Elbaum-Garfinkle S, Holehouse AS, Chen CCH, Feric M, Arnold CB, Priestley RD, Pappu RV, Brangwynne CP. Phase behaviour of disordered proteins underlying low density and high permeability of liquid organelles. Nat Chem. 2017;9(11):1118–1125.

[17] Feric M, Vaidya N, Harmon TS, Mitrea DM, Zhu L, Richardson TM, Kriwacki RW, Pappu RV, Brangwynne CP. Coexisting liquid phases underlie nucleolar subcompartments. Cell. 2016;165(7):1686–1697.

[18] Dignon GL, Zheng W, Best RB, Kim YC, Mittal J. Relation between single-molecule properties and phase behavior of intrinsically disordered proteins. Proc Natl Acad Sci. 2018;115(40):9929–9934.

[19] Das S, Amin AN, Lin YH, Chan HS. Coarse-grained residue-based models of disordered protein condensates: utility and limitations of simple charge pattern parameters. Phys Chem Chem Phys. 2018;20(45):28558–28574.

[20] Das S, Lin YH, Vernon RM, Forman-Kay JD, Chan HS. Comparative roles of charge, π, and hydrophobic interactions in sequence-dependent phase separation of intrinsically disordered proteins. Proc Natl Acad Sci. 2020;117(46):28795–28805.

[21] Fung HYJ, Birol M, Rhoades E. IDPs in macromolecular complexes: the roles of multivalent interactions in diverse assemblies. Curr Opin Struct Biol. 2018;49:36–43.

[22] Harding SE, Johnson P. The concentration-dependence of macromolecular parameters. Biochem J. 1985;231(3):543–547.

[23] Laghmach R, Alshareedah I, Pham M, Raju M, Banerjee PR, Potoyan DA. RNA chain length and stoichiometry govern surface tension and stability of protein-RNA condensates. Iscience. 2022;25(4):104105.

[24] Wadsworth GM, Zahurancik WJ, Zeng X, Pullara P, Lai LB, Sidharthan V, Pappu RV, Gopalan V, Banerjee PR. RNAs undergo Phase Transitions with lower critical solution temperatures. bioRxiv. 2022; p. 2022–10.

[25] Oldfield CJ, Dunker AK. Intrinsically disordered proteins and intrinsically disordered protein regions. Annu Rev Biochem. 2014;83:553–584.

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

[27] Dignon GL, Zheng W, Mittal J. Simulation methods for liquid–liquid phase separation of disordered proteins. Curr Opin Chem. 2019;23:92–98.

[28] Dignon GL, Zheng W, Kim YC, Best RB, Mittal J. Sequence determinants of protein phase behavior from a coarse-grained model. PLoS Comput Biol. 2018;14(1):e1005941.

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

[30] Webb MA, Jackson NE, Gil PS, de Pablo JJ. Targeted sequence design within the coarse-grained polymer genome. Sci Adv. 2020;6(43):eabc6216.

[31] Gormley AJ, Webb MA. Machine learning in combinatorial Polym. Chem. Nat Rev Mater. 2021;.

[32] Smith JS, Nebgen B, Lubbers N, Isayev O, Roitberg AE. Less is more: Sampling chemical space with active learning. J Chem Phys. 2018;148(24):241733.

[33] Kim C, Chandrasekaran A, Jha A, Ramprasad R. Active-

learning and materials design: the example of high glass transition temperature polymers. MRS Communications. 2019;9(3):860–866. doi:10.1557/mrc.2019.78.

[34] Shahriari B, Swersky K, Wang Z, Adams RP, De Freitas N. Taking the human out of the loop: A review of Bayesian optimization. Proceedings of the IEEE. 2015;104(1):148–175.

[35] Wellawatte GP, Seshadri A, White AD. Model agnostic generation of counterfactual explanations for molecules. Chem Sci. 2022;13(13):3697–3705. doi:10.1039/d1sc05259d.

[36] Wellawatte GP, Gandhi HA, Seshadri A, White AD. A Perspective on Explanations of Molecular Prediction Models. J Chem Theory Comput. 2023;19(8):2149–2160.

[37] Hazra MK, Levy Y. Affinity of disordered protein complexes is modulated by entropy–energy reinforcement. Proc Natl Acad Sci. 2022;119(26):e2120456119.

[38] Tuinier R, Vliegenthart GA, Lekkerkerker HN. Depletion interaction between spheres immersed in a solution of ideal polymer chains. J Chem Phys. 2000;113(23):10768–10775.

[39] Rubinstein M, Colby RH, et al. Polymer physics. vol. 23. Oxford university press New York; 2003.

[40] Comer J, Gumbart JC, Hénin J, Lelièvre T, Pohorille A, Chipot C. The adaptive biasing force method: Everything you always wanted to know but were afraid to ask. J Phys Chem B. 2015;119(3):1129–1151.

[41] Binder K, Block BJ, Virnau P, Tröster A. Beyond the van der Waals loop: What can be learned from simulating Lennard-Jones fluids inside the region of phase coexistence. Am J Phys. 2012;80(12):1099–1109.

[42] Hatos A, Hajdu-Soltész B, Monzon AM, Palopoli N, Álvarez L, Aykac-Fas B, Bassot C, Benítez GI, Bevilacqua M, et al. DisProt: intrinsic protein disorder annotation in 2020. Nucleic Acids Res. 2019;.

[43] Li Q, Peng X, Li Y, Tang W, Zhu J, Huang J, Qi Y, Zhang Z. LLPSDB: a database of proteins undergoing liquid–liquid phase separation In Vitro. Nucleic Acids Res. 2020;48(D1):D320–D327.

[44] Rekhi S, Devarajan DS, Howard MP, Kim YC, Nikoubashman A, Mittal J. Role of Strong Localized vs Weak Distributed Interactions in Disordered Protein Phase Separation. J Phys Chem B. 2023;127(17):3829–3838.

[45] Shmilovich K, Mansbach RA, Sidky H, Dunne OE, Panda SS, Tovar JD, Ferguson AL. Discovery of Self-Assembling pi-Conjugated Peptides by Active Learning-Directed Coarse-Grained Mol. Simul. J Phys Chem B. 2020;124(19):3873–3891.

[46] Tamasi MJ, Patel RA, Borca CH, Kosuri S, Mugnier H, Upadhya R, Murthy NS, Webb MA, Gormley AJ. Machine Learning on a Robotic Platform for the Design of Polymer–Protein Hybrids. Adv Mater. 2022; p. 2201809.

[47] Kosuri S, Borca CH, Mugnier H, Tamasi M, Patel RA, Perez I, Kumar S, Finkel Z, Schloss R, et al. Machine-Assisted Discovery of Chondroitinase ABC Complexes toward Sustained Neural Regeneration. Adv Healthc Mater. 2022;11(10):2102101.

[48] Jablonka KM, Jothiappan GM, Wang S, Smit B, Yoo B. Bias free multiobjective active learning for materials design and discovery. Nat Commun. 2021;12(1):2312.

[49] Tran A, Tranchida J, Wildey T, Thompson AP. Multi-fidelity machine-learning with uncertainty quantification and Bayesian optimization for materials design: Application to ternary random alloys. J Chem Phys. 2020;153(7):074705.

[50] Emmerich M, Klinkenberg Jw. The computation of the expected improvement in dominated hypervolume of Pareto front approximations. Rapport technique, Leiden University. 2008;34:7–3.

[51] Zuluaga M, Krause A, Püschel M. ϵ-PAL: An Active Learning Approach to the Multi-Objective Optimization Problem. J Mach Learn Res. 2016;17(1):3619–3650.

[52] Yang K, Emmerich M, Deutz A, Bäck T. Multi-Objective Bayesian Global Optimization using expected hypervolume improvement gradient. Swarm Evol Comput. 2019;44:945–956. doi:10.1016/j.swevo.2018.10.007.

[53] MacLeod BP, Parlane FG, Rupnow CC, Dettelbach KE, Elliott MS, Morrissey TD, Haley TH, Proskurin O, Rooney MB, et al. A self-driving laboratory advances the Pareto front for material properties. Nat Commun. 2022;13(1):995.

[54] Patel RA, Borca CH, Webb MA. Featurization strategies for polymer sequence or composition design by machine learning. Mol Syst Des Eng. 2022;7:661–67610.1039/c4py01217h.

[55] Patel RA, Webb MA. Data-Driven Design of Polymer-Based Biomaterials: High-throughput Simulation, Experimentation, and Machine Learning. ACS Appl Bio Mater. 2023;.

[56] Zheng W, Dignon G, Brown M, Kim YC, Mittal J. Hydropathy patterning complements charge patterning to describe conformational preferences of disordered proteins. J Phys Chem Lett. 2020;11(9):3408–3415.

[57] Statt A, Casademunt H, Brangwynne CP, Panagiotopoulos AZ. Model for disordered proteins with strongly sequence-dependent liquid phase behavior. J Chem Phys. 2020;152(7):075101.

[58] Mao AH, Crick SL, Vitalis A, Chicoine CL, Pappu RV. Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. Proc Natl Acad Sci. 2010;107(18):8183–8188.

[59] Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv JL & Tech. 2017;31:841.

[60] Dandl S, Molnar C, Binder M, Bischl B. Multi-objective counterfactual explanations. In: International Conference on Parallel Problem Solving from Nature. Springer; 2020. p. 448–469.

[61] Salvi N, Abyzov A, Blackledge M. Solvent-dependent segmental dynamics in intrinsically disordered proteins. Sci Adv. 2019;5(6):eaax2348.

[62] Rudzinski JF. Recent Progress towards Chemically-Specific Coarse-Grained Simulation Models with Consistent Dynamical Properties. Computation. 2019;7(3):42.

[63] Dhamankar S, Webb MA. Chemically specific coarse-graining of polymers: Methods and prospects. J of Poly Sci. 2021; p. 2613.

[64] Garaizar A, Espinosa JR. Salt dependent phase behavior of intrinsically disordered proteins from a coarse-grained model with explicit water and ions. J Chem Phys. 2021;155(12):125103.

[65] Jacobs WM. Theory and simulation of multiphase coexistence in biomolecular mixtures. J Chem Theory Comput. 2023;.

[66] Do S, Lee C, Lee T, Kim DN, Shin Y. Engineering DNA-based synthetic condensates with programmable material

properties, compositions, and functionalities. Sci Adv. 2022;8(41):eabj1771.

[67] Chew PY, Joseph JA, Collepardo-Guevara R, Reinhardt A. Physical determinants of multiphase organisation in multi-component protein/RNA condensates. Biophys J. 2023;122(3):295a–296a.

[68] Thompson AP, Aktulga HM, Berger R, Bolintineanu DS, Brown WM, Crozier PS, in't Veld PJ, Kohlmeyer A, Moore SG, et al. LAMMPS-a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. Comput Phys Commun. 2022;271:108171.

[69] Rasmussen CE, Williams CK, et al. Gaussian processes for machine learning. vol. 1. Springer; 2006.

[70] Biau G, Scornet E. A random forest guided tour. Test. 2016;25(2):197–227.

[71] Hatos A, Hajdu-Soltész B, Monzon AM, Palopoli N, Álvarez L, Aykac-Fas B, Bassot C, Benítez GI, Bevilacqua M, et al. DisProt: intrinsic protein disorder annotation in 2020. Nucleic Acids Res. 2020;48(D1):D269–D276.

[72] Zhang Y, Otani A, Maginn EJ. Reliable viscosity calculation from equilibrium molecular dynamics simulations: A time decomposition method. J Chem Theory Comput. 2015;11(8):3537–3546.

SUPPLEMENTARY INFORMATION FOR "ACTIVE LEARNING OF THE THERMODYNAMICS–DYNAMICS TRADEOFF IN PROTEIN CONDENSATES"

V. EXTENDED METHODS

A. Featurization of Protein Sequences

A 30-dimensional feature vector $\vec{x}$ is used as a numerical encoding of a polypeptide sequence and its chemical characteristics. To construct this feature vector, 30 different characteristics of each protein sequence are considered, including

- the composition of each amino acid in the sequence, c_a, where a corresponds to one of the amino acids (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y) and $\sum_a c_a = 1$;
- the sequence length (i.e., number of amino acids), N;
- the net charge per residue, $|\bar{q}|$,

$$|\bar{q}| \equiv \frac{1}{N}\left|\sum_{i=1}^{N} q_i\right|;$$

- the sequence charge decoration, SCD,

$$\mathrm{SCD} \equiv \frac{1}{N}\sum_{i=1}^{N}\sum_{j=i+1}^{N} q_i q_j (j-i)^{1/2};$$

- the average hydrophobicity per residue, $\bar{\lambda}$,

$$\bar{\lambda} \equiv \frac{1}{N}\sum_{i=1}^{N} \lambda_i;$$

- the sequence hydropathy decoration, SHD,

$$\mathrm{SHD} \equiv \frac{1}{N}\sum_{i=1}^{N}\sum_{j=i+1}^{N} (\lambda_i + \lambda_j)(j-i)^{-1};$$

- a mean-field prediction of the second-virial coefficient, $B_2^{(\mathrm{MF})}$,

$$B_2^{(\mathrm{MF})} \equiv \sum_{i}^{N}\sum_{j}^{N} b_{2,ij},$$

  where $b_{2,ij} \equiv 2\pi \int_0^\infty dr\, r^2 \left[1 - e^{-\beta u_{ij}(r)}\right]$ is the second-virial coefficient between monomers i and j, which are assumed to be on different chains, and $u_{ij}(r)$ is the pair potential between monomers i and j;
- the fraction of positively charged residues, $\bar{q}_+$;
- the fraction of negatively charged residues, $\bar{q}_-$;
- the Shannon entropy, S,

$$S \equiv -\sum_a c_a \log c_a;$$

- and the average molar mass of a residue, $\bar{M}$,

$$\bar{M} \equiv \frac{1}{N}\sum_{i=1}^{N} M_i,$$

  where M_i is the molecular weight of the ith amino acid in the sequence.

For the purpose of using these sequence characteristics as features in machine learning and optimization, it is useful to transform the variables such that they have similar magnitudes. Therefore, we apply normalization techniques to each of the last ten sequence characteristics (i.e., all features except the $\{c_a\}$, which are already on a unit scale). With the exception of N and S, standard normalization is used for each feature:

$$\tilde{A} = \frac{A - \mu_A}{\sigma_A}, \tag{11}$$

where A is a given sequence characteristic with an average of μ_A and a standard deviation of σ_A estimated over the sequences that comprise the current dataset $\mathcal{D}$ for training the machine-learning models. Min-max normalization is employed for N,

$$\tilde{N} = \frac{N - \min_{\mathcal{D}} N}{\max_{\mathcal{D}} N - \min_{\mathcal{D}} N}, \tag{12}$$

and normalization by the maximum is employed for S,

$$\tilde{S} = \frac{S}{\max_{\mathcal{D}} S} - 1. \tag{13}$$

We note that similar feature vectors have demonstrated excellent predictive capabilities relating to both structural and dynamic properties of intrinsically disordered protein chains [54, 55].

B. Gaussian process regression models

Gaussian process regression (GPR) is used to estimate expected values and uncertainties for the second-virial coefficient, B_2, and the condensed-phase self-diffusion coefficient, D, as a function of the feature vector, $\vec{x}$. GPR is particularly used for its native uncertainty estimates, which are central to the computation of expected hypervolume improvement (EHVI) acquisition function described in the main text, Eq. (14). Separate GPR models are constructed for B_2 and D. Covariances modeled by the Gaussian processes are calculated using a Matérn kernel ($\nu = 3/2$) with added noise,

$$k(\vec{x}, \vec{x}') = \sigma^2 \left[\frac{1}{\Gamma(3/2)2^{1/2}} \times \left(\frac{\sqrt{3}}{l} ||\vec{x} - \vec{x}'||_2 \right)^{3/2} K_{3/2} \left(\frac{\sqrt{3}}{l} ||\vec{x} - \vec{x}'||_2 \right) \right] + \sigma_n^2. \tag{14}$$

In Eq. (14), σ, l, and σ_n are treated as adjustable hyperparameters, $\Gamma(\cdot)$ denotes the gamma function, and $K_{3/2}$ is a modified Bessel function. Hyperparamater optimization is predicated on minimization of the mean-squared errors. Five-fold cross-validation is employed on the training set to mitigate overfitting and obtain optimal hyperparameters. In particular, five sets of optimized hyperparameters are produced (one for each fold), and these sets are then averaged to yield a final set of parameters. Using this final set, a GPR model is trained over the entirety of the data acquired up to that point for the purpose of surrogate modeling during sequence optimization. The performance of the GPR models at the conclusion of active learning for an 80/20 train/test split is shown in Fig. S10a,b and illustrates good predictive ability.

C. Random forest classifier construction

A random forest (RF) classifier is used to predict whether phase separation is expected to occur as a function of the feature vector, $\vec{x}$. Here, the RF classifier consists of 100 decision trees. The best split in decision trees is determined by the Gini impurity. The minimum number of samples required to split an internal node is set at two, and the minimum number of samples required to be a leaf node is one. The dataset was split to 80/20 for training and testing. Five-fold cross-validation is employed on the training set to mitigate overfitting and to obtain optimal hyperparameters in the same fashion as described in SI Sec. V B. The performance of the RF classifier at the conclusion of active learning for is shown as a confusion matrix in Fig. S10c.

D. Sequence optimization and genetic algorithm

Sequence optimization to identify candidate proteins with high EHVI is facilitated using a genetic algorithm. In addition to conventional mutation and crossover moves, we also employ “deletion” and “growth” moves to improve

diversity of proposed sequences. All moves are constrained as necessary to produce sequences with $20 \leq N \leq 50$. In all discussion, sequences of N residues are index from 1 to N. In each iteration, batches of 96 candidate sequences are produced as described below.

- *Step 1*: A set of n possible parent sequences sorted by their fitness from high to low. In Iteration 0, the parent sequences are sequences from DisProt that exhibit negative B_2. At the beginning of every other iteration, the parent sequences correspond to those sequences that define the current approximation to the Pareto front.
- *Step 2*: To generate "child" sequences, a pair of sequences are randomly selected from the top 30% of parent sequences. These sequences then undergo a series of crossover, mutation, growth and deletion moves:
  - *Crossover*: Suppose the pair of selected sequences have a length of N and M. First, two indices are selected s_1 and s_2 both from the set $\in [1, \min(\{N, M\})$ with uniform probability. Then, another index is selected $s_3 \in [1, \max(\{N, M\}) - (s_2 - s_1)]$. Finally, the sub-sequence with the indices $\in [s_1, s_2]$ from the smaller sequence is exchanged with the sub-sequence with indices $\in [s_3, s_3 + (s_2 - s)1]$ from the larger sequence, accounting for changes to bond connectivity as needed.
  - *Deletion*: For a given sequence of length N, the length of a sub-sequence for deletion $l_{\text{del}} \in [0, N - 20]$ is selected with uniform random probability. Next, an index s is chosen from the set $[0, N - l_{\text{del}}]$. Then, the amino acids with indices in the sequence $\in [s, s + l_{\text{del}} - 1]$ are removed from the chain. If necessary, bonds are established between the amino acids with indices of $s - 1$ and $s + l_{\text{del}}$.
  - *Growth*: For a given sequence, the length of a sub-sequence for addition $l_{\text{gro}} \in [0, 50 - N]$ is selected with uniform random probability. Next, a random position s in the sequence is selected with uniform probability. Then, the sub-sequence with indices $\in [1, l_{\text{gro}}]$ is replicated and inserted at s. As necessary, the bond between amino acids at s and $s+1$ is eliminated, the amino acid at s is bonded to the first amino acid in the replicated sequence, and the amino acid formerly at $s + 1$ is bonded the last amino acid in the replicated sequence.
  - *Mutation*: For a given sequence of length N, an index $s \in [1, N]$ is selected with uniform random probability. Then, the amino acid at position s is exchanged for another amino acid, which is selected from the pool of twenty amino acids with uniform random probability.

  The probabilities for crossover, deletion, growth, and mutation are 0.5, 0.2, 0.5, and 0.5, respectively. The moves are executed independently in the order of crossover, deletion, growth, and mutation. This procedure generates two new "child" sequences.
- *Step 3*: Repeat *Step 2* until $\lfloor 0.7n \rfloor$ new sequences have been produced. These newly generated child sequences are combined with the top 30% of the parent sequences from the prior generation to comprise a new set of n parent sequences, which represent a new generation.
- *Step 4*: The fitness function is evaluated over the new generation of prospective parent sequences. The sequences are ranked by fitness.
- *Step 5*: Repeat starting from Step 2 until 100 generations have been produced.
- *Step 6:* The sequence with the best fitness at the conclusion of all generations is identified and proposed as a candidate for simulation as part of the active learning iteration.

  Steps 1-6 above results in one candidate sequence. To facilitate convergence towards an optimal sequence, *Steps 1-6* are executed 96 times in parallel (see Fig. S13) without communication. The resulting 96 sequences, which differ by the stochastic nature of the genetic algorithm, are ranked by their fitness, and the one with the best overall fitness is marked for characterization by simulation. This overall workflow is then executed again (going back to *Step 1*) 95 additional times to obtain a total of 96 distinct sequences for characterization by simulation. Note that the fitness function changes based on the active learning iteration and the prior history of proposed sequences (see Eqn. 7 of the main text). Although a genetic algorithm was used for this task, any reasonable optimization algorithm should be able to achieve similar results. In addition, different move probabilities or move sets may facilitate better convergence but were not explored here.

E. Determination of Counterfactuals

The task of determining a counterfactual for a sequence requires identifying another sequence that possesses overall similar characteristics but exhibits a different classification. In the context of this study, the classifications correspond

to *Pareto-optimal* versus *near-Pareto-optimal*. To treat our two objective properties (D and B_2) on equal footing, we consider the classification of near-Pareto-optimal to be based on a distance determined according to standard-normalized variables $\tilde{D}$ and $\tilde{B}_2$. Each sequence i can then be represented as a coordinate in the standard-normalized B_2–D plane as $\vec{z}^{(i)} = (\tilde{B}_2^{(i)}, \tilde{D}^{(i)})^\intercal$. To account for statistical uncertainties associated with the quantities obtained from simulations, we associated near-Pareto-optimal with distances $d \in [0.15, 0.3]$, which are measured from a given coordinate to a piecewise-linear approximation to the Pareto front. In particular, we first calculate three sets of 35 distances ($\{d_{v,i}\}$, $\{d_{1,i}\}$ and $\{d_{2,i}\}$) associated with our 35 Pareto-optimal sequences with coordinates $\vec{z}^{(\mathrm{P}_1)}, \ldots, \vec{z}^{(\mathrm{P}_{35})}$ as below:

$$d_{v,i} = \frac{(\vec{z}^{(k)} - \vec{z}^{(\mathrm{P}_i)}) \times (\vec{z}^{(\mathrm{P}_i)} - \vec{z}^{(\mathrm{P}_{i+1})})}{|(\vec{z}^{(\mathrm{P}_i)} - \vec{z}^{(\mathrm{P}_{i+1})}|}, \tag{15}$$

$$d_{1,i} = |\vec{z}^{(k)} - \vec{z}^{(\mathrm{P}_i)}|, \tag{16}$$

and

$$d_{2,i} = |\vec{z}^{(k)} - \vec{z}^{(\mathrm{P}_{i+1})}|, \tag{17}$$

where $d_{v,i}$ is the distance between the point $\vec{z}^{(k)}$ to the piecewise lines determined by two adjacent Pareto-optimal points $\vec{z}^{(\mathrm{P}_i)}$ and $\vec{z}^{(\mathrm{P}_{i+1})}$, $d_{1,i}$ is the distance between $\vec{z}^{(k)}$ and $\vec{z}^{(\mathrm{P}_i)}$, and $d_{2,i}$ is the distance between $\vec{z}^{(k)}$ and $\vec{z}^{(\mathrm{P}_{i+1})}$; for the above $i \in [1, 34]$.

A series of distances $d_k^{(i)}$ for $\vec{z}^{(k)}$ are then determined as

$$d_k^{(i)} = \begin{cases} d_{v,i} & \text{if } \cos\theta_{1,i} > 0, \text{ and } \cos\theta_{2,i} < 0 \\ d_{1,i} & \text{if } \cos\theta_{1,i} < 0, \text{ and } \cos\theta_{2,i} < 0 \\ d_{2,i} & \text{if } \cos\theta_{1,i} > 0, \text{ and } \cos\theta_{2,i} > 0, \end{cases} \tag{18}$$

where $\theta_{1,i}$ is the angle between $(\vec{z}^{(k)} - \vec{z}^{(\mathrm{P}_{i+1})})$ and $(\vec{z}^{(\mathrm{P}_i)} - \vec{z}^{(\mathrm{P}_{i+1})})$, and $\theta_{2,i}$ is the angle between $(\vec{z}^{(k)} - \vec{z}^{(\mathrm{P}_i)})$ and $(\vec{z}^{(\mathrm{P}_i)} - \vec{z}^{(\mathrm{P}_{i+1})})$:

$$\cos\theta_{1,i} = \frac{(\vec{z}^{(k)} - \vec{z}^{(\mathrm{P}_{i+1})}) \cdot (\vec{z}^{(\mathrm{P}_i)} - \vec{z}^{(\mathrm{P}_{i+1})})}{|(\vec{z}^{(k)} - \vec{z}^{(\mathrm{P}_{i+1})})||(\vec{z}^{(\mathrm{P}_i)} - \vec{z}^{(\mathrm{P}_{i+1})})|}, \tag{19}$$

$$\cos\theta_{1,i} = \frac{(\vec{z}^{(k)} - \vec{z}^{(\mathrm{P}_i)}) \cdot (\vec{z}^{(\mathrm{P}_i)} - \vec{z}^{(\mathrm{P}_{i+1})})}{|(\vec{z}^{(k)} - \vec{z}^{(\mathrm{P}_i)})||(\vec{z}^{(\mathrm{P}_i)} - \vec{z}^{(\mathrm{P}_{i+1})})|}. \tag{20}$$

Finally, the shortest distance of $z^{(k)}$ to the Pareto front is set as

$$d_k = \min(\{d_k^{(i)}\}). \tag{21}$$

VI. SUPPLEMENTARY FIGURES

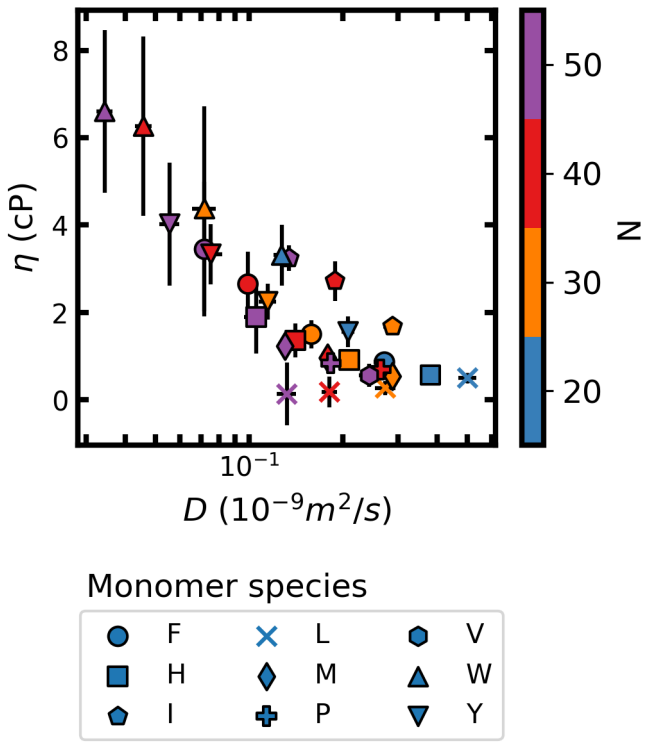

FIG. 5. **Relationship of viscosity and self-diffusivity in the condensed-phase of homomeric polypeptides.** The viscosity, η, is anticorrelated with self-diffusivity, D, although the former possesses larger statistical uncertainties. For all systems, quantities are measured in the canonical ensemble at the estimated coexistence density of the condensed phase as determined by the approximate EOS method (see main manuscript). The viscosity is determined using the Green–Kubo formalism [72]

.

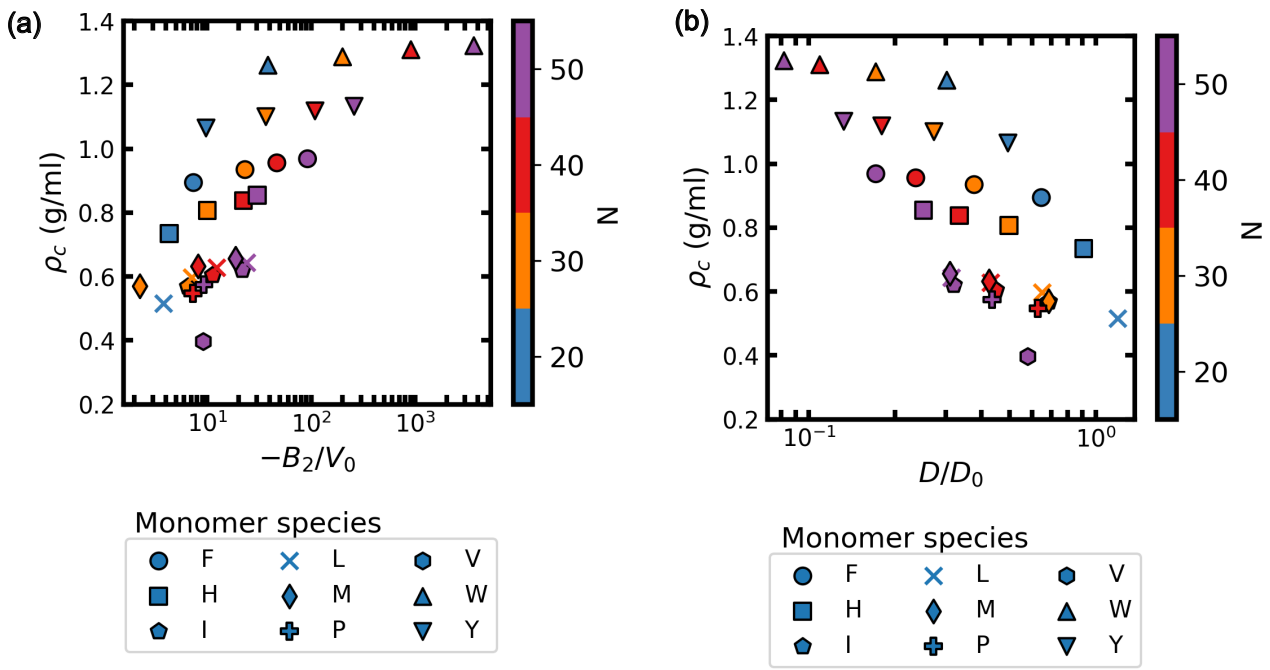

FIG. 6. **Correlation of condensed-phase densities with target properties in phase-separating homomeric polypeptides.** Correlation plots for condensed-phase density, ρ_c with **(a)** the dimensionless second-virial coefficient, $-B_2/V_0$, and **(b)** the dimensionless self-diffusion coefficient in the condensed-phase, D/D_0.

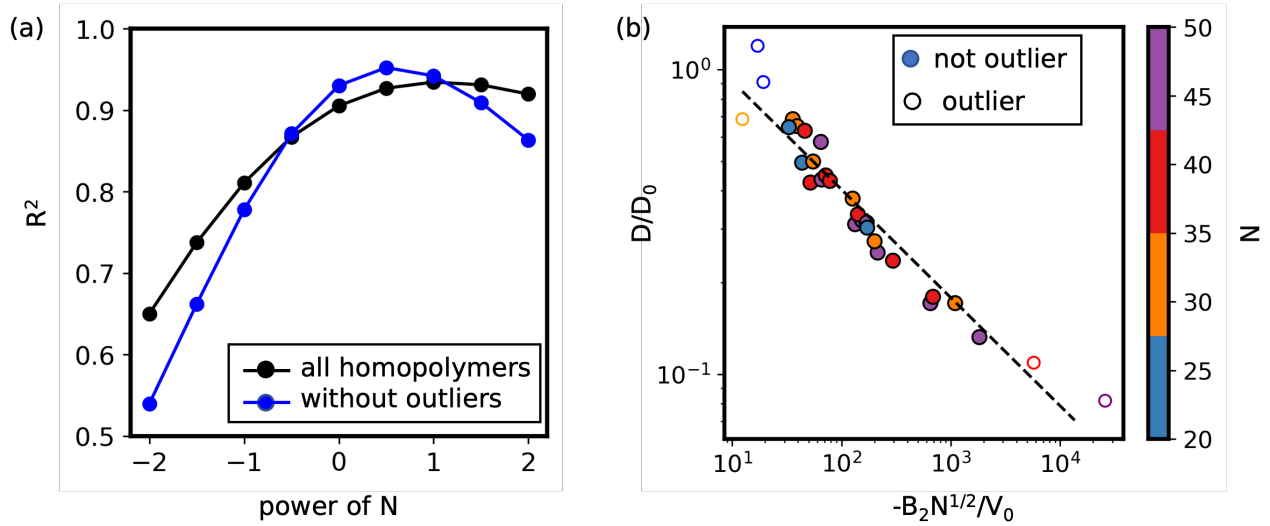

FIG. 7. **Examination of thermodynamics-dynamics tradeoff for homomeric polypeptides. (a)** Coefficient of determination for correlation between dynamics via diffusion coefficient ($\log[D/D_0]$) versus a measure of thermodynamic attraction via $\log[-B_2N^{\nu}/V_0]$ where B_2 is the second virial coefficient, N is the sequence length, and ν is a power-law scaling exponent. The highest correlation is for $\nu = 1/2$ or $\nu \approx 1$. **(b)** Illustration of data collapse with $\nu = 1/2$. In both panels, data points are considered "outliers" if $-B_2 > 10^{6.5}$ or $-B_2 < 10^{4.5}$.

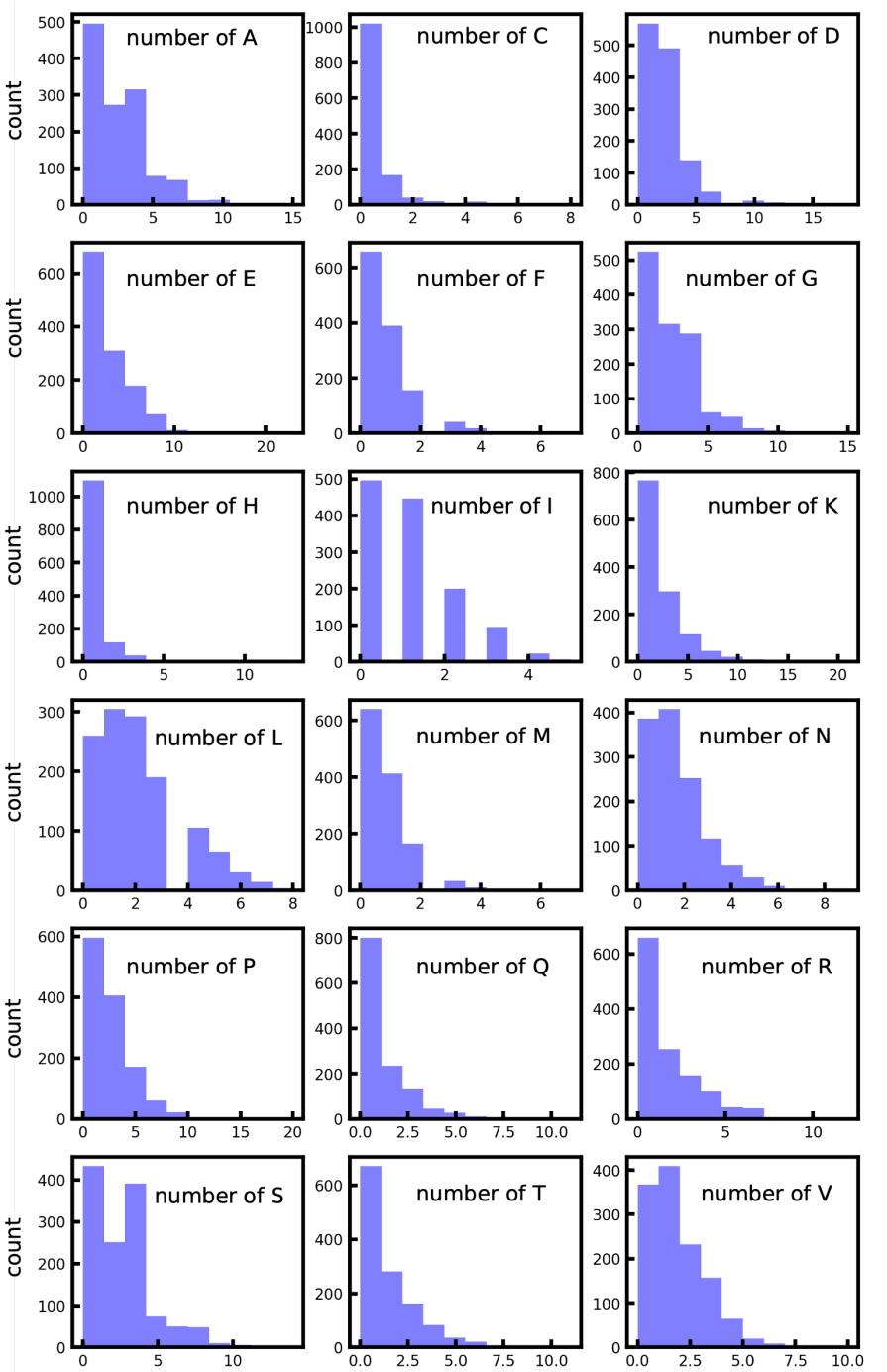

FIG. 8. **Distribution of values for features among sequences selected from the DisProt database.** Panels show the number of each amino-acid type (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, or Y) and the ten sequence characteristics described in SI Sec. V A. Sequences possess lengths of $20 \leq N \leq 50$.

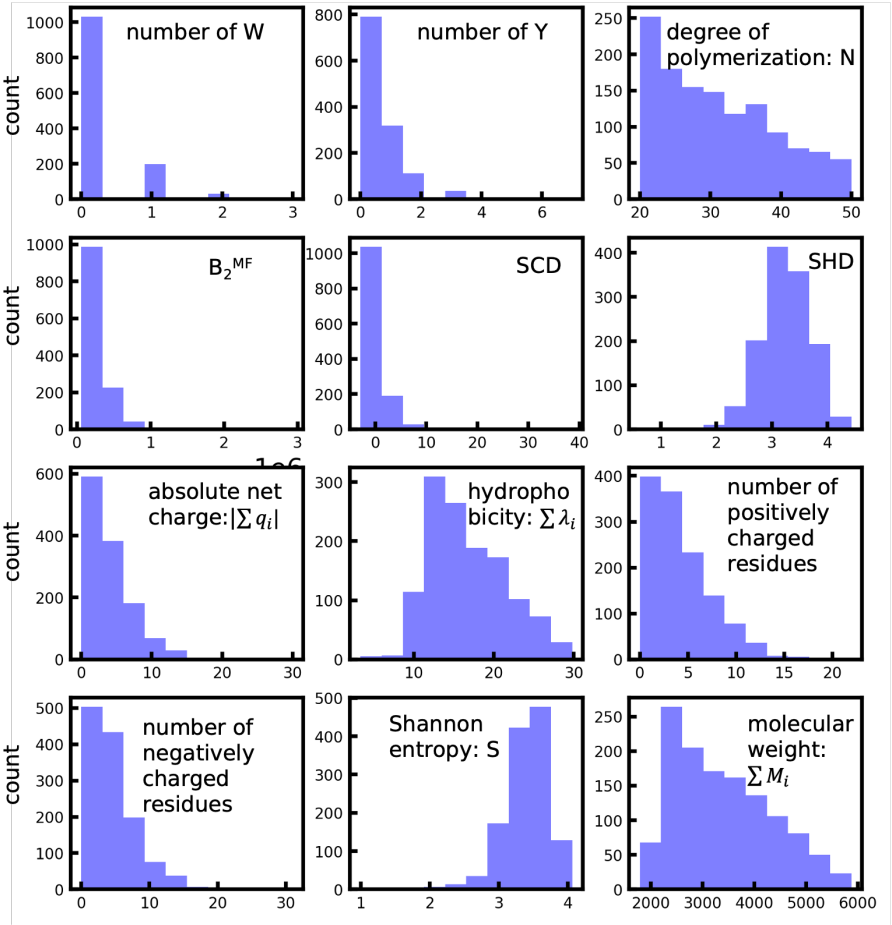

FIG. S8 (Cont.). **Distribution of values for features among sequences selected from the DisProt database.** Panels show the number of each amino-acid type (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, or Y) and the ten sequence characteristics described in SI Sec. V A. Sequences possess lengths of $20 \leq N \leq 50$.

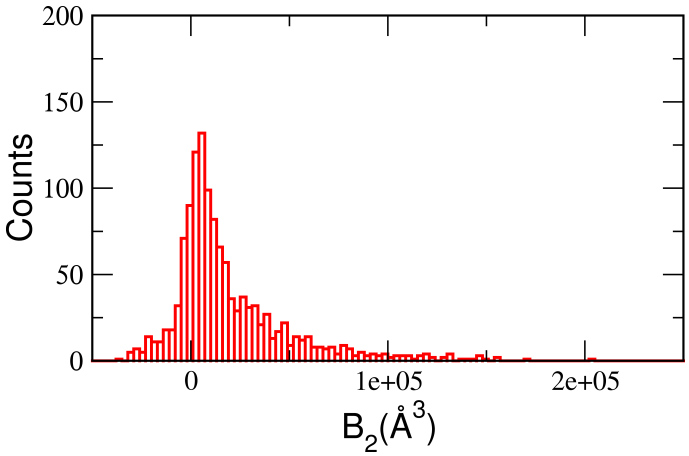

FIG. S9. **Distribution of calculated second virial coefficients among sequences selected from the DisProt database.** Sequences possess lengths of $20 \leq N \leq 50$. The majority of sequences exhibit $B_2 > 0$.

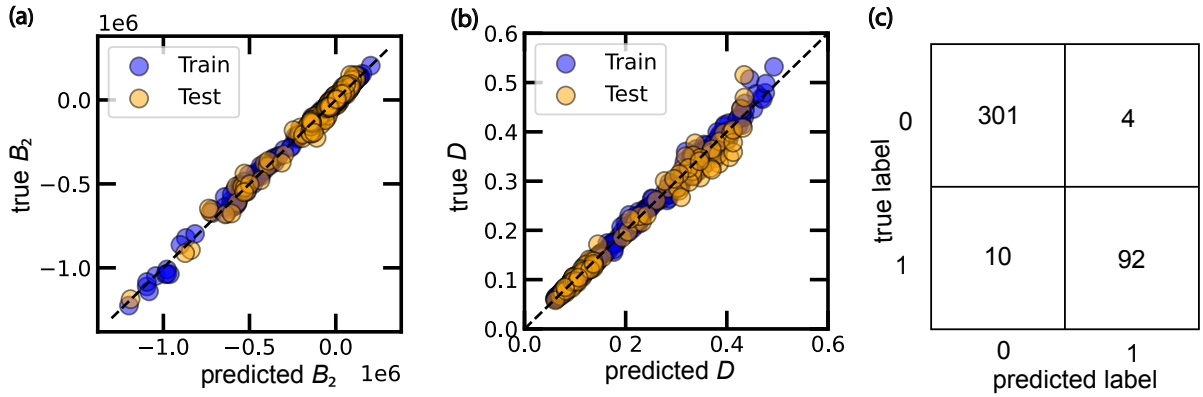

FIG. S10. **Illustration of the quality of machine-learning models.** **a** Parity plot for the Gaussian process regression model that predicts the second-virial coefficient, B_2. **(b)** Parity plot for the Gaussian process regression model that predicts the condensed-phase self-diffusion coefficient, D. **(c)** Confusion matrix (test set) for classification of phase-separation behavior. A value of '0' indicates no phase separation and '1' indicates phase separation. The squares on the diagonal report the number of true classifications while the off-diagonal squares report the number of misclassifications. In all panels, the train-test split is 80% for training and 20% for testing with random selection. In **(a)** and **(c)**, all 2034 heteromeric sequences are included. In **(b)**, only the subset of the 2034 heteromeric sequences for which condensed-phase densities could be extracted were included; this includes 508 sequences.

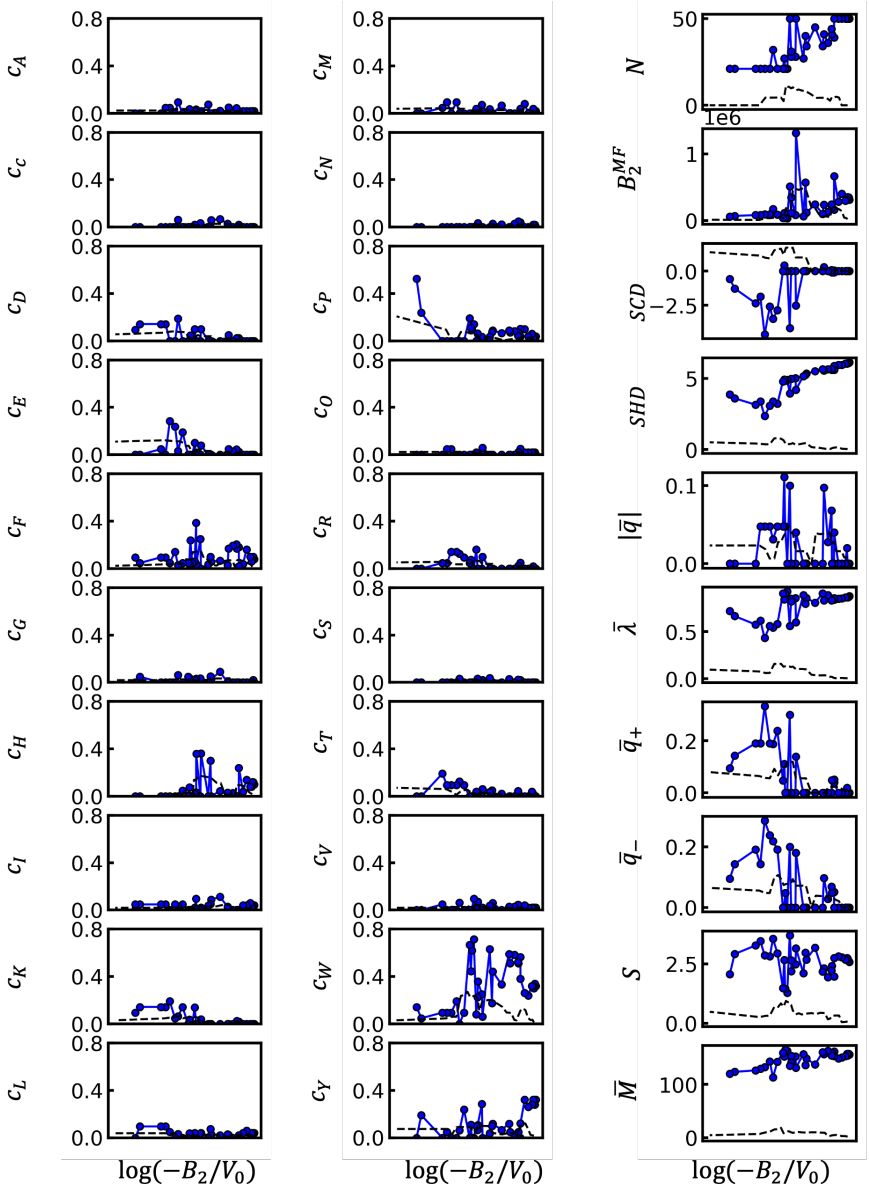

FIG. S11. **Variation of feature values across the Pareto front.** Each marker indicates the value of a specific feature for a Pareto-optimal sequence. The data are presented such that both the diffusion coefficient and $-B_2$ are decreased moving left-to-right. The dashed lines indicate a rolling standard deviation of the feature x (i.e., $\sqrt{\frac{1}{m}\sum_{i=1}^{m}(x_i-\bar{x})^2}$, where $\vec{x}$ is the feature vector, the bin size $m=5$, and $\bar{x} \equiv \frac{1}{m}\sum_{i=1}^{m} x_i$) as measure of the extent of variability of a feature in particular regions of the Pareto front.

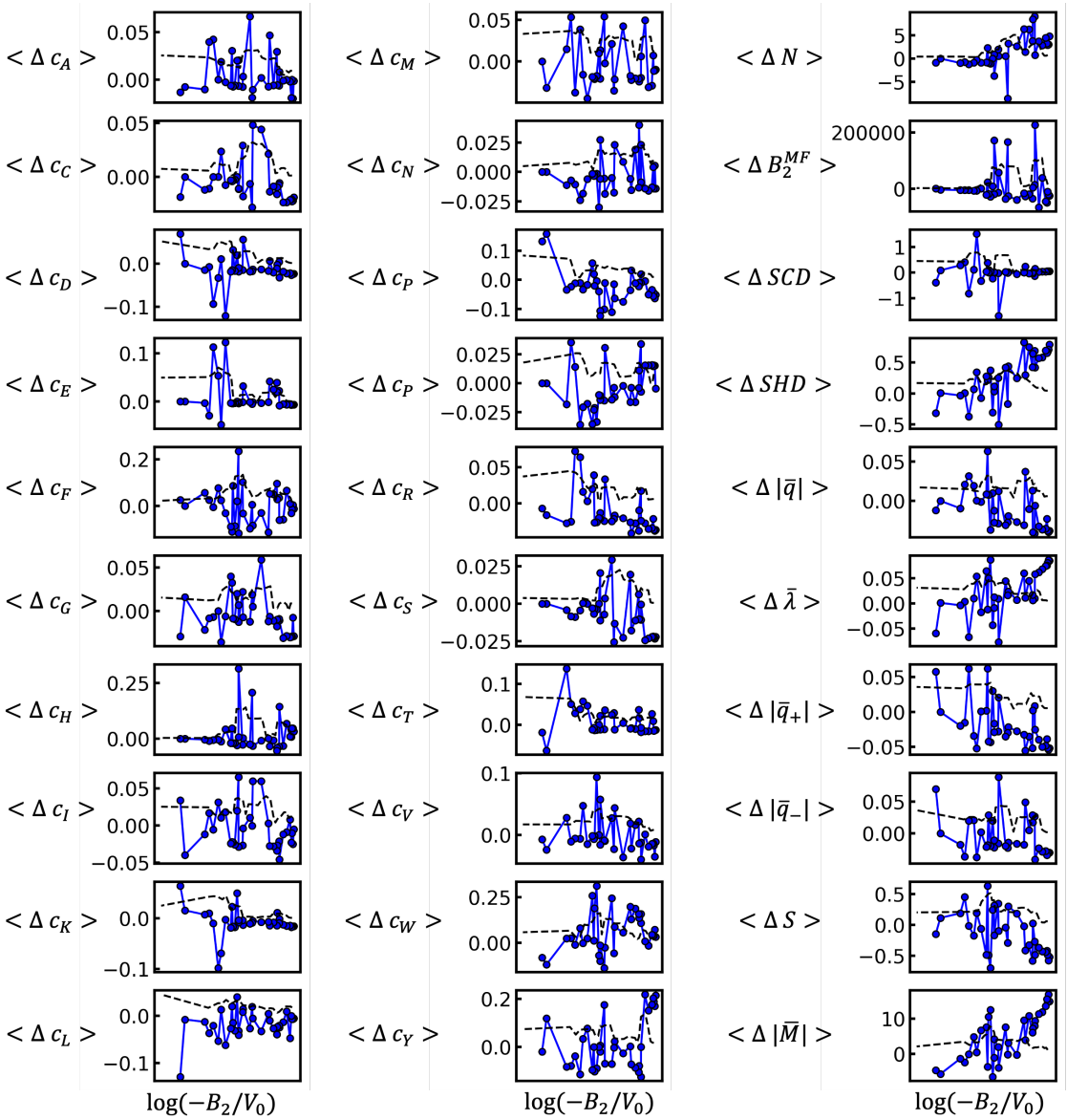

FIG. S12. **Average differences in feature values between Pareto-optimal sequences and their counterfactuals.** Average feature differences determined via Eq. (2) of the main text. The dashed lines indicate a rolling standard deviation as a measure of the extent of variability of the feature differences in particular regions of the Pareto front (see SI Fig. S11).

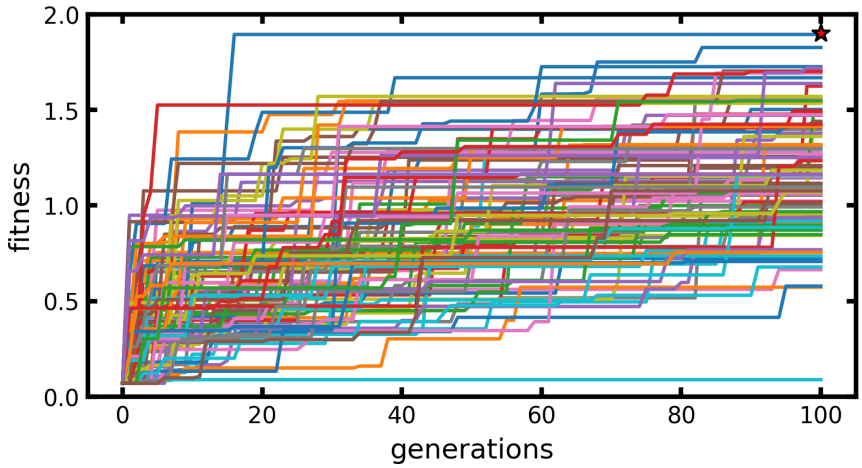

FIG. S13. **Illustration of convergence toward an optimal sequence using parallel execution of the genetic algorithm.** Each solid line indicates the current highest fitness of a sequence produced over the course of 100 generations for a single execution of the genetic algorithm (*Steps 1- 6* in Sec. V D). There are 96 lines corresponding to 96 independent executions with the same set of possible parent sequences. The different executions lead to a range of sequences with different fitness scores. The single best sequence (red star) is proposed for characterization by molecular dynamics simulation. A given active learning iteration produces 96 sequences in this manner.