

Design of intrinsically disordered proteins that undergo phase transitions with lower critical solution temperatures

Cite as: APL Mater. 9, 021119 (2021); <https://doi.org/10.1063/5.0037438>

Submitted: 13 November 2020 • Accepted: 20 January 2021 • Published Online: 18 February 2021

 Xiangze Zeng,  Chengwen Liu, Martin J. Fossat, et al.

COLLECTIONS

Paper published as part of the special topic on [Discovering Patterns in Disorder: Machine Learning for Fluctuating Mesoscopic Materials](#)



View Online



Export Citation



CrossMark

ARTICLES YOU MAY BE INTERESTED IN

[Model for disordered proteins with strongly sequence-dependent liquid phase behavior](#)
The Journal of Chemical Physics **152**, 075101 (2020); <https://doi.org/10.1063/1.5141095>

[Phase separation vs aggregation behavior for model disordered proteins](#)
The Journal of Chemical Physics **155**, 125101 (2021); <https://doi.org/10.1063/5.0060046>

[Polyphasic linkage and the impact of ligand binding on the regulation of biomolecular condensates](#)
Biophysics Reviews **2**, 021302 (2021); <https://doi.org/10.1063/5.0050059>

APL Materials

SPECIAL TOPIC: Emerging Materials
for Spin-Charge Interconversion

Read Now!



Design of intrinsically disordered proteins that undergo phase transitions with lower critical solution temperatures

Cite as: APL Mater. 9, 021119 (2021); doi: 10.1063/5.0037438

Submitted: 13 November 2020 • Accepted: 20 January 2021 •

Published Online: 18 February 2021



Xiangze Zeng,¹ Chengwen Liu,² Martin J. Fossat,¹ Pengyu Ren,² Ashutosh Chilkoti,³ and Rohit V. Pappu^{1,a)}

AFFILIATIONS

¹Department of Biomedical Engineering and Center for Science and Engineering of Living Systems (CELS), Washington University in St. Louis, St. Louis, Missouri 63130, USA

²Department of Biomedical Engineering, The University of Texas at Austin, Austin, Texas 78712, USA

³Department of Biomedical Engineering, Duke University, Durham, North Carolina 27708, USA

Note: This paper is part of the Special Topic on Discovering Patterns in Disorder: Machine Learning for Fluctuating Mesoscopic Materials.

a) Author to whom correspondence should be addressed: pappu@wustl.edu

ABSTRACT

Many naturally occurring elastomers are intrinsically disordered proteins (IDPs) built up of repeating units, and they can demonstrate two types of thermoresponsive phase behavior. Systems characterized by lower critical solution temperatures (LCSTs) undergo phase separation above the LCST, whereas systems characterized by upper critical solution temperatures (UCSTs) undergo phase separation below the UCST. There is congruence between thermoresponsive coil-globule transitions and phase behavior, whereby the theta temperatures above or below which the IDPs transition from coils to globules serve as useful proxies for the LCST/UCST values. This implies that one can design sequences with desired values for the theta temperature with either increasing or decreasing radii of gyration above the theta temperature. Here, we show that the Monte Carlo simulations performed in the so-called intrinsic solvation (IS) limit version of the temperature dependent self-Assembly of Biomolecules Studied by an Implicit, Novel, and Tunable Hamiltonian (ABSINTH) implicit solvation model yield a useful heuristic for discriminating between sequences with known LCST and UCST phase behavior. Accordingly, we use this heuristic in a supervised approach, integrate it with a genetic algorithm, combine this with IS limit simulations, and demonstrate that novel sequences can be designed with LCST phase behavior. These calculations are aided by direct estimates of temperature dependent free energies of solvation for model compounds that are derived using the polarizable atomic multipole optimized energetics for biomolecular applications forcefield. To demonstrate the validity of our designs, we calculate coil-globule transition profiles using the full ABSINTH model and combine these with Gaussian cluster theory calculations to establish the LCST phase behavior of designed IDPs.

© 2021 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>). <https://doi.org/10.1063/5.0037438>

INTRODUCTION

Intrinsically disordered proteins (IDPs) that undergo thermoresponsive phase transitions are the basis of many naturally occurring elastomeric materials.¹ These naturally occurring scaffold IDPs² serve as the basis of ongoing efforts to design thermoresponsive materials.³ Well-known examples of disordered regions derived from elastomeric proteins⁴ include the repetitive sequences from

proteins such as resilins,⁵ elastins,⁶ proteins from spider silks,⁷ titin,⁸ and neurofilament sidearms.⁹ Elastin-like polypeptides have served as the benchmark systems for the development of responsive disordered proteins that can be adapted for use in various biotechnology settings.¹⁰ The multi-way interplay of sequence-encoded intermolecular interactions, chain-solvent interactions, and chain and solvent entropy gives rise to thermoresponsive phase transitions that lead to the formation of coacervates.¹ Here, we show that one can

expand the “materials genome”¹¹ through *de novo* design strategies that are based on heuristics anchored in the physics of thermoresponsive transitions and efficient simulation engines that apply the learned heuristics in a supervised approach. We report the development of a genetic algorithm (GA) and show how it can be applied in conjunction with multiscale computations to design thermoresponsive IDPs with lower critical solution temperature (LCST) phase behavior.

Conformational heterogeneity is a defining hallmark of IDPs.¹² Work over the past decade-and-a-half has shown that naturally occurring IDPs come in distinct sequence flavors.¹² Indeed, IDPs can be distinguished based on their sequence-encoded interplay between intramolecular and chain-solvent interactions that can be altered through changes in solution conditions. Recent studies have shown that IDPs can be drivers or regulators of reversible phase transitions in simple and complex mixtures of protein and nucleic acid molecules.¹³ These transitions are driven primarily by the multivalence of interaction motifs that engage in reversible physical cross-links.¹⁴ IDPs can serve as scaffolds for interaction motifs (stickers), interspersed by spacers. Alternatively, they can modulate multivalent interactions mediated by stickers that are situated on the surfaces¹⁵ of autonomously foldable protein domains.¹⁶

Thermoresponsive phase transitions arise either by increasing the solution temperature above a lower critical solution temperature (LCST) or by lowering the temperature below an upper critical solution temperature (UCST).¹ Many systems are capable of both types of thermoresponsive transitions, although only one of the transitions might be accessible in the temperature range of interest. Here, we leverage our working knowledge of the sequence features that encode driving forces for thermoresponsive phase transitions¹⁷ to develop and deploy a GA for the design of novel IDPs characterized by LCST behavior. Inspired by work on elastin-like polypeptides,³ we focus on designing IDPs that are repeats of pentapeptide motifs. The amino acid composition of each motif contributes to the LCST behavior, and the number of repeats determines the multivalence of stickers that drive phase transitions with LCST behavior.

The GA we adopt for this work is driven by advances that include the following: (a) an improved fundamental understanding of the physics of LCST phase behavior;¹⁸ (b) experiments showing that many IDPs undergo collapse transitions with increased temperature;¹⁹ (c) a generalization of the ABSINTH implicit solvation model and forcefield paradigm²⁰ to account for the temperature dependence of chain solvation; (d) a growing corpus of information regarding the sequence determinants of LCST phase behavior in repetitive IDPs;¹⁷ and (e) the prior demonstration that a GA-based method known as Genetic Algorithm for Design of Intrinsic secondary Structure (GADIS)²¹ can be combined with efficient, ABSINTH-based simulations to design IDPs with bespoke secondary structural preferences.

Studies of synthetic polymer systems have helped in elucidating the origins of the driving forces and the mechanisms of LCST phase behavior.²² A well-known example is poly-*N*-isopropylacrylamide (PNIPAM).²³ Here, the dispersed single phase is stabilized at temperatures below ~32 °C by the favorable hydration of amides in the side chains. Solvation of amides requires that the solvent be organized around the hydrophobic moieties that include the backbone carbon chain and the isopropyl group in the side chain. The entropic cost for organizing solvent molecules around individual chains

increases with the increasing temperature. Accordingly, above the LCST of ~32 °C and for volume fractions that are greater than a threshold value, the system phase separates to form a polymer-rich coacervate phase that coexists with a polymer-poor dilute phase. The driving forces for phase separation are the gain in solvent entropy through the release of solvent molecules from the polymer and the gain of favorable inter-chain interactions, such as hydrogen-bonding interactions between amides in the polymer.

Tanaka and co-workers developed a *cooperative hydration* approach, inspired by the Zimm and Bragg theories for helix-coil transitions,²⁴ to model the physics of phase transitions with LCST.²⁵ Cooperative hydration refers to the cooperative association (below the LCST) or dissociation (above the LCST) of water molecules that are bound to repeating units along the polymer chain.²⁶ Cooperativity is captured using the Zimm-Bragg formalism by modeling each repeating unit as being in one of two states, viz., solvated or desolvated. In the solvated state, the repeating unit has a defined interaction strength with solvent molecules. In the desolvated state, pairs of such repeating units have defined exchange interactions. In addition, desolvation is associated with a gain in solvent entropy. The three-way interplay of direct solvent-chain interactions, the interactions among desolvated pairs of units, and the gain in solvent entropy above the LCST can be captured in a suitable physical framework that can be parameterized to describe system-specific phase transitions. Accordingly, if one has prior knowledge of the interaction energies associated with each repeat unit, one can use the framework of Tanaka and co-workers to design novel sequences with LCST behavior.

An alternative approach, which we adopt in this work, is to leverage the corollary of LCST behavior at the single chain limit.²⁷ At temperatures that are proximal to the LCST, the chain of interest will undergo a coil-to-globule transition in a dilute solution.²⁸ This is because chain collapse is a manifestation of the physics of phase separation in the single chain limit. Here, we leverage this connection between phase separation and chain collapse of isolated polymer chains in ultra-dilute solutions to design novel IDPs that are predicted to undergo phase transitions with LCST phase behavior. We do so by using a multi-pronged approach that starts with improved estimates of the temperature dependencies of free energies of solvation of model compounds that mimic the amino acid side chain and backbone moieties. For this, we use free energy calculations based on the atomic multipole optimized energetics for biomolecular applications (AMOEBA) forcefield,²⁹ which is a second-generation, molecular mechanics based, polarizable model for water molecules and proteins. We incorporate these temperature dependent free energies of solvation into the ABSINTH implicit solvation model and show that thermoresponsive changes to chain dimensions, calculated in the “intrinsic solvation (IS) limit,”³⁰ yield robust heuristics that discriminate sequences with known LCST phase behavior from those that show UCST behavior. We then describe the development of a GA, an adaptation of the GADIS approach, to design novel sequences that rely on all-atom simulations, performed using the ABSINTH model in the IS limit, and learned heuristics as fitness scores. Distinct classes of designed sequences emerge from our approach, and these are screened to filter out sequences with low disorder scores as assessed using the IUPRED2 algorithm.³¹ The resulting set of sequences are analyzed using simulations based on the full ABSINTH model, which show that the designed sequences

do undergo collapse transitions above a threshold temperature. The contraction ratio, defined as the ratio of chain dimensions at temperature T to the dimensions at the theta temperature, computed as a function of simulation temperature, is analyzed to extract temperature dependent two-body interaction parameters and athermal three-body interaction parameters that are used in conjunction with the Gaussian Cluster Theory (GCT)³² to calculate system-specific phase diagrams.²⁸ The upshot is a multiscale pipeline, whereby a GA, aided by a derived heuristic and IS limit simulations, leads to the design of novel sequences with predicted LCST phase behavior. Following a post-processing step that selects for sequences with a high confidence of being intrinsically disordered, we combine all-atom ABSINTH-T based simulations with Gaussian cluster theory to obtain sequence-specific phase diagrams. These last two steps allow further pruning of the sequence space derived from the designs and provide further confidence regarding the authenticity of the predicted LCST phase behavior.

Temperature dependent free energies of solvation are central to accurate descriptions of LCST behavior. Each protein may be viewed as a chain of model compounds, and measured/calculated temperature dependent values of temperature dependent free energies of hydration $\Delta\mu_h$ for fully solvated model compounds can be used as the reference free energies of solvation (rFoS) in implicit solvation models such as EEF1³³ or ABSINTH.²⁰ Where possible, the ABSINTH model^{20,34} uses experimentally measured free energies of solvation for model compounds. In the original formalism, Vitalis and Pappu²⁰ adopted experimentally derived rFoS values at 298 K and assumed these values to be independent of temperature. This approach was generalized by Wuttke *et al.*,¹⁹ to calculate temperature dependent rFoS values, using data from calorimetric measurements made by Makhataдзе and Privalov³⁵ for the enthalpy and heat capacity of hydration at a reference temperature. These values were augmented by those of Cabani *et al.*³⁶ for naphthalene, which is used as a model compound mimic of tryptophan. Wuttke *et al.*¹⁹ incorporated the enthalpy and heat capacity of hydration estimated at a reference temperature into an integrated version of the Gibbs–Helmholtz equation to yield a thermodynamic model for temperature dependent rFoS values for all the relevant model compounds. In this formalism, $rFoS(T)$ or $\Delta\mu_h(T)$ is written as

$$\Delta\mu_h(T) = \frac{[\Delta\mu_h(T_0) - \Delta h]T}{T_0} + \Delta h + \Delta c_p \left[T \left(1 - \ln \frac{T}{T_0} \right) - T_0 \right]. \quad (1)$$

Here, Δh is the enthalpy of solvation (hydration) at a reference temperature T_0 , which is typically set to be 298 K, and Δc_p is the molar heat capacity change associated with the solvation process. Based on measurements, the assumption is that Δc_p is independent of temperature.³⁷

We built on the approach of Wuttke *et al.*¹⁹ to incorporate temperature dependent rFoS values in ABSINTH. This is implemented in a version that we refer to as ABSINTH-T. The issues we faced in developing ABSINTH-T were two-fold. First, the values for Δc_p and Δh that were used by Wuttke *et al.* rely on decompositions of measurements for model compounds into group-specific contributions. In contrast, the $\Delta\mu_h$ values used in ABSINTH are for model compounds and explicitly avoid the group-specific decompositions made by Makhataдзе and Privalov.³⁵ This choice reflects the fact that group-specific decompositions are not measured. Instead, they

are derived quantities that are based on empirical reasoning. This creates a mismatch with the paradigm that underlies the ABSINTH framework.²⁰ Put simply, we require values of $\Delta\mu_h$, Δc_p , and Δh that correspond to model compounds as opposed to group-specific decompositions.

Second, model compounds that mimic the side chains of ionizable residues pose unique challenges. For any solute, including ions, the free energy of hydration at a specific temperature and pressure is defined as the change in free energy associated with transferring the solute of interest from a dilute vapor phase into water.³⁸ The accommodation of the solute into liquid water is associated with the cost to create a cavity in the solvent,³⁸ the electronegative cavity potential,³⁹ the work to add soft dispersion interactions,⁴⁰ and distribute charges uniformly or non-uniformly across the solute.⁴¹ Vapor pressure osmometry with radioactive labeling, as used by Wolfenden⁴² to measure free energies of hydration for polar solutes, including neutral forms of ionizable species, cannot be used to measure free energies of hydration of ions because of the ultra-low vapor pressures and the confounding effects of ion-pairing in the gas phase. Calorimetry, as used by Makhataдзе, Privalov, and colleagues, provides an alternative approach.^{35,43} However, the large magnitudes of free energies of hydration, which are expected to be on the order 10² kcal/mol,^{44,45} giving rise to even larger magnitudes for enthalpies of hydration, make it impossible to obtain the numbers of interest directly from calorimetric measurements. Measurements of activity coefficients as a function of the concentration of whole salts in aqueous solutions can be used to place bounds on the values of $\Delta\mu_h$,⁴⁶ but these are not direct measurements of $\Delta\mu_h$.

A key challenge is that stable solutions are electroneutral.⁴⁵ Accordingly, all measurements aimed at estimating the free energies of hydration of ionic species have to rely on parsing numbers derived from measurements on whole salts against those of reference salts⁴⁴—see the work of Grossfield *et al.*⁴¹ and references therein. Alternatives rely on referencing measurements for whole salts against the free energy of hydration of the proton⁴⁷—a quantity plagued by considerable uncertainty, given the interplay between the Zundel form and the Eigen form for the hydronium ion.⁴⁸ One can also use direct measurements of neutralized versions of ionic species;^{35,42,43} however, extracting the parameters of interest ends up relying on explicit or implicit assumptions regarding proton hydration free energies to extract estimates of the desired free energies of hydration of ionic species. The upshot is that direct measurements of free energies of hydration of ionic species are not feasible, and hence, one has to rely on the validity of models that are used to parse experimental data.

In 1996, in their work aimed at accounting for reaction-field effects in calculations of hydration free energies in continuum models, Marten *et al.*⁴⁹ compiled a set of values for hydration free energies for all the relevant model compounds. In the original ABSINTH model,²⁰ the values tabulated by Marten *et al.* were used for all uncharged solutes. For charged species, specifically the protonated versions of Arg and Lys side chains and deprotonated versions of Asp and Glu side chains, Vitalis and Pappu²⁰ used the numbers tabulated and parsed by Marcus⁵⁰ for a reference temperature of 298 K.

Wuttke *et al.*¹⁹ used the rFoS values tabulated by Vitalis and Pappu²⁰ at 298 K and tested three different models for generating T-dependent rFoS values of model compounds used to mimic

the charged versions of Arg, Asp, Lys, and Glu. Model 1 uses the measured enthalpies and heat capacities measured for the neutral compounds,³⁵ i.e., protonated Asp and Glu and deprotonated Lys and Arg. These were then scaled by the rFoS values used by Vitalis and Pappu²⁰ for the charged variants. The scaled enthalpies and heat capacities were then deployed in Eq. (1). Model 2 of Wuttke *et al.*¹⁹ uses the enthalpies of hydration estimated by Marcus⁵¹ and the heat capacities of hydration tabulated by Abraham and Marcus.⁵² As noted above, these numbers are not direct measurements. Instead, they were derived from measurements of whole salts and then parsed using different models to arrive at a consensus set of estimates for the enthalpies and heat capacities. Model 3 of Wuttke *et al.*¹⁹ uses the same heat capacities as model 2, and empirical choices were made for the enthalpies based on “expectations for a variety of charged model compounds.”

The preceding discussion emphasizes the fact that direct measurements of the rFoS values as a function of temperature or of the enthalpies and heat capacities of hydration at reference temperatures are unavailable for model compounds that mimic charged versions of the side chains of Arg, Asp, Lys, and Glu. To put the challenge into perspective, we note that models 1 and 2 of Wuttke *et al.*¹⁹ yield values of 50.37 cal mol⁻¹ K⁻¹ and 5.30 cal mol⁻¹ K⁻¹, respectively, for ΔC_p of the acetate ion. The large variations are a reflection of the challenges associated with estimating temperature-independent and temperature rFoS values for charged species.

Here, we pursue a different approach: we use AMOEBA, which is a second generation molecular mechanics based polarizable force-field,²⁹ in direct calculations of T-dependent rFoS values for all the relevant model compounds. The AMOEBA water model reproduces the temperature dependent anomalies of liquid water⁵³ and yields accurate free energies of solvation for model compounds in aqueous solvents.^{29,54,55} Our goal was to have a common source for T-dependent rFoS values of the key model compounds that are used in ABSINTH. The free energy calculations were performed at specific temperatures, and the integrated version of the Gibbs–Helmholtz equation was used to extract Δh and ΔC_p . The values of Δh and ΔC_p in conjunction with Eq. (1) are used to calculate T-dependent rFoS values in ABSINTH-T.

RESULTS AND DISCUSSION

Results from AMOEBA-based free energy calculations for model compounds: We performed temperature dependent free energy calculations based on the Bennett Acceptance Ratio (BAR) free energy estimator⁵⁶ for the direct investigation of how $\Delta\mu_h$ varies with temperature. These calculations were performed for 19 different model compounds that mimic the 20 side chain moieties and the backbone peptide unit. The details of the parameterization of the AMOEBA forcefield for model compounds used in this study and the design of the free energy calculations are provided in the section titled “Methods.”

The temperature dependent values for $\Delta\mu_h$ with error bars are shown in Table S1 of the [supplementary material](#). Figure S1 shows two sets of plots that compare the AMOEBA-derived rFoS values at 298 K to direct measurements for uncharged molecules and to inferred values from parsing of data for charged compounds. The calculated values are in good agreement with experimental data for uncharged molecules. This is reassuring because

AMOEBA is parameterized directly from *ab initio* quantum mechanical calculations, and no knowledge is used with regard to condensed phases or experimental data in condensed phases. We do observe deviations between the AMOEBA-derived rFoS values of charged species and the inferred values from experimental data for whole salts (Fig. S1). These deviations are in accord with the concerns expressed in the Introduction. Inasmuch as the AMOEBA-derived values are direct calculations, we use these numbers as a self-consistent set for uncharged and charged molecules alike.

The results from temperature dependent calculations of $\Delta\mu_h$ for the 19 relevant model compounds are shown in Fig. 1. The enthalpy of hydration (Δh) at $T_0 = 298$ K and the temperature independent heat capacities of hydration (ΔC_p) were extracted for each model compound by fitting the calculated temperature dependent free energies of solvation to the integral of the Gibbs–Helmholtz equation. The results are summarized in Table I. As expected,³⁷ the large positive heat capacity of hydration combined with the favorable enthalpies and unfavorable entropies leads to non-monotonic temperature dependencies for model compound mimics of the

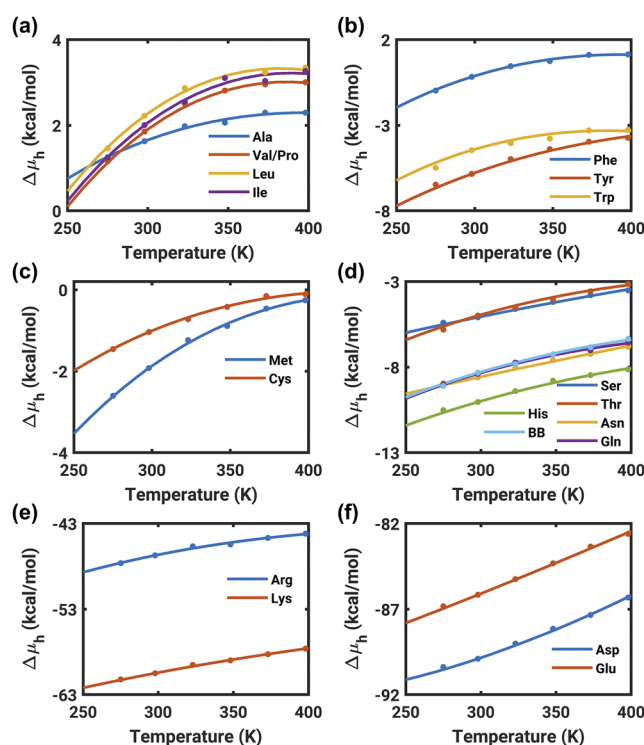


FIG. 1. Temperature dependent free energies of solvation $\Delta\mu_h$ for model compounds that mimic the side chain and backbone moieties. The dots show results from free energy calculations based on the AMOEBA forcefield. These values are then fit to the integral of the Gibbs–Helmholtz equation (see the main text), and the results of the fits are shown as solid curves. Parameters from the fits, which include estimates for Δh and ΔC_p , are shown in Table I. In the legends, we use the three letter abbreviations for each of the amino acids. Here, BB in panel (d) refers to the backbone moiety, modeled using *N*-methylacetamide, that mimics the peptide unit.

TABLE I. Results from free energy calculations that summarize values obtained for $\Delta\mu_h$ at 298 K. Data for the temperature dependence of $\Delta\mu_h$ were fit to Eq. (1), setting $T_0 = 298$ K, to extract values for Δh and Δc_p .

Residue/unit	Model compound	$\Delta\mu_h$ kcal/mol	Δh kcal/mol	Δc_p cal/molK
Ala	Methane	1.63	−2.57	48.93
Val/Pro	Propane	1.85	−6.33	105.80
Leu	2-methylpropane	2.22	−5.92	109.38
Ile	<i>n</i> -butane	2.00	−6.34	105.23
Met	Ethyl methyl thioether	−1.92	−10.11	71.10
Phe	Toluene	−0.17	−8.68	102.24
Cys	Methanethiol	−1.04	−5.84	43.61
Tyr	<i>p</i> -Cresol	−5.85	−15.62	71.09
Trp	3-methylindole	−4.46	−12.67	108.10
Ser	Methanol	−5.08	−10.41	10.43
Thr	Ethanol	−4.98	−12.55	50.06
Asn	Acetamide	−8.61	−14.37	6.18
Gln	Propionamide	−8.39	−16.06	51.47
His	4-methylimidazole	−10.04	−17.60	38.01
Backbone/Gly	<i>N</i> -methylacetamide	−8.33	−16.10	44.73
Arg	<i>n</i> -propylguanidine	−46.72 ^a	−57.4 ^a	69.39
Lys	1-butylamine	−60.49 ^a	−70.37 ^a	29.98
Asp	Acetic acid	−89.91 ^a	−98.65 ^a	−44.97
Glu	Propionic acid	−86.16 ^a	−96.62 ^a	−8.75

^aAs with the default ABSINTH model, in ABSINTH-T, the rFoS values and, therefore, the Δh values we used for ionizable residues are offset from calculated $\Delta\mu_h$ by a fixed constant of −30 kcal/mol. This, as was shown in the original work, is required to avoid the chelation of solution ions around ionizable residues. This “feature” remains a continuing weakness of the ABSINTH paradigm and one that we hope to remedy through suitable generalization of the model used in ABSINTH to interpolate between fully solvated and fully desolvated states.

side chain moieties of Ala, Val, Leu, Ile, and Pro. Similar results are observed for mimics of Phe, Tyr, and Trp. Of import are the differences in hydration thermodynamics of the model compounds that mimic side chains of Lys, Arg, Asp, and Glu. The model compounds 1-butylamine and *n*-propylguanidine that mimic the side chains of Lys and Arg feature a duality of favorable enthalpy of hydration and large positive values for Δc_p . Finally, the deprotonated versions of acetic acid and propionic acid that mimic the deprotonated versions of Asp and Glu, respectively, have the most favorable free energies of hydration across the temperature range studied. Interestingly, these two solutes stand out for their distinctive negative heat capacities of hydration. Inferences based on integral equation theories³⁷ suggest that negative heat capacities of hydration derive from a weakening of the favorable solute–solvent interactions and a reduction in the extent to which water molecules are orientationally distorted within and in the vicinity of the first hydration shell.

Incorporation of *T*-dependent rFoS values into ABSINTH: In the ABSINTH model, each polyatomic solute is parsed into a set of solvation groups.^{20,34} These groups are model compounds for which the free energies of solvation rFoS are known *a priori*. In this work, we follow the work of Wuttke *et al.*¹⁹ and generalize the ABSINTH model to incorporate temperature dependencies of model compound rFoS values. In this ABSINTH-T model, the total solvent-mediated energy associated for a given configuration of the

protein and solution ions is written as

$$E_{\text{total}} = W_{\text{solv}}(T) + W_{\text{el}}(T) + U_{\text{LJ}} + U_{\text{corr}}. \quad (2)$$

Here, $W_{\text{solv}}(\{\text{rFoS}(T)\}, \{\mathbf{r}\})$ is the many-body direct mean field interaction (DMFI) with the continuum solvent that depends on $\{\text{rFoS}(T)\}$, the set of temperature dependent rFoS values of model compounds that make up the solute and solution ions, and $\{\mathbf{r}\}$ is the set of configurational coordinates for polypeptide atoms and solution ions. The term $W_{\text{solv}}(\{\text{rFoS}(T)\}, \{\mathbf{r}\})$ quantifies the free energy change associated with transferring the polyatomic solute into a mean field solvent while accounting for the temperature dependent modulation of the reference free energy of solvation for each solvation group due to other groups of the polyatomic solute and the solution ions. Additional modulations to the free energy of solvation of the solute due to interactions with charged sites on the polyatomic solute are accounted for by the W_{el} term. In ABSINTH-T, the term $W_{\text{el}}(\{\mathbf{r}\}, \{\mathbf{v}\}, \epsilon(T))$ is a function of the set of configurational coordinates $\{\mathbf{r}\}$, solvation states $\{\mathbf{v}\}$ of the solute atoms and solution ions, and the temperature dependent dielectric constant $\epsilon(T)$. For $\epsilon(T)$, we used the parameterization of Wuttke *et al.*¹⁹ The effects of dielectric inhomogeneities, which are reflected in the configuration dependent solvation states, are accounted for without making explicit assumptions regarding the distance or spatial dependencies of dielectric saturation. The term U_{LJ} is a standard 12-6 Lennard-Jones potential, and U_{corr} models specific torsion and bond angle-dependent stereoelectronic effects that are not captured by the U_{LJ} term. The ABSINTH paradigm is optimally interoperable with the Optimized Potentials for Liquid Simulations-All Atom/with LMP2 corrections (OPLS-AA/L) and the Chemistry at Harvard Molecular Mechanics (CHARMM)⁵⁸ family of forcefields, and we use the OPLS-AA/L⁵⁹ forcefield.

Intrinsic solvation (IS) approximation of ABSINTH-T as an efficient heuristic for discriminating IDPs with LCST vs UCST behavior

In the single chain limit, accessible in dilute solutions, polypeptides that show LCST phase behavior undergo collapse above a system-specific theta temperature, whereas polypeptides that show UCST phase behavior expand above the system specific theta temperature.^{1,28} A GADIS-like strategy²¹ for the *de novo* design of polypeptide sequences with LCST phase behavior would involve ABSINTH-T-based all-atom simulations to evaluate whether an increase in temperature leads to chain collapse. In effect, the fitness function in a GA comes from the evaluation of the simulated ensembles as a function of temperature. Computationally, this becomes prohibitively expensive. Accordingly, we pursued a pared down version of ABSINTH-T, which is referred to as the intrinsic solvation (IS) limit of the model.³⁰ The IS limit was introduced to set up sequence and composition specific reference models with respect to which one can use mean-field models to uncover how desolvation impacts IDP ensembles.^{30,60} In effect, the IS limit helps us map conformations in the maximally solvated ensemble and assess how this ensemble changes as a function of temperature. In the IS limit, the energy in a specific configuration for the sequence of interest is written as

$$E_{\text{IS-limit}} = W_{\text{solv}}(T) + U_{\text{LJ}} + U_{\text{corr}}. \quad (3)$$

The only difference between the full model [see Eq. (2)] and the IS limit is the omission of the W_{el} term. This increases the speed of simulations by 1–2 orders of magnitude depending on the system. Next, we asked if ensembles obtained from temperature dependent simulations performed in the IS limit could be used to obtain a suitable heuristic that discriminates sequences with LCST vs UCST behavior. These simulations were performed for a set of 30 sequences (see Table S2 of the [supplementary material](#)) that were previously shown by Quiroz and Chilkoti to have LCST and UCST phase behavior.¹⁷ The results are summarized in Fig. 2 and Figs. S2 and S3 of the [supplementary material](#). As shown in panel (a) of Fig. 2, the radii of gyration (R_g), suitably normalized for comparisons across different sequences of different lengths, appear to be segregated into two distinct classes. To test this hypothesis, we computed the slopes m for each of the profiles of normalized R_g vs temperature. These slopes were calculated in the interval of simulation temperatures between 230 K and 380 K. The results, shown in panel (b) of Fig. 2, clearly indicate that there, indeed, are two categories of sequences. Those that are known to show LCST phase behavior are colored in red, and they fall into a distinct group characterized by negative values of the slope m with an average value of $-5.9 \times 10^{-3} \text{ \AA K}^{-1}$. Here, we use $\text{\AA}$ to denote the units of R_g values normalized by the square root of the chain length N . In contrast, the slope for sequences that show UCST behavior is $-1.4 \times 10^{-3} \text{ \AA K}^{-1}$. Given the range of sequences covered in the calibration based on the IS limit, we pursued an approach, whereby we use slopes of $R_g N^{-0.5}$ vs T as a heuristic to guide the design of a genetic algorithm to find new sequences with LCST phase behavior. It is worth noting that we use the slopes of $R_g N^{-0.5}$ vs T plots instead of specific values of slopes of $R_g N^{-0.5}$ because (a) *a priori* we would not know which temperature to choose for comparison of the R_g values and (b) there is the formal possibility that the curves for $R_g N^{-0.5}$ vs T obtained for different constructs might cross one another, making the issue raised in (a) more confounding.

GA for the design of IDPs that are likely to have LCST phase behavior

We adopted the GADIS algorithm²¹ to explore sequence space and discover candidate IDPs with predicted LCST phase behavior.

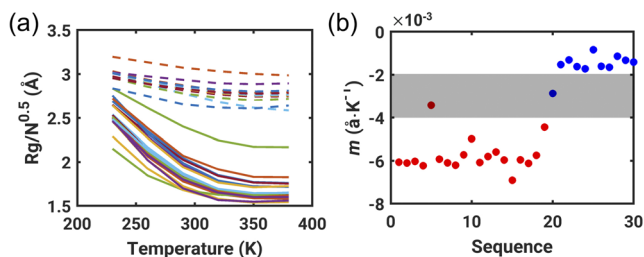


FIG. 2. Analysis of IS limit simulations yields a heuristic that discriminates sequences with UCST vs LCST phase behavior. (a) Plots of $R_g N^{-0.5}$ vs temperature, extracted from IS limit simulations, for sequences shown by Quiroz and Chilkoti to have UCST (dashed lines) vs LCST (solid lines) phase behavior. The sequences are shown in Table S1 of the [supplementary material](#). (b) The slope m of $R_g N^{-0.5}$ vs temperature profiles. These slopes fall into two distinct categories, one for those with LCST phase behavior (blue) and another for those with UCST phase behavior (red). The gray region corresponds to the values of m that clearly demarcate the two categories of sequences.

To introduce the GA and demonstrate its usage, we set about designing novel sequences that are repeats of pentapeptide motifs. We focused on designing 55-mers, i.e., sequences with 11 pentapeptides. To keep the exercise simple, we focused on designing polymers that are perfect repeats of the pentapeptide in question. The GA used in this work is summarized in Fig. 3, and the details are described below.

The GA-based design process is initiated by choosing a random set of 200 sequences. Next, for each of the random sequences, we performed temperature-based replica exchange⁶¹ Metropolis Monte Carlo simulations in the IS limit. The simulation temperatures range from 200 K to 375 K with an interval of 25 K. From each converged IS limit simulation, we computed the ensemble averaged R_g values as a function of simulation temperature T . These data were then used to evaluate the initial set of 200 values for the slope m using the following relationship:

$$m = \frac{1}{N^{0.5}(n-1)} \sum_{i=1}^{n-1} \frac{R_g(T_{i+1}) - R_g(T_i)}{T_{i+1} - T_i}. \quad (4)$$

Here, N is the number of amino acids in each sequence, n is the number of replicas used in the simulation, and T_i is the temperature associated with replica i . The slope m was used to select 100 out of the 200 sequences that were chosen at random initially. The picking probability p was based on the following criterion:

$$p \propto \exp[-c(m - m_0)]. \quad (5)$$

Here, $c = 400$ in units that are reciprocal to m and m_0 is set to $-6.9 \times 10^{-3} \text{ \AA K}^{-1}$. This choice enables an efficient evolution of the GA and a strong selection for sequences with negative values

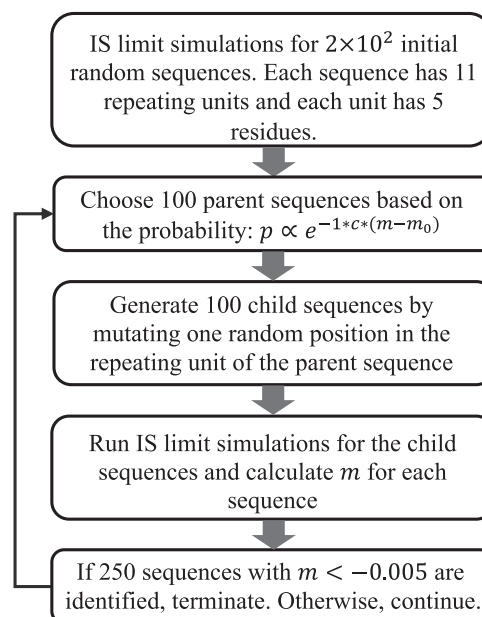


FIG. 3. Workflow of the GA. We use this approach to design sequences that are predicted to have LCST phase behavior. A final post-processing step is added to filter our sequences that do not have high disorder scores (see the main text).

of m . The parameter c ensures numerical stability, guarding against the unnormalized value of p becoming too large or too small.

The chosen parent sequences were used to generate 100 child sequences by mutating a single, randomly chosen position to a randomly chosen residue in the repeating unit. To avoid the prospect of introducing spurious disulfide bonds, we do not include Cys residues either in the original parent pool or for propagating the child sequences. The GA was allowed to evolve for multiple iterations until the convergence criteria were met. These include the generation of at least 250 new sequences, each with a value of m being less than $-5.0 \times 10^{-3} \text{ \AA K}^{-1}$. For the results presented here, six iterations were sufficient to meet the prescribed convergence criteria. The picking probability p determines the selection pressure encoded into the GA. There needs to be an optimal balance between the two extremes in selection pressure. High selection pressures can lead to early convergence to a local optimum, whereas low selection pressures can drastically slow down convergence.⁶² The use of a single evolutionary operator can lead to a single sequence, becoming the dominant choice. The number of iterations that pass before the emergence of a single sequence is known as the takeover time.⁶² High selection pressures lead to low takeover times and vice versa. The issue of a single dominant individual emerging is less of a concern in the sequence design, given the high dimensionality of sequence space. We tuned the choices for c and m_0 to ensure that candidate sequences with putative UCST phase behavior can be part of the offspring, thus lending diversity to sequence evolution by the GA.

Panel (a) in Fig. 4 quantifies the progress of the GA through each iteration of the design process. The quantification is performed in terms of cumulative distribution functions, which for each iteration will quantify the probability that the emerging sequences have associated slope values that are less than or equal to a specific value. The rightward shift in each iteration is indicative of the improved fitness vis-à-vis the selection criterion, which is the lowering of m .

Finally, we added a post-processing step to increase the likelihood that the designed sequences are *bona fide* IDPs. We used the

disorder predictor IUPRED2³¹ to quantify disorder scores for each of the designed sequences. IUPRED2 yields a score between 0 and 1 for each residue, and only sequences where over half of the residues in the repeat are above 0.5 were selected as the final set of designed IDPs that are predicted to have LCST phase behavior. A particular concern with designing sequences for experimental prototyping is the issue of aggregation/precipitation. To ensure that designs were unlikely to create such problems, we calculated predicted solubility scores using the CamSol program⁶³ and found that all sequences that were selected after the post-processing step also have high solubility scores. This provides confidence that the designed IDPs are likely to show phase behavior via liquid–liquid phase separation above system-specific LCST values without creating problems of precipitation/aggregation.

Panel (b) in Fig. 4 summarizes the mean number of each amino acid type observed across the final tally of 64 designed sequences that survive the post-processing step. These statistics are largely in accord with the observations of Quiroz and Chilkoti.¹⁷ Essentially every sequence has at least once Pro residue in the repeat. The beta branched polar amino acid Thr is the other prominent feature that emerges from the selection. The remaining selection preferences fall into four distinct categories that include the following: (i) a clear preference for at least one polar amino acid, viz., His, Ser, Thr, Asn, and Gln; (ii) a clear preference for the inclusion of at least one hydrophobic amino acid, viz., Ala, Ile, Met, and Val; (iii) negligible selection, essentially an avoidance of the acidic residues Asp and Glu, as well as the aromatic residues Phe, Trp, and Tyr; and finally, (iv) a weak preference for Arg over Lys, which is concordant with the distinct temperature dependent profiles for $\Delta\mu_h$ (Fig. 1) and the large positive heat capacity of Arg (Table I). Interestingly, if we fix the positions of Pro and Gly and select for residues in XPXXG or other types of motifs that are inspired by previous work on elastin-like polypeptides, the design process often converges on repeats that are known to be generators of polypeptides with *bona fide* LCST phase behavior (see Fig. S4 of the supplementary material). This observation and the statistics summarized in Fig. 4(b) indicate that the design process uncovers sequences that are likely to have LCST phase behavior.

The designed sequences fall into distinct sequence classes: To quantify the degree of similarity among the set of designed sequences, we computed pairwise Hamming distances between all pairs of the 64 sequences. The resulting Hamming distances were then sorted, and sequences were clustered into distinct groups. Highly similar sequences have low Hamming distances, whereas the converse is true for dissimilar sequences. The resultant Hamming distance map is shown in Fig. 5. The 64 sequences are unevenly distributed across nine major clusters. The actual sequences of the repeats, color-coded by their Hamming distance-based groupings, are shown in Fig. 6. There are two features that stand out. First, sequences deviate from being repeats of VPGVG, which is the elastin-like motif. Second, we find that different sequence permutations on identical or similar composition manifolds emerge as candidates for LCST phase behavior. This observation suggests that at least in the IS limit, it is the composition of each motif rather than the precise sequence that underlies adherence to the selection pressure in the GA. Interestingly, our observations are in accord with results from large-scale *in vitro* characterizations of sequences with

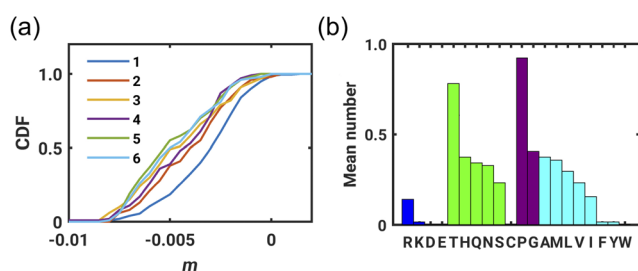


FIG. 4. Calibration of the performance of the GA and statistics for compositional biases that emerge from the application of the design protocol. (a) The cumulative distribution function (CDF) of the slope for sequences in each iteration. There is an overall shift for these CDFs toward smaller m -values with each iteration of the GA. (b) The mean number of each residue in the 64 designed IDPs that are predicted to show LCST phase behavior. Residues in panel (b) are grouped into categories based on their side chain chemistries, i.e., basic residues in blue bars, acidic residues in red bars (although these are not visible since they are not selected), polar residues in green, Pro and Gly in purple, and aliphatic and aromatic residues in cyan. Within each group, the bars are sorted in the descending order of the mean numbers of occurrences in the designs.

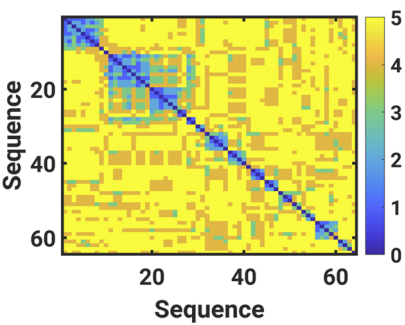


FIG. 5. Identification of distinct sequence classes using a Hamming distance-based assessment of pairwise sequence similarities.

LCST phase behavior.⁶⁴ These experiments show that composition, rather than the precise sequence, is a defining feature of LCST phase behavior—a feature that is distinct from sequences that show UCST phase behavior.³

ABSINTH-T simulations of coil-to-globule transitions for select sequences: We selected four sequence repeats, viz., (TPTGM)₁₁, (PTPLV)₁₁, (LTPTA)₁₁, and (RTAMG)₁₁, for characterization using the full ABSINTH-T model and the calculation of phase diagrams. These sequences were chosen because they are representatives from each of the four major classes that emerge from the design process. Additionally, these sequences bear minimal resemblance to extant designs or naturally occurring sequences that are known to have LCST phase behavior.

Using all-atom, thermal replica exchange Monte Carlo simulations and the full ABSINTH-T model, we performed simulations to test for the presence of a collapse transition for each of the four sequences. The results are shown in Fig. 7. All sequences show a clear tendency to form collapsed conformations as temperature increases. This is diagnosed by there being a clear preference for values of $R_g N^{-0.5}$ being less than the theta state reference value of 2.5 at higher temperatures and values of $R_g N^{-0.5}$ being greater than 2.5 at lower temperatures.

(QPTGM) ₁₁	(MTTGN) ₁₁	(PTPLS) ₁₁	(LTPSA) ₁₁	(INSPH) ₁₁	(AHQRI) ₁₁	(SGNHM) ₁₁	(VMRPG) ₁₁
(GPTGM) ₁₁	(PNKVV) ₁₁	(QTPLS) ₁₁	(LTGGQ) ₁₁	(INAPH) ₁₁	(VHNPP) ₁₁	(HGNHM) ₁₁	(VMPPQ) ₁₁
(QPTGL) ₁₁	(PTPLT) ₁₁	(PAPLS) ₁₁	(LNGGQ) ₁₁	(ANAPH) ₁₁	(MHNPP) ₁₁	(PGTAT) ₁₁	(VGRPM) ₁₁
(QPTVM) ₁₁	(PTNLT) ₁₁	(LTPTQ) ₁₁	(PTPQI) ₁₁	(GNIPH) ₁₁	(MQNPP) ₁₁	(PHYAN) ₁₁	(HMAPQ) ₁₁
(TPTGM) ₁₁	(PTNLS) ₁₁	(LTPTT) ₁₁	(PMPQI) ₁₁	(RTAMT) ₁₁	(GHTTL) ₁₁	(QGHSA) ₁₁	(TQIGH) ₁₁
(TPTVM) ₁₁	(PTPLV) ₁₁	(LTPTA) ₁₁	(APASS) ₁₁	(RTAMG) ₁₁	(GHRTL) ₁₁	(NHMSA) ₁₁	(TQQVH) ₁₁
(GPTVM) ₁₁	(PTPLA) ₁₁	(LQPTA) ₁₁	(APSSM) ₁₁	(RTAPI) ₁₁	(GHMTP) ₁₁	(NHMVS) ₁₁	(SFQNH) ₁₁
(GPTVT) ₁₁	(PNPLV) ₁₁	(TTPTA) ₁₁	(AHPPH) ₁₁	(RTHAI) ₁₁	(PGNIN) ₁₁	(VMRPO) ₁₁	(QAGQT) ₁₁

FIG. 6. Sequences of 64 designed IDPs that emerge from the application of the GA. Different colors except black are used to label sequences in the same group.

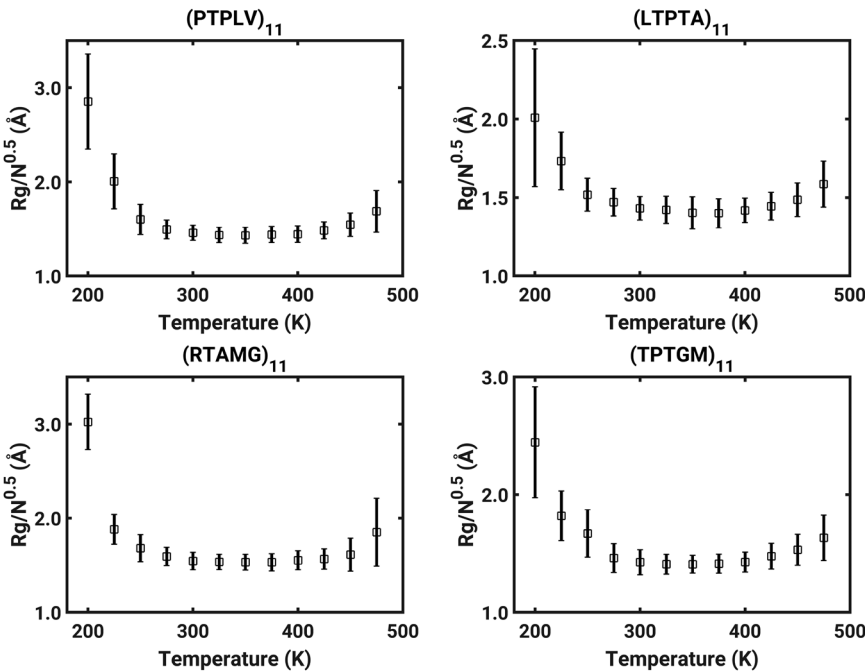


FIG. 7. Profiles of normalized $R_g N^{-0.5}$ vs temperature for four IDPs designed using the GA. The results shown here use the full ABSINTH-T model. The theta temperatures extracted from these simulations are presented in the main text.

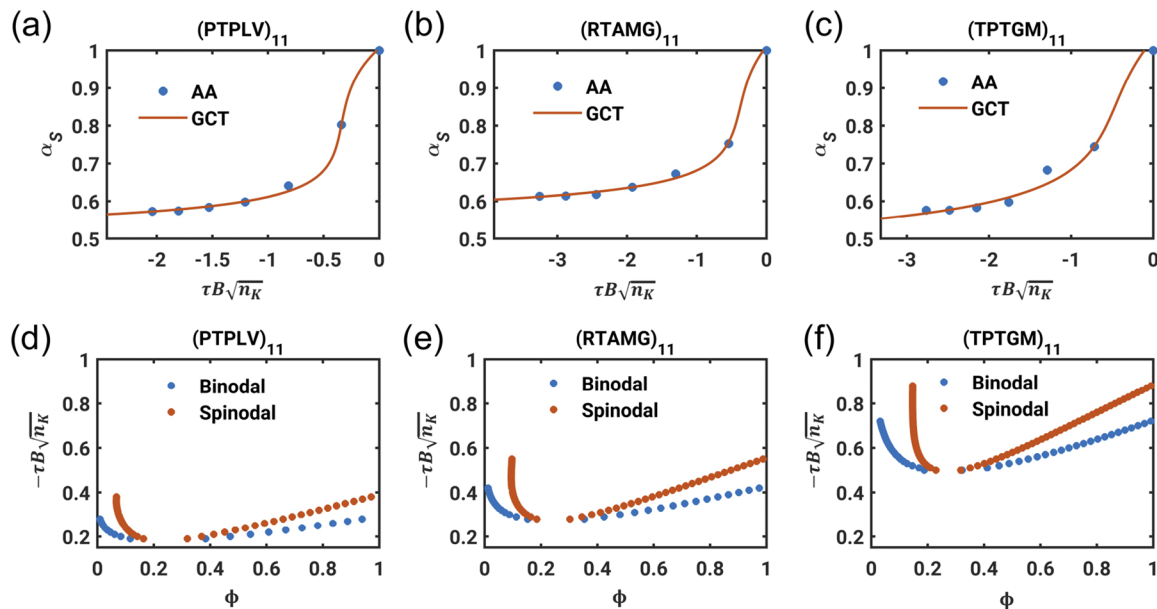


FIG. 8. Results from the application of the Gaussian cluster theory for calculating full phase diagrams. Panels (a)–(c) show the contraction ratio profiles for (PTPLV)₁₁, (RTAMG)₁₁, and (TPTGM)₁₁, respectively. Blue dots represent the contraction ratios calculated from all atom simulations with ABSINTH-T at temperatures from 200 K to 350 K, and red curves represent fits to these data using the Gaussian cluster theory that lead to estimates of the sequence-specific values for the temperature dependent two-body interaction coefficient B and the temperature independent three-body interaction parameter w . Panels (d)–(f) show the full phase diagrams, including the binodal, spinodal, and the estimated location of the critical point for (PTPLV)₁₁, (RTAMG)₁₁, and (TPTGM)₁₁, respectively.

Analysis of coil-globule transitions, extraction of parameters, and calculation of phase diagrams using the Gaussian cluster theory: The profiles of $R_g N^{-0.5}$ vs T were analyzed to extract the theta temperature (T_θ) for each of the four sequences. For this, we used a method that was described recently by Zeng *et al.*²⁸ Only three of the four sequences have coil-globule transition profiles for which a robust estimate of the theta temperature can be made. The extent of expansion at low temperatures is modest and suggests that apparent T_θ for (LTPTA)₁₁ is outside the window where converged simulations can be performed. For the other three sequences, namely, (PTPLV)₁₁, (RTAMG)₁₁, and (TPTGM)₁₁, the estimated T_θ values are 210 K, 210 K, and 200 K, respectively.

Next, we used the estimates of T_θ in conjunction with the Gaussian cluster theory of Guido and Allegra.³² We extracted the two- and three-body interaction coefficients by fitting the contraction ratio α_s calculated from simulations using the formalism of the Gaussian cluster theory, and this yields sequence-specific estimates of B , the two-body interaction coefficient, and w , the three-body interaction coefficient [see panels (a)–(c) of Fig. 8]. These parameters were then deployed to compute full phase diagrams using the numerical approach developed by Zeng *et al.*²⁸ and adapted by others.⁶⁵ The results are shown in panels (d)–(f) of Fig. 8. The abscissas in these diagrams denote the bulk polymer volume fractions, whereas the ordinates quantify temperature in terms of the thermal interaction parameter $\tau B \sqrt{n_K}$. Here, $\tau = \left(\frac{T - T_\theta}{T}\right)$, which is positive for $T > T_\theta$, B is the temperature dependent two-body interaction coefficient inferred from analysis of the contraction ratio, and n_K is the number of the Kuhn segment in the single chain, which we set

to be 11. Note that B is negative for temperatures above T_θ . Accordingly, the thermal interaction parameter is positive above T_θ and the critical temperature T_c . Therefore, comparative assessments of the driving forces for LCST phase behavior can be gleaned by comparing the sequence-specific values of $\tau B \sqrt{n_K}$ and the volume fraction at the critical point. It follows that the sequences can be arranged in the descending order of the driving forces as (TPTGM)₁₁, (RTAMG)₁₁, and (PTPLV)₁₁, respectively. Importantly, full characterization of the phase behavior using a combination of all-atom simulations and numerical adaptation of the Gaussian cluster theory shows that, in general, sequences designed to have LCST phase behavior do match the predictions (see Fig. 8).

DISCUSSION

In this work, we have adapted a GA to design novel sequences of repetitive IDPs that we predict to have LCST phase behavior. Our method is aided by a learned heuristic that was shown to provide clear segregation between sequences with known LCST vs UCST phase behavior. This heuristic is the slope m of the change in $R_g N^{-0.5}$ vs T from simulations of sequences performed in the IS limit of the ABSINTH-T model. We use the heuristic in conjunction with IS limit simulations to incorporate a selection pressure into the GA, thereby allowing the selection of sequences that are “fit” as assessed by the heuristic to be predictive of LCST phase behavior.

Here, we presented one instantiation of the GA and used it to uncover 64 novel sequences that can be grouped into four major classes and several minor classes (Fig. 6). We then focused on four

sequences, one each from each of the four major classes, and characterized temperature dependent coil–globule transitions. These profiles, analyzed in conjunction with recent adaptations of the Gaussian cluster theory,³² allowed us to extract sequence-specific values for theta temperatures, temperature dependent values of the two-body interaction coefficients, and three-body interaction coefficients. We incorporated these parameters into our numerical implementation²⁸ of the Gaussian cluster theory to calculate full phase diagrams for three sequences. These affirm the predictions of LCST phase behavior and demonstrate sequence-specificity in control over the driving forces for thermoresponsive phase behavior.

Our overall approach is aided by the following advances: We used the AMOEBA forcefield²⁹ to obtain direct estimates of temperature dependent free energies of solvation for model compounds used to mimic the side chain and backbone moieties. These temperature dependent free energies of solvation were used in conjunction with the integral of the Gibbs–Helmholtz equation to obtain model compound specific values for the enthalpy and heat capacity of hydration.

The methods we present here are a start toward the integration of supervised learning to leverage information gleaned from systematic characterizations of IDP phase behavior and physical chemistry based computations that combined all-atom simulations with improvements such as ABSINTH-T and theoretical calculations that allow us to connect single chain coil–globule transitions to full phase diagrams.²⁸ The heuristic we have extracted from IS limit simulations helps in discriminating sequences with LCST vs UCST phase behavior. These simulations are sufficient for IS limit driven and GA aided designs of sequences that are expected to have LCST phase behavior. This is because composition as opposed to the syntactic details of sequences plays a determining role of LCST phase behavior.³ Recent studies have shown that even the simplest changes to sequence syntax can have profound impacts on UCST phase behavior.⁶⁶ This makes it challenging to guide the design of sequences with predicted UCST phase behavior that relies exclusively on IS limit simulations. We will need to incorporate simulations based on either transferrable⁶⁷ or learned coarse-grained models⁶⁸ as a substitution for the IS limit simulations. This approach comes with challenges because one has to be sure that the coarse-grained models afford the requisite sequence specificity without compromising efficiency. The work of Dignon *et al.*⁶⁹ is noteworthy in this regard. Their coarse-grained model, which is based on knowledge-based potentials parameterized to have temperature dependent interactions, has been shown to be very effective in discriminating sequences that are shown to have UCST vs LCST phase behavior.⁶⁹ The conceptual underpinnings of their approach and that presented here derive from the work of Wuttke *et al.*¹⁹ It would be interesting to combine or compare our approach to that of Dignon *et al.* in the context of designing novel IDPs and characterizing their phase behavior. We view these approaches as being complementary rather than competing ones, and we expect that the approaches will have distinct advantages in different settings. The specific feature of our approach is that the calculations, at least for designing sequences with LCST phase behavior, do not ever become more complex than single chain simulations. This has the value for achieving design objectives. It also has the value for designing sequences that are not only thermoresponsive but also responsive to changes in pH, pressure, and other solvent parameters, especially since recent

studies suggest that solution space scanning is a way to obtain efficient delineation of the desirable conformational and phase equilibria for IDPs.⁷⁰

The design of sequences with UCST phase behavior or sequences that combine UCST and LCST phase behavior, going beyond simple block copolymeric designs, will be of utmost interest for developing new IDP based materials. Additionally, we hope to build on improved understanding⁷¹ of the impact of pH on conformational⁷² and phase equilibria⁷³ of IDPs and the impact of metal chelation sites on phase behavior⁷⁴ to design sequences that combine the ability to exhibit phase behavior in response to orthogonal stimuli. Such efforts are of direct relevance to engineering orthogonal biomolecular condensates into simple unicellular prokaryotic and eukaryotic cells, as has been demonstrated recently with the engineering a protein translation circuit into protocells based on a thermoresponsive elastin-like polypeptide.⁷⁵ Of course, the proof of the validity/accuracy of designs and predictions will have to come from experimental work geared toward testing the predictions/designs. These efforts—that leverage the high-throughput expression of these *de novo* sequences in *E. coli* and *in situ* characterization of their phase behavior—are underway.⁷⁶ The initial experimental investigations suggest that the designs reported here and those that will emerge from the application of the methods deployed in this work do indeed show LCST phase behavior. Detailed reports of these experimental characterizations will follow in separate work.

METHODS

AMOEBA forcefield parameterization for the model compounds of interest

To obtain values of free energies of solvation from AMOEBA simulations, we first derived the AMOEBA forcefield parameters for the model compounds listed in Table I of the main text. The parameters for *N*-methylacetamide, methane, methanol, ethanol, toluene, and *p*-Cresol are taken from previous work,⁵⁵ which was released in the amoeba09.prm parameter file in the TINKER package.⁷⁷ The parameters for other model compounds are derived following the standard automated protocol that has been established for the AMOEBA forcefield.⁷⁸ Briefly, the protocol involves the following steps: Quantum chemical calculations were utilized to derive the electrostatic parameters; these include atom-centered partial charges and dipole and quadrupole moments. The molecular structures were fully optimized at the MP2/6-31G* level of theory⁷⁹ followed by MP2/cc-pvtz calculations to obtain the electron density of the molecules. Then, the initial multipole parameters were determined via distributed multipole analysis calculation via the Gaussian Distributed Multipole Analysis (GDMA) program.⁸⁰ With the charges being fixed, the dipole and quadrupole moments were further fit to the electrostatic potential generated at the MP2/aug-cc-pvtz level on a grid of points outside of the molecules, where the least-squares restrained optimization was used to keep the multipole moments close to their Distributed Multipole Analysis (DMA) derived values while providing improved electrostatic potentials. The poedit and potential programs of the TINKER package⁷⁷ were used in this process.

The Thole damping⁸¹ value of 0.39 and the standard AMOEBA atomic polarizabilities were assigned for each atom. Valence and

van der Waals (vdW) parameters were directly assigned from the existing small molecule library and Molecular Mechanics 3 (MM3) forcefield, and the equilibrium values for bond lengths and bond angles were calculated from the above QM-optimized geometry. Torsional parameters of rotatable bonds were obtained by comparing the conformational energy profile of QM and AMOEBA models, which includes electrostatics, polarization, vdW, and valence terms. The dihedral angle was scanned by minimizing all torsions about the rotatable bond of interest at 30° intervals with restrained optimization at the HF/6-31G* level of theory. The QM conformational energy was obtained as the single point energy at the ω B97XD/6-311++G(d,p) level of theory.⁸² Torsions about the same rotatable bond that are also in-phase are collapsed into one set of parameters for the fitting, and the contributions are distributed evenly among the parameters. AMOEBA uses the traditional Fourier expansion up to six-fold. Here, the force constant parameters were fit using one-, two-, and three-fold trigonometric forms. All the quantum calculations were performed using the Gaussian 09 software package.⁸³ The parameterization process has been automated in the Poltype (version 2) software.⁷⁸ All the parameters derived above are appended as part of a separate text file in the [supplementary material](#).

Setup of molecular dynamics simulations using AMOEBA

All AMOEBA simulations were performed using the TINKER-OpenMM package.⁸⁴ Each model compound was solvated in a cubic water box with periodic boundary conditions. The initial dimensions of the central cell were set to be $30 \times 30 \times 30 \text{ \AA}^3$. Following energy minimization, molecular dynamics simulations were performed using the reversible reference system propagator algorithm (RESPA) integrator⁸⁵ with an inner time step of 0.25 ps and an outer time step of 2.0 fs in the isothermal–isobaric ensemble (NPT) with the target temperature being between 273 K and 400 K depending on the temperature of interest and the target pressure being 1 bar. The temperature and pressure were controlled using a stochastic velocity rescaling thermostat⁸⁶ and a Monte Carlo constant pressure algorithm,⁸⁷ respectively. The particle mesh Ewald (PME) method⁸⁸ with PME-GRID being 36 grid units, an order 8 B-spline interpolation,⁸⁹ with a real space cutoff of 7 Å was used to compute long-range corrections to electrostatic interactions. The cutoff for van der Waals interactions was set to be 12 Å. This combination of a shorter cutoff for PME real space and a longer cutoff for buffered-14-7 potential has been verified⁹⁰ for AMOEBA free energy simulations.⁹¹ Snapshots were saved every ps. In simulations performed along a prescribed schedule for the Kirkwood coupling parameters (see below), we use the same solvent box across the schedule. However, the velocities were randomized at the start of each simulation, and the first 1 ns of data were set aside as equilibration and not used in the free energy estimations.

Free energy calculations

We used the Bennett Acceptance Ratio (BAR)⁵⁶ method to quantify the free energies of solvation for the model compounds of interest. This method has been shown to be superior to other free energy estimators in terms of reducing the statistical errors in calculations of free energies of solvation.⁹² The solute is grown using two different Kirkwood coupling parameters λ_{vdW} and λ_{el} that scale

the strengths of solute–solute and solute–solvent van der Waals and electrostatic interactions. A series of independent molecular dynamics simulations were performed in the NPT ensemble for different combinations of λ_{vdW} and λ_{el} . A soft-core modification of the buffered-14-7 function was used to scale the vdW interactions as implemented in Tinker-OpenMM.⁸⁴ We used the following combinations for the scaling coefficients: $[\lambda_{\text{vdW}}, \lambda_{\text{el}}] \lambda \equiv [0, 0], [0.1, 0], [0.2, 0], [0.3, 0], [0.4, 0], [0.5, 0], [0.6, 0], [0.7, 0], [0.8, 0], [0.9, 0], [1, 0], [1, 0.1], [1, 0.2], [1, 0.3], [1, 0.4], [1, 0.5], [1, 0.6], [1, 0.7], [1, 0.8], [1, 0.9], [1, 1]$. For each pair of λ values, we performed simulations, each of length 6 ns, at the desired temperature and a pressure of 1 bar. We then used the TINKER bar program to calculate the free energy difference between neighboring windows defined in terms of the scaling coefficients. For every combination of λ_{vdW} and λ_{el} , we set aside the first 1 ns simulation as part of the equilibration process. Finally, for each model compound, we computed free energies of solvation at six different temperatures, viz., 275 K, 298 K, 323 K, 348 K, 373 K, 398 K, thus giving us the direct estimates of temperature dependent free energies of solvation that we sought from the AMOEBA-based simulations. Note that 398 K is above the boiling point of water. However, although the physical properties of water are accurately captured by the AMOEBA model, the finite size of the system, the starting conditions, and the finite duration of the simulations, even though they are in the NPT ensemble, imply that water at 398 K and 1 bar corresponds to superheated liquid water.

The temperature dependent free energies of solvation were fit to the integral of the Gibbs–Helmholtz equation—see Eq. (1) in the main text. The free energy calculations provide us with direct estimates for $r\text{FoS}(T)$ at specific values for T . We set $T_0 = 298 \text{ K}$ and use the non-linear regression to fit Eq. (1) to the calculated values for $r\text{FoS}(T)$. The regression analysis provides estimates of Δh and Δc_p , which we then use, in conjunction with Eq. (1) in the manner prescribed by Wuttke *et al.*,¹⁹ for all the ABSINTH-T-based simulations.

Setup of Monte Carlo simulations in the IS limit and using ABSINTH-T

Thermal replica exchange⁶¹ Monte Carlo simulations were performed using version 2.0 of the CAMPARI modeling software (<http://campari.sourceforge.net/>). The temperature schedule for thermal replica exchange simulations that use the full ABSINTH-T model ranges from 200 K to 470 K with an interval of 25 K. A total of 6×10^7 independent moves were attempted per replica. For systems in Fig. 7, we performed three independent sets of thermal replica exchange simulations. All the simulations are performed within a spherical droplet with the radius of 100 Å. The other settings were identical to those used by Zeng *et al.*²⁸

The details of the simulations including parameters, move sets, analyses, and the design of the simulations are identical to those published in the recent work of Zeng *et al.*²⁸ Briefly, we used the ABSINTH-T implicit solvent model and forcefield paradigm. The forcefield parameters are based on the `abs_ops_3.2.prm` set, and they include the parameters for proline residues that were developed by Radhakrishnan *et al.*⁹³ However, they do not include the CMAP corrections introduced by Choi and Pappu.³⁴ The AMOEBA-based $r\text{FoS}$ values at 298 K were incorporated into the standard

parameter file, and the temperature dependent rFoS values were calculated using the model compound specific values for Δh and Δc_p that were derived using the AMOEBA-based calculations of $\Delta\mu_h(T_0)$ —see Table I. As in our recent work,²⁸ we used the temperature dependent dielectric constant prescribed by Wuttke *et al.*¹⁹ Neutralizing counterions were added to the simulation droplet for polypeptides with net charges to neutralize the system. For the Na^+ and Cl^- ions, we use the following values for $\Delta\mu_h$ at 298 K, Δh , and Δc_p , respectively: $\{-74.6 \text{ kcal/mol}, -80.2 \text{ kcal/mol}, -18.4 \text{ cal/mol K}\}$ and $\{-87.2 \text{ kcal/mol}, -99.2 \text{ kcal/mol}, -11.7 \text{ cal/mol K}\}$. The Lennard-Jones parameters for Na^+ and Cl^- ions are default parameters in the original work of Vitalis and Pappu.²⁰

For the IS limit simulations, we turned off the W_{el} term by setting the keyword SC_POLAR to be 0 in the key file. For each of the systems shown in Fig. 2, we performed one set of replica exchange simulations, and a total of 6×10^7 independent moves were attempted per replica. The temperature schedule for the replica exchange simulation is from 230 K to 380 K with an interval of 30 K. Error bars in Fig. 7 and in Figs. S2 and S3 are reported as standard deviations of the distribution of mean R_g values for each simulation temperature.

SUPPLEMENTARY MATERIAL

See the [supplementary material](#) for (a) the sequences shown by Quiroz and Chilkoti to have UCST and LCST phase behaviors that were used for IS limit simulations in this study, (b) the temperature dependent free energies of solvation derived from free energy calculations using the AMOEBA forcefield, (c) figures showing the temperature dependent R_g profiles calculated in the IS limit, and (d) figures quantifying the statistics for residues selected as part of the design process directed toward the XPXXG system. A zip archive appended to the [supplementary material](#) includes the parameter file for the AMOEBA forcefield and a sample key file for performing free energy calculations based on molecular dynamics simulations.

ACKNOWLEDGMENTS

This work was supported by Grant Nos. DMR 1729783 from the US National Science Foundation (A.C. and R.V.P.), RGP0034/2017 from the Human Frontier Science Program (R.V.P.), and R01GM114237 from the US National Institutes of Health (P.R.). Resources from the Center for High Performance Computing (CHPC) and the Research Infrastructure Services (RIS) at Washington University in St. Louis were used for some of the simulations. X.Z. and R.V.P. are grateful to Dr. Andreas Vitalis for several discussions over the years regarding issues that arise with regard to T-dependent free energies of solvation and the parameters for Δh and Δc_p .

DATA AVAILABILITY

The data that support the findings of this study are available within the article and its [supplementary material](#).

REFERENCES

¹K. M. Ruff, S. Roberts, A. Chilkoti, and R. V. Pappu, *J. Mol. Biol.* **430**(23), 4619 (2018).

- ²M. S. Cortese, V. N. Uversky, and A. Keith Dunker, *Prog. Biophys. Mol. Biol.* **98**(1), 85 (2008).
- ³M. Dzuricky, S. Roberts, and A. Chilkoti, *Biochemistry* **57**(17), 2405 (2018).
- ⁴S. Rauscher and R. Pomès, “Fuzziness: Structural disorder in protein complexes,” in *Monika Fuxreiter and Peter Tompa* (Springer US, New York, NY, 2012), p. 159.
- ⁵I. Weitzhandler, M. Dzuricky, I. Hoffmann, F. G. Quiroz, M. Gradzielski, and A. Chilkoti, *Biomacromolecules* **18**(8), 2419 (2017).
- ⁶J. R. Simon, N. J. Carroll, M. Rubinstein, A. Chilkoti, and G. P. López, *Nat. Chem.* **9**(6), 509 (2017).
- ⁷M. Saric and T. Scheibel, *Curr. Opin. Biotechnol.* **60**, 213 (2019).
- ⁸J. Hsin, J. Strümpfer, E. H. Lee, and K. Schulten, *Annu. Rev. Biophys.* **40**(1), 187 (2011).
- ⁹R. Lei, J. P. Lee, M. B. Francis, and S. Kumar, *Biochemistry* **57**(27), 4019 (2018).
- ¹⁰A. K. Varanko, J. C. Su, and A. Chilkoti, *Annu. Rev. Biomed. Eng.* **22**(1), 343 (2020).
- ¹¹J. J. de Pablo, N. E. Jackson, M. A. Webb, L.-Q. Chen, J. E. Moore, D. Morgan, R. Jacobs, T. Pollock, D. G. Schlom, E. S. Toberer, A. James, I. Dabo, D. M. DeLongchamp, G. A. Fiete, G. M. Grason, G. Hautier, Y. Mo, K. Rajan, E. J. Reed, E. Rodriguez, V. Stevanovic, J. Sunitich, K. Thornton, and J.-C. Zhao, *npj Comput. Mater.* **5**(1), 41 (2019).
- ¹²R. K. Das, K. M. Ruff, and R. V. Pappu, *Curr. Opin. Struct. Biol.* **32**, 102 (2015).
- ¹³J. Guillén-Boixet, A. Kopach, A. S. Holehouse, S. Wittmann, M. Jahnel, R. Schlüsler, K. Kim, I. R. E. A. Trussina, J. Wang, D. Mateju, I. Poser, S. Maharana, M. Ruer-Gruß, D. Richter, X. Zhang, Y.-T. Chang, J. Guck, A. Honigmann, J. Mahamid, A. A. Hyman, R. V. Pappu, S. Alberti, and T. M. Franzmann, *Cell* **181**(2), 346 (2020).
- ¹⁴J.-M. Choi, A. S. Holehouse, and R. V. Pappu, *Annu. Rev. Biophys.* **49**(1), 107 (2020).
- ¹⁵S. Banjade, Q. Wu, A. Mittal, W. B. Peeples, R. V. Pappu, and M. K. Rosen, *Proc. Natl. Acad. Sci. U. S. A.* **112**(47), E6426 (2015).
- ¹⁶M. C. Cohan and R. V. Pappu, *Trends Biochem. Sci.* **45**(8), 668 (2020).
- ¹⁷F. G. Quiroz and A. Chilkoti, *Nat. Mater.* **14**(11), 1164 (2015).
- ¹⁸F. Tanaka, T. Koga, I. Kaneda, and F. M. Winnik, *J. Phys.: Condens. Matter* **23**(28), 284105 (2011).
- ¹⁹R. Wuttke, H. Hofmann, D. Nettels, M. B. Borgia, J. Mittal, R. B. Best, and B. Schuler, *Proc. Natl. Acad. Sci. U. S. A.* **111**(14), 5213 (2014).
- ²⁰A. Vitalis and R. V. Pappu, *J. Comput. Chem.* **30**(5), 673 (2009).
- ²¹T. S. Harmon, M. D. Crabtree, S. L. Shammass, A. E. Posey, J. Clarke, and R. V. Pappu, *Protein Eng., Des. Sel.* **29**(9), 339 (2016).
- ²²H. Kojima, *Polym. J.* **50**(6), 411 (2018).
- ²³G. Zhang and C. Wu, *Adv. Polym. Sci.* **195**, 101 (2006).
- ²⁴B. H. Zimm and J. K. Bragg, *J. Chem. Phys.* **31**(2), 526 (1959).
- ²⁵Y. Okada and F. Tanaka, *Macromolecules* **38**(10), 4465 (2005).
- ²⁶H. Kojima and F. Tanaka, *Macromolecules* **43**(11), 5103 (2010).
- ²⁷F. Tanaka, *Macromolecules* **33**(11), 4249 (2000).
- ²⁸X. Zeng, A. S. Holehouse, A. Chilkoti, T. Mittag, and R. V. Pappu, *Biophys. J.* **119**(2), 402 (2020).
- ²⁹J. W. Ponder, C. Wu, P. Ren, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D. S. Lambrecht, R. A. DiStasio, M. Head-Gordon, G. N. I. Clark, M. E. Johnson, and T. Head-Gordon, *J. Phys. Chem. B* **114**(8), 2549 (2010).
- ³⁰R. K. Das and R. V. Pappu, *Proc. Natl. Acad. Sci. U. S. A.* **110**(33), 13392 (2013).
- ³¹B. Mészáros, G. Erdős, and Z. Dosztányi, *Nucleic Acids Res.* **46**(W1), W329 (2018).
- ³²R. Guido and G. Allegra, *J. Chem. Phys.* **104**(4), 1626 (1996).
- ³³T. Lazaridis and M. Karplus, *Proteins: Struct., Funct., Bioinf.* **35**(2), 133 (1999).
- ³⁴J.-M. Choi and R. V. Pappu, *J. Chem. Theory Comput.* **15**(2), 1367 (2019).
- ³⁵G. I. Makhataдзе and P. L. Privalov, *J. Mol. Biol.* **232**(2), 639 (1993).
- ³⁶S. Cabani, P. Gianni, V. Mollica, and L. Lepori, *J. Solution Chem.* **10**(8), 563 (1981).
- ³⁷N. V. Prabhu and K. A. Sharp, *Annu. Rev. Phys. Chem.* **56**(1), 521 (2005).
- ³⁸A. Ben-Naim, K.-L. Ting, and R. L. Jernigan, *Biopolymers* **28**(7), 1309 (1989).
- ³⁹D. Asthagiri, L. R. Pratt, and H. S. Ashbaugh, *J. Chem. Phys.* **119**(5), 2702 (2003).

- ⁴⁰H. S. Ashbaugh and M. E. Paulaitis, *J. Am. Chem. Soc.* **123**(43), 10721 (2001).
- ⁴¹A. Grossfield, P. Ren, and J. W. Ponder, *J. Am. Chem. Soc.* **125**(50), 15671 (2003).
- ⁴²R. Wolfenden, *Biochemistry* **17**(1), 201 (1978).
- ⁴³G. I. Makhatadze, M. M. Lopez, and P. L. Privalov, *Biophys. Chem.* **64**(1), 93 (1997).
- ⁴⁴Y. Marcus, *J. Chem. Soc., Faraday Trans. 1* **82**(1), 233 (1986).
- ⁴⁵Y. Marcus, *J. Phys. Chem. B* **109**(39), 18541 (2005).
- ⁴⁶Z. A. Panagiotopoulos, *J. Chem. Phys.* **153**(1), 010903 (2020).
- ⁴⁷G. A. Krestov, *Thermodynamics of Solvation, Solution and Dissolution; Ions and Solvents; Structure and Energetics* (Ellis Horwood Ltd., New York, NY, 1991).
- ⁴⁸O. Markovitch, H. Chen, S. Izvekov, F. Paesani, G. A. Voth, and N. Agmon, *J. Phys. Chem. B* **112**(31), 9456 (2008).
- ⁴⁹B. Marten, K. Kim, C. Cortis, R. A. Friesner, R. B. Murphy, M. N. Ringnalda, D. Sitkoff, and B. Honig, *J. Phys. Chem.* **100**(28), 11775 (1996).
- ⁵⁰Y. Marcus, *J. Chem. Soc., Faraday Trans.* **87**(18), 2995 (1991).
- ⁵¹Y. Marcus, *J. Chem. Soc., Faraday Trans. 1* **83**(2), 339 (1987).
- ⁵²M. H. Abraham and Y. Marcus, *J. Chem. Soc., Faraday Trans. 1* **82**(10), 3255 (1986).
- ⁵³P. Ren and J. W. Ponder, *J. Phys. Chem. B* **107**(24), 5933 (2003).
- ⁵⁴Y. Shi, C. Wu, J. W. Ponder, and P. Ren, *J. Comput. Chem.* **32**(5), 967 (2011).
- ⁵⁵P. Ren, C. Wu, and J. W. Ponder, *J. Chem. Theory Comput.* **7**(10), 3143 (2011).
- ⁵⁶C. H. Bennett, *J. Comput. Phys.* **22**(2), 245 (1976).
- ⁵⁷M. Kinoshita and T. Yoshidome, *J. Chem. Phys.* **130**(14), 144705 (2009).
- ⁵⁸A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus, *J. Phys. Chem. B* **102**(18), 3586 (1998).
- ⁵⁹G. A. Kaminski, R. A. Friesner, J. Tirado-Rives, and W. L. Jorgensen, *J. Phys. Chem. B* **105**(28), 6474 (2001).
- ⁶⁰S. Lucas, J. Huihui, and K. Ghosh, *J. Chem. Theory Comput.* **13**(10), 5065 (2017).
- ⁶¹A. Mitsutake, Y. Sugita, and Y. Okamoto, *J. Chem. Phys.* **118**(14), 6664 (2003).
- ⁶²M. P. Thompson, J. D. Hamann, and J. Sessions, *Int. J. For. Res.* **2009**, 527392.
- ⁶³P. Sormanni and M. Vendruscolo, *Cold Spring Harbor Perspect. Biol.* **11**(12), a033845 (2019).
- ⁶⁴M. Amiram, F. G. Quiroz, D. J. Callahan, and A. Chilkoti, *Nat. Mater.* **10**(2), 141 (2011).
- ⁶⁵H.-Y. Chou and A. Aksimentiev, *J. Phys. Chem. Lett.* **11**(12), 4923 (2020).
- ⁶⁶F. G. Quiroz, N. K. Li, S. Roberts, P. Weber, M. Dzuricky, I. Weitzhandler, Y. G. Yingling, and A. Chilkoti, *Sci. Adv.* **5**(10), eaax5177 (2019).
- ⁶⁷Z. Monahan, V. H. Ryan, A. M. Janke, K. A. Burke, S. N. Rhoads, G. H. Zerze, R. O'Meally, G. L. Dignon, A. E. Conicella, W. Zheng, R. B. Best, R. N. Cole, J. Mittal, F. Shewmaker, and N. L. Fawzi, *EMBO J.* **36**(20), 2951 (2017).
- ⁶⁸J.-M. Choi, F. Dar, and R. V. Pappu, *PLoS Comput. Biol.* **15**(10), e1007028 (2019).
- ⁶⁹G. L. Dignon, W. Zheng, Y. C. Kim, and J. Mittal, *ACS Cent. Sci.* **5**(5), 821 (2019).
- ⁷⁰A. S. Holehouse and S. Sukenik, *J. Chem. Theory Comput.* **16**(3), 1794 (2020).
- ⁷¹F. G. Quiroz, V. F. Fiore, J. Levorse, L. Polak, E. Wong, H. A. Pasolli, and E. Fuchs, *Science* **367**(6483), eaax9554 (2020).
- ⁷²M. J. Fossat and R. V. Pappu, *J. Phys. Chem. B* **123**(32), 6952 (2019).
- ⁷³O. Adame-Arana, C. A. Weber, V. Zaburdaev, J. Prost, and F. Jülicher, *Biophys. J.* **119**(8), 1590 (2020).
- ⁷⁴K. Hong, D. Song, and Y. Jung, *Nat. Commun.* **11**(1), 5554 (2020).
- ⁷⁵J. R. Simon, S. A. Eghtesadi, M. Dzuricky, L. You, and A. Chilkoti, *Mol. Cell* **75**(1), 66 (2019).
- ⁷⁶M. Dzuricky, B. A. Rogers, A. Shahid, P. S. Cremer, and A. Chilkoti, *Nat. Chem.* **12**(9), 814 (2020).
- ⁷⁷J. A. Rackers, Z. Wang, C. Lu, M. L. Laury, L. Lagardère, M. J. Schnieders, J.-P. Piquemal, P. Ren, and J. W. Ponder, *J. Chem. Theory Comput.* **14**(10), 5273 (2018).
- ⁷⁸J. C. Wu, G. Chattree, and P. Ren, *Theor. Chem. Acc.* **131**(3), 1138 (2012).
- ⁷⁹T. H. Dunning, Jr., *J. Chem. Phys.* **90**(2), 1007 (1989).
- ⁸⁰A. J. Stone, *J. Chem. Theory Comput.* **1**(6), 1128 (2005).
- ⁸¹B. T. Thole, *Chem. Phys.* **59**(3), 341 (1981).
- ⁸²J.-D. Chai and M. Head-Gordon, *J. Chem. Phys.* **128**(8), 084106 (2008).
- ⁸³G. W. Trucks, M. J. Frisch, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox, G09: Gaussian (Gaussian, Inc., Wallingford, CT, 2009).
- ⁸⁴M. Harger, D. Li, Z. Wang, K. Dalby, L. Lagardère, J.-P. Piquemal, J. Ponder, and P. Ren, *J. Comput. Chem.* **38**(23), 2047 (2017).
- ⁸⁵M. E. Tuckerman, B. J. Berne, and G. J. Martyna, *J. Chem. Phys.* **94**(10), 6811 (1991).
- ⁸⁶G. Bussi, T. Zykova-Timan, and M. Parrinello, *J. Chem. Phys.* **130**(7), 074101 (2009).
- ⁸⁷J. Åqvist, P. Wennerström, M. Nervall, S. Bjelic, and B. O. Brandsdal, *Chem. Phys. Lett.* **384**(4), 288 (2004).
- ⁸⁸T. Darden, D. York, and L. Pedersen, *J. Chem. Phys.* **98**(12), 10089 (1993).
- ⁸⁹E. Ulrich, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, *J. Chem. Phys.* **103**(19), 8577 (1995).
- ⁹⁰Z. Jing, C. Liu, R. Qi, and P. Ren, *Proc. Natl. Acad. Sci. U. S. A.* **115**(32), E7495 (2018).
- ⁹¹D. Jiao, P. A. Golubkov, T. A. Darden, and P. Ren, *Proc. Natl. Acad. Sci. U. S. A.* **105**(17), 6290 (2008).
- ⁹²M. A. Wycalkowski, A. Vitalis, and R. V. Pappu, *J. Phys. Chem. B* **114**(24), 8166 (2010).
- ⁹³A. Radhakrishnan, A. Vitalis, A. H. Mao, A. T. Steffen, and R. V. Pappu, *J. Phys. Chem. B* **116**(23), 6862 (2012).