Inclusion of persistence length-based secondary structure in replica field theoretic models of heteropolymer freezing

Jeffrey K. Weber and Vijay S. Pande^{a)}

Department of Chemistry, Stanford University, Stanford, California 94305, USA

(Received 15 April 2013; accepted 10 July 2013; published online 31 July 2013)

The protein folding problem has long represented a “holy grail” in statistical physics due to its physical complexity and its relevance to many human diseases. While past theoretical work has yielded apt descriptions of protein folding landscapes, recent large-scale simulations have provided insights into protein folding that were impractical to obtain from early theories. In particular, the role that non-native contacts play in protein folding, and their relation to the existence of misfolded, β -sheet rich trap states on folding landscapes, has emerged as a topic of interest in the field. In this paper, we present a modified model of heteropolymer freezing that includes explicit secondary structural characteristics which allow observations of “intramolecular amyloid” states to be probed from a theoretical perspective. We introduce a variable persistence length-based energy penalty to a model Hamiltonian, and we illustrate how this modification alters the phase transitions present in the theory. We find, in particular, that inclusion of this variable persistence length increases both generic freezing and folding temperatures in the model, allowing both folding and glass transitions to occur in a more highly optimized fashion. We go on to discuss how these changes might relate to protein evolution, misfolding, and the emergence of intramolecular amyloid states.

© 2013 AIP Publishing LLC. [<http://dx.doi.org/10.1063/1.4816633>]

INTRODUCTION

Protein folding has long captured the attention of physicists not only because of its pertinence to the fields of polymer physics, biophysics, and the physics of disordered systems, but also because the protein folding problem offers physicists a platform to help untangle the mysteries of human disease. Attempts to explain how a linear chain of amino acids assembles into its unique native state in time to perform a meaningful biological function have yielded elegant theories of rugged, native-biased free energy surfaces and hub-like network kinetics.^{1,2} The desire for comprehensive understanding of protein folding landscapes is driven in part by the perplexities of protein misfolding, and the pathological role misfolding and protein aggregation play in a wide array of diseases.³ In particular, evidence suggests that amyloid fibrils, protein aggregates rich in β -sheet structure, are at the locus for toxicity in a number of neurodegenerative disorders including Alzheimer's, Parkinson's, and other diseases. Connections between properly folded protein, misfolded states, and protein aggregates have proven to be difficult to elucidate *in vivo* and have made designing effective treatments a difficult task.³

Early theoretical approaches to describing protein folding were focused on simple, phenomenological models aimed at capturing the relevant physics involved in the folding process.^{4–7} Coupled to results from lattice protein folding simulations, these models were quite successful in characterizing the general properties of the folding phase transition.^{8,9}

In the decades since the development of these theories, however, the exponential march of Moore's law has allowed rigorous simulation on biophysical systems of extraordinary extent to reach once unthinkable time scales. With respect to protein folding, distributed computing networks like Folding@Home, and specialized supercomputing hardware like Anton, have pushed the boundaries of fully atomistic molecular dynamics (MD) simulations applied to protein folding systems.^{10–13} The folding dynamics of numerous single-domain protein systems have now been elucidated, often near biological temperatures and in the presence of explicit solvent molecules. Using sophisticated statistical techniques like Markov state models (MSMs), protein simulations have been amassed to capture processes like folding and conformational change that occur on the order of tens of milliseconds in time scale.

While MD simulations have provided great insight into specific protein systems, such calculations have also informed general observations about the protein folding process that were not evident from early theories. Access to detailed information about kinetics has inspired descriptions of ubiquitous native hub-like network dynamics on folding landscapes, leading to a characterization of the role non-native contacts play in protein folding.^{2,14} The idea that non-native contacts have a generic impact on the folding process relates intimately to the risk of protein misfolding, as formation of contacts outside the native fold can promote transitions into long-lived “trap” conformations. Indeed, misfolded trap states have emerged as general features of protein folding landscapes studied with comprehensive MD. In particular, non-specific β -sheet rich states exhibiting little overlap with the native state have appeared in the dynamics of numerous

^{a)} Author to whom correspondence should be addressed. Electronic mail: pande@stanford.edu. Present address: James H. Clark Center, Stanford University, 318 Campus Drive, Stanford, California 94305, USA.

systems.^{10,12} Given their apparent homology to intermolecular amyloid structures, these β -rich states have been dubbed “intramolecular” amyloid conformations, owing to their occurrence on the single molecule level.^{12,15} Subsequent experiments have provided data consistent with the existence of intramolecular amyloids in systems comparable to those simulated.¹⁵ While some evidence suggests these states can be highly stable, their full effects on protein folding and misfolding processes remain to be elucidated.¹⁶ Notions of the intramolecular amyloid are largely absent from traditional theories of protein folding. Can we revisit old theoretical approaches and find hallmarks of these novel states?

The class of models on which we will focus here concern “heteropolymer freezing,” the process by which a polymer composed of different monomer species (e.g., a protein) collapses onto its unique ground state in the thermodynamic limit.^{5,17} Models of heteropolymer freezing often take advantage of the physics of disordered systems, using disorder-averaged partition functions to derive phase transitions like those seen for spin glasses.¹⁸ In the case of heteropolymers, the randomness, or “disorder,” in the system arises from variability in sequence: differing monomer compositions and arrangements obviously lead to different physical behavior. Since polymer sequences do not change on the time scales of polymer dynamics (but might rather change on much slower evolutionary time scales), we say that this disorder is “quenched,” and we avail ourselves of rigorous techniques like the replica trick to compute disorder-averaged quantities.¹⁸

The traditional heteropolymer freezing models discussed here start with a simple coarse-grained representation of monomers that interact in a pairwise fashion through a contact energy matrix.^{5,17} Such models usually make high temperature and mean-field approximations, consider only maximally compact, uniform globules, and assume the validity of the random energy model (REM) (i.e., that the energies of different conformations are uncorrelated). Often, models of heteropolymer freezing include a “sequence design” element, wherein sequences that have low energies in a given “target” conformation are somehow preferentially selected.^{19,20} This feature is important for application to protein systems, as proteins have unique native states with structures determined by evolution. Introduction of sequence design gives rise to two separate phase transitions in heteropolymer freezing: a generic “freezing” transition into a non-specific ground state, and a “folding” transition that represents collapse onto the target conformation.

In this paper, we aim to find the hallmarks of the intramolecular amyloid state in models of heteropolymer freezing. To do so, we need to characterize the effects of secondary structure on the phase transitions that occur within the system. Established approaches, however, rarely include explicit descriptions of secondary structure, but rather assume that relevant effects are captured through the coarse-graining process.²¹ Some models treat secondary structure by including a propensity to extend the compact polymer globule; more often, however, secondary structural properties are encoded in an entropic polymer stiffness parameter, s .^{17,21,22} To account for secondary structural effects, the entropic penalty for

bending is simply increased (i.e., the polymer is made to be uniformly stiffer). Estimates from simulation suggest that s should be reduced by a factor of four from the fully flexible polymer case to match data for polymers with secondary structure.⁹

Secondary structural elements, however, are not “uniformly stiff,” but have elastic properties related to the specific characteristics of their formation. Such variance in stiffness has important implications for long-range contact formation: using proteins as an example, long-range contacts are much more easily formed within a β -sheet motif than within a given stretch of α -helix. Here, we posit that the important aspects of secondary structure can be captured with a “variable persistence length,” where some regions of the polymer chain are made harder to bend than others. Borrowing from elastic beam theory, we build a modified model of heteropolymer freezing wherein polymer bending incurs a quadratic energy penalty proportional to varying persistence lengths within the chain.²³ The entropic stiffness parameter, s , will be set at its fully flexible value, allowing for the effects of variable stiffness to be treated with an explicit energy penalty term in the Hamiltonian. Apart from this polymer stiffness modification, we will make many of the same approximations (high temperature, mean-field, maximally compact globules, etc.) featured in traditional heteropolymer freezing models. In parallel to purely contact-based models, one might expect this additional energy variance term to contribute to increasing the freezing temperature of the system; the extent of this contribution and the subtleties surrounding it, however, remain to be derived.

Below, we present the setup and outline the derivation of our modified model. We go on to characterize how the variable persistence length modification impacts the phase transitions present in the model, and we discuss the implications of these changes with respect to the inclusion of secondary structural effects in the Hamiltonian. Finally, we describe the possible evidence for the prominence of intramolecular amyloid states in protein folding landscapes that this theory might provide.

THE MODEL

Setup

To simplify descriptions of our expressions throughout this derivation, we adopt much of the notation of Pande *et al.*¹⁷ described in depth in the literature. In the context of that previous work, we begin our derivation with a zeroth-order Hamiltonian, $\mathcal{H}^0$, for a length- N heteropolymer sequence that includes only monomer-monomer contact energies, presented in the form of the interaction matrix Ω :

$$\mathcal{H}^0(\{s_I\}, \{\mathbf{r}_I\}) = \sum_{I < J}^N \Omega_{s_I, s_J} \Delta(\mathbf{r}_I - \mathbf{r}_J). \quad (1)$$

Here, capital Roman indices indicate a monomer’s position in the heteropolymer sequence. The terms s_I and $\mathbf{r}_I$ represent the chemical identity and position of monomer I , respectively. The sets $\{s_I\}$ and $\{\mathbf{r}_I\}$ thus indicate a specific polymer sequence and conformation over which the energy is evaluated. The function $\Delta(\mathbf{r})$ is a softened δ -function: $\Delta(\mathbf{r})$ is unity if $\mathbf{r}$

is less than some cutoff r_0 (i.e., if the residues are in contact), and zero otherwise. Importantly, we reiterate that capital Roman indices (I, J) represent positions along the polymer chain; small Roman indices (i, j , etc.) will later be used to denote monomer identities independent of chain connectivity.¹⁷

To introduce semi-flexibility into our model, we say that each monomer can assume one of two states, α or β , which are each defined by a different intrinsic stiffness. In quantifying stiffness, we say regions of the polymer in the α state have persistence length l_α , and those regions in the β state have persistence length l_β . To aid in energy evaluation, we introduce tangent vectors along the polymer chain, defined simply by $\mathbf{t}_I = \mathbf{r}_I - \mathbf{r}_{I-1}$. Borrowing from thin-beam elastic theory, we apply a quadratic energy penalty to bending the polymer chain:²³

$$\Xi(\{I\}, \{\mathbf{r}_I\}) = \sum_{I=3}^N \frac{\delta_{I,\alpha} l_\alpha + \delta_{I,\beta} l_\beta}{2} (\mathbf{t}_I - \mathbf{t}_{I-1})^2, \quad (2)$$

where $\delta_{I,\beta} = (1 - \delta_{I,\alpha})$. By inspecting Eq. (2), one finds that if the chain is nearly straight (i.e., $(\mathbf{t}_I - \mathbf{t}_{I-1})$ is small), the bending penalty over that section of polymer is negligible; the penalty is significant, however, if the tangent vectors point in different directions. The energy penalty is different, of course, depending on whether the monomer I is in the α or β state.

To transform Eq. (2) into a more vector-friendly expression, we introduce the energy penalty kernel, A :

$$A = \frac{\delta_{I,\alpha} l_\alpha + \delta_{I,\beta} l_\beta}{2} \times I^{(3)}, \quad (3)$$

where $I^{(3)}$ is a three-dimensional identity matrix. We can now express the term Ξ as a quadratic form in Dirac notation:

$$\Xi(\{I\}, \{\mathbf{r}_I\}) = \sum_{I=3}^N \langle \mathbf{t}_I - \mathbf{t}_{I-1} | A | \mathbf{t}_I - \mathbf{t}_{I-1} \rangle, \quad (4)$$

meaning our full system Hamiltonian is now given by

$$\mathcal{H}(\{s_I\}, \{\mathbf{r}_I\}, \{I\}) = \sum_{I < J}^N \Omega_{s_I, s_J} \Delta(\mathbf{r}_I - \mathbf{r}_J) + \sum_{I=3}^N \langle \mathbf{t}_I - \mathbf{t}_{I-1} | A | \mathbf{t}_I - \mathbf{t}_{I-1} \rangle. \quad (5)$$

Figure 1 provides a cartoon illustration of the physical effects encoded in this Hamiltonian. As dictated by the model, the heteropolymer is represented in a fully compact conformation, and monomers are coarse-grained to remove fine chemical details. At left in Figure 1, we illustrate a standard expression included in models of heteropolymer freezing: energies of conformations are determined in part by monomer-monomer contact energies, ϵ_{IJ} .

The second effect we include in our model, shown at right, is modulated by the variable persistence length. Each monomer unit is assigned one of two persistence lengths, indicated by the alternating thick and thin links of the chain. The energy penalty to bending into this conformation is determined by the degree of difference between tangent vectors along the chain, and is proportional to the persistence lengths, l_p , of individual links. One can imagine how this variance

FIG. 1. Cartoon illustration of energetic components included in the modified heteropolymer freezing Hamiltonian. The model is valid only for maximally compact polymer configurations, and monomers are coarse-grained to remove fine chemical detail. Left: the Hamiltonian contains the standard monomer-monomer contact energies, ϵ_{IJ} . Right: the modified Hamiltonian now includes a bending energy penalty term, where each component penalty, ϵ_p , is proportional to the persistence length of a given chain link.

in persistence length might impact long-range contact formation, as such contacts within “stiff” regions of the chain would be disfavored. In analogy to the range of monomer sequences we will consider in the model, persistence lengths will also be randomly distributed throughout given sequences and controlled in the same fashion as monomer identities.

Given a particular sequence, we can express its partition function simply by summing over Boltzmann weights of polymer conformations:

$$Z(seq) = \sum_{\{\mathbf{r}_I\}} \exp \left[-\frac{\beta}{2} \left(\sum_{I \neq J}^N \Omega_{s_I, s_J} \Delta(\mathbf{r}_I - \mathbf{r}_J) + \sum_{I=3}^N \langle \mathbf{t}_I - \mathbf{t}_{I-1} | A | \mathbf{t}_I - \mathbf{t}_{I-1} \rangle \right) \right]. \quad (6)$$

Model evaluation

With this sequence-dependent partition function in hand, the next steps in our derivation involve averaging over the possible sequences and so-called “target” conformations of our heteropolymer chain. The procedure for conducting these two averaging programs is somewhat technical, and the process used here largely follows the arguments given by Pande *et al.* in the derivation of their contact-based model of heteropolymer freezing.¹⁷ Owing to these facts, the full derivation for our modified model and its concomitant properties are given in the Appendix. In lieu of providing the technical details here, we will outline the qualitative arguments made throughout the averaging process, highlighting the differences brought about by the introduction of the variable persistence length.

In order to arrive at a description of the phase behavior that occurs in our model, our derivation proceeds in several steps. Completion of our averaging protocol will rely on powerful methods taken from vector field theory; as such, our first task entails converting the discrete sequence and secondary structural information into a continuous field

representation. To do so, we introduce two distinct vector fields: the q -dimensional “monomer density” field $\vec{\rho}$, where q is the number of monomer types available to the polymer, and the 2-dimensional “stiffness density” field, $\vec{\gamma}$, which accounts for the two possible persistence length values that can be taken by each segment of the chain. As indicated by Eqs. (A1) and (A2), these fields can simply collapse back onto their discrete representations through the evaluation of a set δ -functions.

After constructing our fields of interest, we next compute the properties of the model as an average over ensembles of possible polymer sequences. Since, in analogy to proteins, we do not expect our sequences to change on time scales relevant to polymer folding, we need to compute averages over the free energies defined by each fixed sequence:

$$F = -T \langle \log Z(seq) \rangle_{seq}. \quad (7)$$

We refer to this operation as “quenched” averaging, since the relevant degrees of freedom in the average are frozen throughout the process we choose to investigate. Here, we encounter the first consequence of our theory’s modification: just as with the sequence variables, the secondary structural characteristics of the chain are made to be quenched throughout the folding process. This feature amounts to making an extreme approximation regarding secondary structural propensities for each residue. In principle, these propensities could be controlled by the physical properties of the particular amino acids in each sequence. Incorporating such correlations between sequence and secondary structure, however, would greatly complicate the evaluation of the present model; we thus leave this more challenging adaptation for future work. Here, we reduce our model to include fixed probabilities for each secondary structural identity. This reduction will allow us to focus on the basic physics that variable stiffness introduces into the system, without yet delving into the subtleties of real protein structure.

Computing averages over log-sums like that shown in Eq. (7) is not trivial; to do so here, we invoke a method called the replica trick. Using the replica trick involves completing two central tasks: (1) performing an average over the Taylor expansion of a partition function exponent, and (2) equating that expression to the logarithm average through analytical continuation of the expansion parameter. This process can be carried out by interrogating powers of our sequence-dependent partition function, effectively creating “replicas” of the system for analysis. We can sum over the sequence-dependent variables in our expressions by making a simple high-temperature approximation, leaving a “quenched-averaged” partition function that describes the mean character of heteropolymers with quenched sequences.

It is important to note that the sequences considered in this average are not randomly constructed, nor are they restricted to any particular level of energetic optimization for a given target conformation of the polymer. Instead, we extract average properties at various degrees of sequence optimization using concepts from canonical sequence design. In analogy to the canonical ensemble, we introduce an artificial “design temperature,” T^{des} , that sets the natural scale for fluctuations in the types of sequences that contribute significantly

to the average. A low T^{des} will restrict our ensemble to sequences that have low energies in a chosen target conformation; the sequence ensemble at higher temperatures, by contrast, will be nearly random and largely independent of any target conformation. The parameter T^{des} will be juxtaposed with the physical temperature, T , in our discussion of the effects of sequence design on heteropolymer freezing. In our modified model, keep in mind that T^{des} controls design properties with respect to both the sequence and the new secondary structural identities. A detailed description of the replica trick, sequence averaging, and canonical sequence design can be found in the Appendix.

Once the average over sequences is completed, we direct our efforts toward elucidating the effects of different target conformations on the properties of our model. Protein native states are optimized by evolution to perform some specific mechanical task. Our model, of course, cannot incorporate the minutia of this sort of specific evolutionary design behavior. Therefore, we would like our model’s properties to be agnostic to any particular target conformation of the heteropolymer; satisfying this desire amounts to performing an average over possible target conformations.

To carry out this second average, we first need to study how the “replicas” of the system interact through particular non-physical couplings. Specifically, we need to scrutinize the extent to which these $(n + 1)$ replicas “overlap” under a natural measure of conformational similarity. Since we have two distinct vector quantities in our modified model, we define two so-called “replica overlap parameters,” $\hat{Q}$ and $\hat{W}$, which describe the extent of overlap between the sequence and stiffness density fields, respectively, of different replicas in the system. To average over the system’s possible conformations, it can be argued that one simply needs to integrate over the exponent of the model’s free energy as a function of $\hat{Q}$ and $\hat{W}$:

$$\langle Z^n \rangle = \int \mathcal{D}\{\hat{Q}\} \mathcal{D}\{\hat{W}\} \exp[-\mathcal{F}\{\hat{Q}, \hat{W}\}]. \quad (8)$$

Here, we add a detail important for the evaluation of our theory: in utilizing the free energy, we define an entropy within the model, $S\{\hat{Q}, \hat{W}\}$, which will oppose the “freezing” of the heteropolymer onto specific conformations (target or otherwise). The nature of these overlap parameters and their relation to the entropy and free energy are discussed in detail in the Appendix.

Completing this full integral over the free energy exponent, however, is also not a trivial task. To make its evaluation tractable, we first make a mean field approximation by converting the integral to its saddle point expression. We then employ a variational approach to carry out this saddle point integration, using arguments from polymer theory to understand the scaling of the free energy with increased conformational overlap. In the thermodynamic limit, we find that replicas either overlap completely, or not at all. This realization yields a simple binary matrix that describes the overlaps between replicas in our system. To reach this point, we make one additional approximation based on our modified Hamiltonian: we reasonably assume that when two conformations overlap under $\hat{Q}$ (a measure based on monomer positions), they also

overlap under $\hat{W}$ (a measure based on tangent vectors along the chain).

The process that occurs with this saddle point evaluation is referred to as “replica symmetry breaking,” wherein replicas collapse in groups onto particular polymer conformations. The sizes of these groups are balanced by the energies and entropies of the replicas’ interactions.

With this matrix of simple overlap parameters in hand, we say that our averaging procedure over possible target conformations is complete. The final step in our derivation involves optimizing the model’s free energy with respect to these conformational overlap parameters to extract the theory’s equilibrium behavior. To do so, we introduce two order parameters, x and y , which indicate the extent to which replicas collapse onto either “spontaneous” or “target” ground state conformations, respectively. In practice, y indicates the number of replicas that collapse onto the system’s target conformation.

With these two order parameters in place, we expect to see phase transitions from our model’s random heteropolymer globule into two distinct “frozen” phases. The y parameter, which indicates how many replicas coincide with the system’s target conformation, describes something akin to the protein folding phase transition to the native state. By contrast, the x parameter describes how replicas might spontaneously group onto an arbitrary ground state conformation: when x becomes smaller than unity, replicas start to collapse onto non-specific conformations. Since these conformations have no formal relationship with the target conformation, this phase transition represents something like a generic glass transition, wherein the system becomes trapped in stable, misfolded states.

As illustrated in the Appendix, it is simple, though somewhat laborious, to write expressions for the energy and entropy of the model in terms of these order parameters. Truncating the free energy to quadratic order and taking derivatives with respect to x and y individually, we derive our phase coexistence equations for the model:

$$s = \frac{x^2}{4T^2} (\delta\Omega^2 + \delta A^2), \quad (9)$$

$$s = \frac{x}{2T T_{\text{des}}} (\delta\Omega\delta\Omega_{\text{des}} + \delta A\delta A_{\text{des}}) - \frac{x^2}{4T^2} (\delta\Omega^2 + \delta A^2). \quad (10)$$

Here, s is the microscopic entropy parameter discussed in the Introduction, and $\delta\Omega$ and δA (and their sequence design counterparts) represent the variances of our two interaction kernels. The expressions for these variance terms are rigorously defined in the Appendix.

We see the full effects of our variable persistence length modification in Eqs. (9) and (10): the stiffness kernel, A , contributes to the folding and freezing transitions in equal measure with the contact kernel, Ω . An analysis of these equations yields the expressions needed to draw a phase diagram for our modified model of heteropolymer freezing.

RESULTS AND DISCUSSION

Equations for phase diagram

In deriving our phase diagram, we first look to the freezing transition between random globules and a glassy ground state, described by Eq. (9). Recall that spontaneous symmetry breaking will occur when x starts to deviate from unity, causing replicas to group onto non-specific conformations. The critical temperature, or freezing temperature, can thus be obtained by setting $x = 1$ in Eq. (9) and solving for T :¹⁷

$$T_{\text{freeze}} = \sqrt{\frac{(\delta\Omega^2 + \delta A^2)}{4s}}. \quad (11)$$

It is clear that x will decrease linearly with T below the freezing temperature, implying the phase transition is first-order in nature. In general, the value of x is related to the number of thermodynamically relevant conformations at equilibrium; as re-derived by Pande, Grosberg, and Tanaka, the number of relevant microstates quickly decays to $\mathcal{O}(1)$ below the freezing temperature.^{17,18} Notably, Eq. (9) is independent of the design temperature, T^{des} , which governs the extent to which an ensemble of sequences is optimized for the target conformation.

To investigate the effect of T^{des} , we first recast Eq. (9) in terms of the design temperature:¹⁷

$$T_{\text{freeze}}^{\text{des}} = \sqrt{\frac{(\delta\Omega_{\text{des}}^2 + \delta A_{\text{des}}^2)}{4s}}, \quad (12)$$

where we have explicitly indicated that the design interaction matrices, Ω_{des} and A_{des} , are responsible for the freezing transition. The second freezing transition (caused by collapse onto the target replica), which we will call the “folding” transition, is governed by the additional order parameter y . Equation (10) indicates the conditions under which the value of y switches from 0 to n , i.e., when the system collapses onto the target replica. Reordering Eq. (10) (and keeping $x = 1$), we find the folding temperature to be¹⁷

$$\frac{T_{\text{fold}}^{\text{des}}}{T_{\text{freeze}}^{\text{des}}} = \begin{cases} \frac{2T/T_{\text{freeze}}}{1 + T^2/T_{\text{freeze}}^2} & \text{for } T \geq T_{\text{freeze}}, \\ 1 & \text{for } T \leq T_{\text{freeze}}. \end{cases} \quad (13)$$

Colloquially, $T_{\text{fold}}^{\text{des}}$ represents the value of the design temperature below which sequences become good enough to make the target conformation the system’s ground state. Equation (13) indicates that transitions between the “glassy” and “folded” phases are forbidden: below the freezing temperature, the critical line between the two phases becomes independent of temperature.^{17,18,27} Since T^{des} is a quenched property, one cannot drive a given system between the folded and glassy phases.

If we assume T and T^{des} are close to the tricritical point between the random, glassy, and folded phases, we can rearrange Eq. (13) to yield a simpler, approximate expression for the folding temperature:¹⁷

$$\frac{T_{\text{fold}}^{\text{des}}}{T_{\text{freeze}}^{\text{des}}} = 1 + \sqrt{2} \left[1 - \frac{T^{\text{des}}}{T_{\text{freeze}}^{\text{des}}} \right]^{1/2}. \quad (14)$$

FIG. 2. Phase diagram for unmodified model for heteropolymer freezing, derived from Eqs. (13) and (14). Sequence design introduces three distinct phases into the model: a random phase (with $\mathcal{O}(\exp(N))$ relevant conformations); a glassy, misfolded phase (with $\mathcal{O}(1)$ relevant conformations); and a target, folded phase (also with $\mathcal{O}(1)$ relevant conformations).

Figure 2 illustrates the phase diagram implied by the Eqs. (13) and (14).¹⁷ Note that T_{fold} is higher than T_{freeze} below the critical design temperature: we see a significant gap between the folding of designed and random sequences. This gap represents the advantage gained in designing sequences to fold into a particular target conformation.

Impact of the variable persistence length

At this point, we have yet to emphasize how the inclusion of a variable persistence length impacts heteropolymer freezing phase behavior. Indeed, Eqs. (13) and (14), which describe the shape of the folding coexistence curve, are perfectly valid for traditional heteropolymer freezing models in which only tertiary contact energies are taken into account.

As indicated by Eqs. (11) and (12), however, the persistence length-based energy penalty adds a second term that contributes to the freezing temperature. Given that Ω represents a matrix of tertiary contact energies (like the Miyazawa-Jernigan matrix), the freezing temperature in traditional models is set by the variance in Ω .²⁸ This variance in contact energies corresponds to the heteropolymer effects important in the model: the differences in how different monomers interact determine the temperature at which heteropolymers freeze.

In our case, the freezing temperature is also affected by variance in the polymer stiffness, and these two variance terms add in quadrature. Using Eqs. (11)–(14) and the definition of A (Eq. (3)), we can derive an expression for δA^2 :

$$\delta A^2 = \sum_{lk} p_l \left(\frac{l_l - l_k}{4} \right)^2 p_k. \quad (15)$$

This variance term, of course, must be greater or equal to zero for any values of the two persistence lengths. The introduction of a persistence length difference in the model, therefore, appears to have the general effect of increasing the heteropolymer freezing temperature.

It is interesting that the simple addition of stiffer and floppier segments in the polymer makes both generic freezing and folding easier to accomplish. The variable persistence length adds a second energy variance term to the system's Hamiltonian; the variation in stiffness, just like the variation in contact energies, corresponds in turn to a “heteropolymeric” effect, wherein increased energetic flexibility contributes favorably to freezing properties. In principle, any such energy term with non-zero variance, as long as it can be evaluated as a quadratic form, could contribute to an increased freezing temperature to some extent. This observation suggests that generic variations in energetic interactions can be used by evolution in concert to improve the optimization of heteropolymer freezing properties.

Just how much would this persistence length variation impact models for real proteins? If we interpret l_α and l_β as the persistence lengths for α -helices and β -sheets, the difference in stiffness between the two structures is well known. In general, α -helices are more difficult to bend, prohibiting the formation of long-range tertiary contacts within stretches of helical residues. Computational and experimental estimates for the α -helix persistence length are on the order of ≈ 20 nm.²⁹ By contrast, the turns in β -sheets occur on the order of the persistence length of the polymer itself (≈ 1 nm), and large register shifts allow for long-range contacts between almost any two residues.²⁹

To redefine the persistence length on a natural energy scale, we say $l_p = A/L \times kT$, where A is the persistence length in nm, L is a “minimum bending radius” for the polymer, in nm, and kT defines the thermal energy scale.²³ Under this definition, one might say $l_\alpha \approx 20$ kT and $l_\beta \approx 1$ kT. Letting β -sheets and α -helices be equally probable, we can plug these values into Eq. (A39) (in the Appendix) to estimate an energy variance contribution

$$\sqrt{\delta A^2} \approx \sqrt{0.5 \left(\frac{l_\alpha - l_\beta}{4} \right)^2} \approx 3.5 kT. \quad (16)$$

For reference, the variance of the Miyazawa-Jernigan matrix is about 2 kT.²⁸ The effects of persistence length variance on heteropolymer freezing, thus, could be just as or even more pronounced than those derived from contact energy variance.

We have mentioned previously that the parameter s that appears in the freezing temperature is an entropic term related to polymer stiffness.^{9,21} For this work, we have set $s \approx 1.6$, near its standard value for flexible polymers, and we have rigorously treated differences in persistence length in terms of energetic penalties.¹⁷ However, using simulations of lattice models, Pande *et al.*¹⁷ have estimated a value for s in which secondary structure is taken into account: $s \approx 0.4$. If we reframe our energetic contribution to T_{freeze} as an entropic restriction, we see the following:

$$\begin{aligned} T_{\text{freeze}} &= \sqrt{\frac{(\delta\Omega^2 + \delta A^2)}{4s}} \approx \sqrt{\frac{(2)^2 + (3.5)^2}{4(1.6)}} \\ &\approx \sqrt{\frac{(4 \times \delta\Omega^2)}{4s}} = \sqrt{\frac{\delta\Omega^2}{4s'}}, \end{aligned} \quad (17)$$

where $s' \approx 0.4$. Thus, our inclusion of persistence length-based secondary structure suggests a rescaled s -parameter in near perfect agreement with previous computational predictions. Since some of the above calculations are no doubt imprecise, this exact agreement should be interpreted with measured enthusiasm. Nonetheless, this general consistency with previous work is encouraging.

Emergence of “intramolecular” amyloids

Figure 3 shows an overlay of phase diagrams for the traditional freezing model ($\delta\Omega^2 \approx 4$) and our model under the conditions described above ($\delta A^2 \approx 12$, $\delta\Omega^2 \approx 4$). We see that two new “folding” and “freezing” gaps emerge on the diagram where the inclusion of persistence length variance has driven up the freezing temperature. Based on the nature of the bending energy penalty, one might surmise that the emergence of these gaps is driven by conformations rich in β -secondary structure, which suffer a lesser cost to bending into the maximally compact globule. While regions with high stiffness might play a structural role in the target conformation, the exceptionally large energetic price to compacting stretches of α -content would likely outpace the entropic loss suffered with β -sheet formation in generic conformations of real proteins. It is also important to remember that, in real protein systems, formation of α -helices pays its own entropic price compared to the fully flexible chain. Considering these arguments, the new folding and freezing gaps are labelled as the “ β -induced folded region” and the “ β -induced misfolded region” in Figure 3. States in the misfolded region, in particular, might be expected to be dominated by β -structure, as no bias toward a specific (and possibly α -containing) target conformation is felt at high design temperatures.

We should note that under this model, the folded and misfolded regions do not interconvert for specific sequences at

FIG. 3. Phase diagram for secondary structure-modified version of our heteropolymer freezing model (derived from Eqs. (11)–(14)), overlaid on the phase diagram for the unmodified system. Inclusion of persistence length-based secondary structure causes both the freezing temperature, T_{freeze} , and the folding temperature, T_{fold} , to increase, introducing freezing and folding “gaps” in the modified model. The larger extent of the folding gap suggests an evolutionary reason for risking misfolding at higher temperatures.

equilibrium and in the thermodynamic limit. Real proteins, however, are finite in size and display heterogeneous behavior in their dynamics.¹⁷ In physical systems that go beyond the REM, thus, one might observe slow interconversion between the folded, target region and the β -induced misfolded region, a kinetic process analogous to misfolding seen for real proteins.

We must also mention that, formally, persistence lengths are quenched within particular polymers in our model, meaning that a polymer’s secondary structure rigorously cannot change throughout the freezing process. However, given that our theory assumes the validity of the REM, the limitations associated with quenched disorder can also be avoided to some extent.¹⁷ Since we assume that different polymer conformations are uncorrelated in energy, we can speculate that some “cross-talk” will occur in non-REM polymers exhibiting different secondary structural characteristics. In fact, it can be shown that the properties of the model under quenched disorder (in which structural parameters are fixed) and annealed disorder (in which structural parameters can change) are equivalent under the REM.¹⁷ When discussing our model in the context of real systems, these caveats need to be taken into account; nonetheless, the REM affords us with some flexibility when conjecturing about the effects of annealed secondary structure.

The discussion of β -rich phases in heteropolymers brings to mind a common structural motif in protein aggregates: the intermolecular amyloid. In intermolecular amyloids, many molecules are woven together into β -sheet structures that reach macroscopic sizes. The implications of β -amyloid structures in a number of neurological diseases are well documented.³

In this model, however, we suggest that a β -rich misfolded phase might appear for the conformations of single heteropolymer globules. This observation aligns nicely with the notion of the “intramolecular” amyloid, a phrase coined by Gruebele and co-workers to describe β -rich traps that appear in the simulated folding landscape of a λ -repressor mutant.^{12,15} As noted in the Introduction, Prigozhin *et al.* have recently identified slow phases in λ -repressor dynamics that could be consistent with conversion between these amyloid-like traps and folded structures.¹⁵

If the β -rich misfolded phase in our theory truly aligns well with the observations of intramolecular amyloids, we can provide some insight into the fundamental driving force for intramolecular amyloid formation. With the introduction of two different persistence lengths, both folding transitions and freezing transitions occur at higher temperatures; in some sense, both folding and non-specific freezing become “easier.” However, based on the shape of the folding coexistence curve, the new folding gap that emerges is larger than the new generic freezing gap. While both folding and misfolding become easier, the folding process is given the greater enhancement.

As an evolutionary gamble, it may well be worthwhile to optimize folding at an increased risk of misfolding into β -sheet rich traps. Sequence design for conformations with both stiff and fluid structural elements appears to be much easier than for uniform heteropolymers: sequence-conformation

pairs that allow for variable persistence lengths are “more designable” through the lens of our theory, as folding occurs at higher design and physical temperatures.

The cost of taking this misfolding risk, of course, lies in the consequences of intermolecular amyloid formation. In finite systems out of equilibrium, one can expect interconversion between the target and misfolded phases to occur on the folding landscape. While intramolecular amyloid structures might be able to interconvert with the native state at near-folding time scales, intermolecular amyloids are essentially indestructible on biological time scales. One might imagine that these intramolecular amyloids have the potential to be “amyloidogenic,” serving as seeds for the formation of their intermolecular counterparts. In rare instances, these intermolecular amyloids can grow to dominate the folding landscape and wreak havoc on processes needed to sustain important biological functions.

Though our model is interpreted in the thermodynamic limit, it may be possible to use finite size scaling arguments to study how a polymer’s length might modulate its freezing properties. Even beyond the random energy model, the degree to which the crossovers between random globule and frozen phases resemble first order phase transitions should scale approximately linearly with chain length. Shorter polymers should thus exhibit more “flexible” target states, as transitions between the folded and unfolded basins would be relatively broad and smooth. The same arguments hold for a shorter polymer’s misfolded ground state conformations, implying that the shorter polymer’s risk of getting trapped in amyloidogenic states is reduced. In longer polymers, these crossovers are sharper and more pronounced, and the folded and misfolded phases of the model each become more stable.

Given that larger proteins are more “difficult” to fold in terms of their increased native topological complexity and entropic constraints, it may not be unreasonable to speculate that larger proteins rely to a greater extent on the folding bonus introduced by a variable persistence length. If larger proteins utilize a more diverse secondary structural ensemble to optimize the folding process, our model suggests that the risk of misfolding is thereby also heightened. This observation may hint at why many larger proteins require chaperones to facilitate folding: though greater variability in secondary structure (and thus persistence length) might make folding easier, the increased risk of misfolding this variability brings could force the system to be more reliant on cellular machinery that rescues molecules from misfolded traps.

CONCLUSION

In summary, we have modified a simple model of heteropolymer freezing to include a persistence length-based representation of secondary structure, and in characterizing that system’s properties we have found intriguing connections between a variable persistence length and enhanced folding behavior in designed sequences. Most notably, we find that a variable persistence length increases the folding and generic freezing temperatures in the model, making both folding and misfolding processes better optimized in that respect. The de-

rived effects of secondary structure agree well with previous estimates from computational work, and suggest an interpretation of our results through the lens of evolution and the pathology of amyloid-like structures.

In evaluating our theory, it is important to remember that our model only rigorously describes heteropolymers of a certain type; the present theory cannot capture all the subtleties of real protein systems. While we have argued that the variable persistence length represents an important characteristic of protein secondary structure, certainly other factors come into play. Perhaps most dramatically, our model treats sequence and secondary structural variables as independent degrees of freedom. In real proteins, however, secondary structure is to a large degree determined by the sequence: particular arrangements of side chains can obviously help and hinder the formation of hydrogen bonds along the protein backbone. Residues like glycine and proline are known to severely inhibit the propagation of α -helices and can facilitate β -hairpin formation. As noted earlier, simple residue-specific effects might be incorporated into our model by explicit inclusion of secondary structural propensities for individual monomers. Such adaptations are certainly worthy of development and appraisal in future work. We can speculate that these effects would add new energy variance terms to the Hamiltonian (and thus impact freezing properties); just how drastic an effect these terms would have remains to be seen. Based on the present theory’s agreement with computational predictions, however, it seems that inclusion of a simple variable persistence length represents a good first step toward understanding the effects of secondary structure in heteropolymer systems.

We can thus argue that our current theory can provide some useful insight into the behavior of real proteins. As we noted in the Introduction, the motivation for completing this theoretical work came in large part from observations taken from simulations and experiments. Given the theoretical results presented in this paper, can we in turn offer any testable predictions to the experimental and simulation communities?

At the very least, our results suggest that “intramolecular amyloid” formation in heteropolymers may be a more general physical phenomenon than previously thought. We propose that the search for such structures in simulation, and their hallmarks in experiment, should be extended to a broader class of systems.

Our results also harken to notions of “glassiness” in protein folding landscapes. Recall that two important phase transitions are observed in models of heteropolymer freezing: a collapse onto the target conformation (the “folding” transition) and collapse onto a non-specific globular conformation (the “freezing,” or “glass,” transition).²⁷ This glassy phase emerges from the ruggedness of energetic landscapes in random or poorly designed sequences, wherein barriers between competing ground states are difficult to traverse on relevant time scales.

While proteins, of course, have been subjected to an evolutionary design procedure, ideas of glassiness on protein folding landscapes are manifest in long-lived “trap” states that can result in misfolding and aggregation. The theory presented here suggests a driving force for these glassy trap

states to form and to be rich in β -sheet structure. Recent perturbative techniques on master equation-based simulation models have shown promise for probing glassiness in protein folding dynamics.¹⁶ Such techniques might be used further to characterize the structure of glassy states on protein folding landscapes and to test the ubiquity of β -rich trap states in protein biology.

ACKNOWLEDGMENTS

We thank National Science Foundation (NSF) (MCB-0954714) and National Institutes of Health (NIH) (R01-GM062868) for their support of this work. J.K.W. was supported by the Fannie and John Hertz Foundation on the endowed Professor Yaser S. Abu-Mostafa Fellowship.

APPENDIX: FULL DERIVATION OF REPLICA FIELD MODEL

The full derivation for the theory presented in this article is included in the text below. The derivation starts with our transition to field coordinates, details the procedures used for averaging over sequences and conformations, and illustrates the free energy optimization that allows us to arrive at the phase properties for our model.

1. Transition to field coordinates

In evaluating the partition function given in Eq. (6), it will prove useful to transform our discrete coordinates $\{\mathbf{r}_I\}$ into two specific continuous vector fields. We will call these fields (1) the monomer density field, $\tilde{\rho}_i(\mathbf{R})$, and (2) the “stiffness” density field, $\tilde{\Upsilon}_z(\mathbf{T})$. We define these fields as follows:¹⁷

$$\tilde{\rho}_i(\mathbf{R}) \rightarrow \sum_{I=1}^N \delta(s_I, i) \delta(\mathbf{r}_I - \mathbf{R}), \quad (\text{A1})$$

$$\tilde{\Upsilon}_z(\mathbf{R}) \rightarrow \sum_{I=3}^N \delta(ss_I, z) (\mathbf{t}_I - \mathbf{t}_{I-1}) \delta((\mathbf{t}_I - \mathbf{t}_{I-1}) - \mathbf{T}). \quad (\text{A2})$$

The spaces denoted by $\mathbf{R}$ and $\mathbf{T}$ are conceptually similar and in practice identical; we distinguish between the two separate volumes merely to remind the reader that $\tilde{\Upsilon}$ exists in a reduced tangent vector space. Our sequence-specific partition function is given in this field representation by

$$\begin{aligned} Z_{seq} = \sum_{\{\mathbf{r}_I\}} \exp \left[-\frac{\beta_k}{2} \int d\mathbf{R}_1 d\mathbf{R}_2 \sum_{i,j}^q \tilde{\rho}_i(\mathbf{R}_1) \right. \\ \times (\Omega_{i,j} \times \delta(\mathbf{R}_1 - \mathbf{R}_2)) \tilde{\rho}_j(\mathbf{R}_2) \\ \left. + \int d\mathbf{T}_1 d\mathbf{T}_2 \sum_{y,z}^2 \Upsilon_y(\mathbf{T}_1) (A_{y,z} \times \delta(\mathbf{T}_1 - \mathbf{T}_2)) \Upsilon_z(\mathbf{T}_2) \right], \end{aligned} \quad (\text{A3})$$

where we sum over the q possible monomer species and the 2 possible secondary structure identities. Here, the operator A has been recast as a $3 \times 3 \times 2$ -dimensional tensor, with

each 3×3 sub-matrix holding either all α -dependent or all β -dependent terms. For simplicity, we will transform this expression into a generalized vector space notation, introducing the operator $\hat{I}^{(\infty)}$ to represent the identity in an infinite Hilbert space of functions of $\mathbf{r}$:

$$\begin{aligned} Z_{seq} = \sum_{\{\mathbf{r}_I\}} \exp \left[-\frac{\beta_k}{2} (\langle \tilde{\rho} | \hat{\Omega}^{(q)} \otimes \hat{I}^{(\infty)} | \tilde{\rho} \rangle \right. \\ \left. + \langle \tilde{\Upsilon} | \hat{A}^{(3 \times 2)} \otimes \hat{I}^{(\infty)} | \tilde{\Upsilon} \rangle) \right]. \end{aligned} \quad (\text{A4})$$

Operators in this Hilbert representation are now topped with the “hat” notation; parenthetical superscripts indicate the dimensionality of the operator in question. The “ $\otimes$ ” operator indicates that a “direct product” is to be taken between the two surrounding operators. Given two arbitrary operators B^u and G^v , $B^u \otimes G^v$ yields an $(u \times v)$ -dimensional operator whose elements are the components of B projected onto the full G matrix.¹⁷ The implied sum of bracket notation in Eq. (A4), coupled with the $\otimes$ operator, means that the integrals in Eq. (A3) are represented as matrix elements over the fields and operators.

2. Replica trick

We are now ready to analyze this partition function in the context of heteropolymer design. To do so, we introduce a particular “target” conformation, labelled $\star$, in which sequence energy will be minimized. To obtain a $\star$ -dependent free energy for the model, we simply average the free energy over all sequences when folded into the target conformation.^{17–20} We need to take a step further, however, and compute the free energy not just for one conformation, but over all possible target conformations $\star$. We thus need to carry out the average

$$F = -T \langle \log Z(seq) \rangle, \quad (\text{A5})$$

where the brackets indicate averaging over both sequence and target conformation.

Computing averages over log-sum quantities is no simple task. To do so here, we employ the so-called “replica trick,” using the following identity:

$$\langle \log Z \rangle = \lim_{n \rightarrow 0} \frac{\langle Z^n \rangle - 1}{n}. \quad (\text{A6})$$

Employing this identity, our problem is reduced to averaging the partition function, $Z(seq)^n$, for a system containing $n + 1$ copies, or “replicas,” of the original system. Once this average quantity is in hand, we can use techniques from analytical continuation to study the $n = 0$ limit.¹⁸

Introducing $n + 1$ system replicas, and replica indices κ and γ , we calculate our average of $Z(seq)^n$ using the following general formula:

$$\begin{aligned} \langle Z(seq)^n \rangle = \sum_{\text{sequence}} \prod_{I=1}^N p_{s_I} (p_{\alpha_I} \delta_{I,\alpha} + p_{\beta_I} \delta_{I,\beta}) \sum_{C_0, \dots, C_n} \\ \times \exp \left[-\frac{1}{2} \sum_{\kappa=0}^n \beta_{k,\kappa} \mathcal{H}_\kappa(C_\kappa, seq) \right], \end{aligned} \quad (\text{A7})$$

where the sum over C_i runs over all possible compact conformations of each replica. We designate the $\kappa = 0$ replica as the “target” replica, for which $\beta_{k,\kappa} = \beta_k^{\text{des}}$, the inverse design temperature, and $C_0 = \star$, the target conformation. Here, we have introduced the concept of canonical sequence design, wherein the parameter $T^{\text{des}} = 1/\beta^{\text{des}}$ acts as an artificial temperature for selecting sequences relative to their energies in the target conformation.^{19,20} Sequence ensembles at high T^{des} are nearly random, while sequences that appear at low T^{des} have been subjected to a higher degree of optimization. The design temperature is plotted as a parameter alongside the physical temperature in the body of this work. Note further that the sum over C_0 in Eq. (A7) runs over all possible target conformations. For replicas that are not the design replica, $\beta_{k,\kappa}$ is simply the inverse temperature, β_k .¹⁷

In our vector space notation introduced above, the averaged partition function now reads

$$\begin{aligned} \langle Z(seq)^n \rangle = & \sum_{\text{sequence}} \prod_{l=1}^N p_{s_l} (p_{\alpha_l} \delta_{l,\alpha} + p_{\beta_l} \delta_{l,\beta}) \sum_{\{\mathbf{r}_l^\kappa\}} \\ & \times \exp \left[-\frac{1}{2} \left(\sum_{\kappa,\gamma=0}^n \langle \vec{\rho}^\kappa | \hat{\Omega}^{\kappa\gamma(q)} \otimes \hat{I}^{(\infty)} | \vec{\rho}^\gamma \rangle \right. \right. \\ & \left. \left. + \langle \vec{\Upsilon}^\kappa | \hat{A}^{\kappa\gamma(3 \times 2)} \otimes \hat{I}^{(\infty)} | \vec{\Upsilon}^\gamma \rangle \right) \right], \quad (\text{A8}) \end{aligned}$$

where we have pulled the temperature inside the operators $\hat{\Omega}$ and $\hat{A}$:

$$\hat{\Omega}_{ij}^{\kappa\gamma} = \beta_k^{\text{des}} \Omega_{ij}^{\text{des}} \delta_{\kappa 0} \delta_{0\gamma} + \beta_k \Omega_{ij} (\delta_{\kappa\gamma} - \delta_{\kappa 0} \delta_{0\gamma}), \quad (\text{A9})$$

$$\hat{A}_{ij}^{\kappa\gamma} = \beta_k^{\text{des}} A_{ij}^{\text{des}} \delta_{\kappa 0} \delta_{0\gamma} + \beta_k A_{ij} (\delta_{\kappa\gamma} - \delta_{\kappa 0} \delta_{0\gamma}). \quad (\text{A10})$$

We are now ready to sum over possible sequences (both with regard to monomer identity and stiffness parameters). In order to perform this sum, we use a common trick in field theory: we carry out a so-called Hubbard-Stratonovich transformation on the field variables $\vec{\rho}$ and $\vec{\Upsilon}$. Introducing the fields $\vec{\phi}$ and $\vec{\psi}$ conjugate to $\vec{\rho}$ and $\vec{\Upsilon}$, we invert a Gaussian functional integral to find¹⁷

$$\begin{aligned} \langle Z(seq)^n \rangle_{seq} = & \sum_{\{\mathbf{r}_l^\kappa\}} \int \mathcal{D}\{\vec{\phi}\} \mathcal{D}\{\vec{\psi}\} \\ & \times \exp \left[\frac{1}{4} (\langle \vec{\phi} | (\hat{\Omega}^{-1})^{(q(n+1))} \otimes \hat{I}^{(\infty)} | \vec{\phi} \rangle \right. \\ & \left. + \langle \vec{\psi} | (\hat{A}^{-1})^{((3 \times 2)(n+1))} \otimes \hat{I}^{(\infty)} | \vec{\psi} \rangle) \right] \\ & \times \sum_{\text{sequence}} \prod_{l=1}^N p_{s_l} (p_{\alpha_l} \delta_{l,\alpha} + p_{\beta_l} \delta_{l,\beta}) \\ & \times \exp[\langle \vec{\phi} | \vec{\rho} \rangle^{q(n+1)\infty} + \langle \vec{\psi} | \vec{\Upsilon} \rangle^{((3 \times 2)(n+1)\infty)}], \quad (\text{A11}) \end{aligned}$$

where replica indices are omitted for simplicity and summation over replicas is implied by the $(n+1)$ superscripts on

various quantities. The purpose of this transformation was to “decouple” monomer and stiffness densities: Eq. (A11) is independent of $\vec{\rho}$ and $\vec{\Upsilon}$ except for the linear “source” terms $\langle \vec{\phi} | \vec{\rho} \rangle$ and $\langle \vec{\psi} | \vec{\Upsilon} \rangle$. Instead of coupled monomers, however, we have introduced a new coupling of replicas through the fields $\vec{\phi}$ and $\vec{\psi}$. We will discuss the consequences of the new replica-space coupling later.¹⁷

For our summation over sequences, however, we now just need to manipulate the two linear source terms. We can write the source terms as

$$\begin{aligned} \exp\{\text{source}\} = & \prod_{l=1}^N \sum_{i=1}^q \sum_{z=1}^2 p_i p_z \exp[\langle \vec{\phi} | \vec{\rho} \rangle^{q(n+1)\infty} \\ & + \langle \vec{\psi} | \vec{\Upsilon} \rangle^{(3 \times 2)(n+1)\infty}] \\ = & \prod_{l=1}^N \sum_{i=1}^q \sum_{z=1}^2 p_i p_z \exp \left[\sum_{\kappa=0}^n \int d\mathbf{R} \vec{\phi}_i^\kappa(\mathbf{R}) \delta(\mathbf{r}_l^\kappa - \mathbf{R}) \right. \\ & \left. + \int d\mathbf{T} \vec{\psi}_z^\kappa(\mathbf{T}) (\mathbf{t}_l^\kappa - \mathbf{t}_{l-1}^\kappa) \delta((\mathbf{t}_l^\kappa - \mathbf{t}_{l-1}^\kappa) - \mathbf{T}) \right]. \quad (\text{A12}) \end{aligned}$$

To make the integrals in Eq. (A11) tractable, we need to make a simplifying expression to Eq. (A12). We perform Taylor expansions on the source exponential in $\vec{\phi}$ and $\vec{\psi}$, and we truncate the series at $\mathcal{O}(\vec{\phi}^2)$ and $\mathcal{O}(\vec{\psi}^2)$. This high-temperature approximation leaves us with the following source term expression:

$$\begin{aligned} = & \sum_{i=1}^q \sum_{z=1}^2 \sum_{\kappa=0}^n \int d\mathbf{R} p_i \vec{\rho}_\kappa \vec{\phi}_i^\kappa + \int d\mathbf{T} p_z \vec{\Upsilon}_\kappa \vec{\psi}_z^\kappa \\ & + \frac{1}{2} \sum_{i,j=1}^q \sum_{y,z=1}^2 \sum_{\kappa,\gamma=0}^n \left[\int d\mathbf{R}_1 \int d\mathbf{R}_2 [p_i \delta_{ij} - p_i p_j] \right. \\ & \times (\vec{\phi}_i^\kappa(\mathbf{R}_1) Q^{\kappa\gamma}(\mathbf{R}_1, \mathbf{R}_2) \vec{\phi}_j^\gamma(\mathbf{R}_2)) \\ & + \int d\mathbf{T}_1 \int d\mathbf{T}_2 [p_y \delta_{yz} - p_y p_z] \\ & \left. \times (\vec{\psi}_y^\kappa(\mathbf{T}_1) W^{\kappa\gamma}(\mathbf{T}_1, \mathbf{T}_2) \vec{\psi}_z^\gamma(\mathbf{T}_2)) \right], \quad (\text{A13}) \end{aligned}$$

where we have introduced two replica overlap parameters, $Q^{\kappa\gamma}(\mathbf{R}_1, \mathbf{R}_2)$ and $W_{tt}^{\kappa\gamma}(\mathbf{T}_1, \mathbf{T}_2)$:

$$Q^{\kappa\gamma}(\mathbf{R}_1, \mathbf{R}_2) = \sum_{l=1}^N \delta(\mathbf{r}_l^\kappa - \mathbf{R}_1) \delta(\mathbf{r}_l^\gamma - \mathbf{R}_2), \quad (\text{A14})$$

$$\begin{aligned} W_{tt}^{\kappa\gamma}(\mathbf{T}_1, \mathbf{T}_2) = & \sum_{l=3}^N (\mathbf{t}_{l+1}^\kappa - \mathbf{t}_l^\kappa) \delta((\mathbf{t}_{l+1}^\kappa - \mathbf{t}_l^\kappa) - \mathbf{T}_1) \\ & \times \delta((\mathbf{t}_{l+1}^\gamma - \mathbf{t}_l^\gamma) - \mathbf{T}_2) (\mathbf{t}_{l+1}^\gamma - \mathbf{t}_l^\gamma). \quad (\text{A15}) \end{aligned}$$

The parameter $Q_{\kappa\gamma}(\mathbf{R}_1, \mathbf{R}_2)$ measures the similarity between the conformations of two replicas, κ and γ , based on the

spatial distribution of monomers.^{5,17,18} In an analogous fashion, $W_{\kappa\gamma}(\mathbf{T}_1, \mathbf{T}_2)$ measures the similarity between replica conformations, but this time based on differences of polymer tangent vectors. As all globules in the theory are maximally compact, the additional weights $(\mathbf{t}_{l+1}^\kappa - \mathbf{t}_l^\kappa)$ placed on $W_{\kappa\gamma}(\mathbf{T}_1, \mathbf{T}_2)$ have little impact on the overlap.

Before proceeding with our calculation, we introduce two new probability matrices, $\hat{\Delta}^{(q)}$ and $\hat{\Gamma}^{(2)}$.¹⁷

$$(\hat{\Delta}^{(q)})_{ij} \equiv p_i \delta_{ij} - p_i p_j, \quad (\text{A16})$$

$$\hat{\Gamma}^{(2)} = \begin{pmatrix} p_\alpha - p_\alpha^2 & -p_\alpha p_\beta \\ -p_\alpha p_\beta & p_\beta - p_\beta^2 \end{pmatrix}, \quad (\text{A17})$$

and we further redefine our density fields with products of single probabilities:

$$\tilde{\rho}_i^\kappa(\mathbf{R}) \rightarrow p_i \sum_{l=1}^N \delta(s_l, i) \delta(\mathbf{r}_l^\kappa - \mathbf{R}), \quad (\text{A18})$$

$$\tilde{\Upsilon}_z^\kappa(\mathbf{T}) \rightarrow p_z \sum_{l=3}^N (\mathbf{t}_l^\kappa - \mathbf{t}_{l-1}^\kappa) \delta((\mathbf{t}_l^\kappa - \mathbf{t}_{l-1}^\kappa) - \mathbf{T}). \quad (\text{A19})$$

With these new definitions in hand, we re-exponentiate our source term expansion to find a simplified form for the $(n+1)$ -replica partition function:

$$\begin{aligned} \langle Z(seq)^n \rangle_{seq} &= \sum_{\{\mathbf{r}_j^\kappa\}} \int \mathcal{D}\{\vec{\phi}\} \mathcal{D}\{\vec{\psi}\} \\ &\times \exp \left[\langle \vec{\phi} | \frac{1}{4} (\hat{\Omega}^{-1})^{(q(n+1))} \otimes \hat{I}^{(\infty)} \right. \\ &+ \frac{1}{2} \hat{Q}^{((n+1)\infty)} \otimes \hat{\Delta}^{(q)} | \vec{\phi} \rangle \\ &+ \langle \vec{\psi} | \frac{1}{4} (\hat{A}^{-1})^{((3 \times 2)(n+1))} \otimes \hat{I}^{(\infty)} \\ &+ \frac{1}{2} \hat{W}^{(3(n+1)\infty)} \otimes \hat{\Gamma}^{(2)} | \vec{\psi} \rangle \\ &\left. + \langle \vec{\phi} | \vec{\rho} \rangle^{q(n+1)\infty} + \langle \vec{\psi} | \vec{\Upsilon} \rangle^{(3 \times 2)(n+1)\infty} \right]. \end{aligned} \quad (\text{A20})$$

We now evaluate Gaussian functional integrals over each of the coordinates $\mathcal{D}\{\vec{\phi}\}$ and $\mathcal{D}\{\vec{\psi}\}$, and we find¹⁷

$$\begin{aligned} \langle Z^n \rangle &= \frac{1}{2} \sum_{\{\mathbf{r}_j^\kappa\}} \exp[-\mathcal{K}_{\text{eff}}], \\ \mathcal{K}_{\text{eff}} &= \langle \vec{\rho} | \hat{\Omega}_{\text{eff}} | \vec{\rho} \rangle + \langle \vec{\Upsilon} | \hat{A}_{\text{eff}} | \vec{\Upsilon} \rangle + \frac{1}{2} \ln(\det \hat{\Omega}_{\text{eff}} \det \hat{A}_{\text{eff}}), \end{aligned} \quad (\text{A21})$$

$$\begin{aligned} \hat{\Omega}_{\text{eff}} &= (\hat{\Omega}^{1/2} \otimes \hat{I}) [\hat{I} + (\hat{Q} \otimes \hat{\Delta}) (\hat{\Omega} \otimes \hat{I})]^{-1} (\hat{\Omega}^{1/2} \otimes \hat{I}), \\ \hat{A}_{\text{eff}} &= (\hat{A}^{1/2} \otimes \hat{I}) [\hat{I} + (\hat{W} \otimes \hat{\Gamma}) (\hat{A} \otimes \hat{I})]^{-1} (\hat{A}^{1/2} \otimes \hat{I}), \end{aligned}$$

where the dimensionality of each operator is implied by its corresponding dimensionality in Eq. (A20).

3. Replica overlap and replica symmetry breaking

In keeping with Pande, Grosberg, and Tanaka, it is important to note that Eq. (A21) now contains four types of interactions: (1) pairwise monomer-monomer couplings, brought about through $\hat{\Omega}$, (2) energy penalties due to polymer bending, held in the matrix $\hat{A}$, (3) interactions between replicas in standard real (replica) space, given by $\hat{Q}$, and (4) interactions between replicas in the reduced tangent vector space, included in the matrix $\hat{W}$.¹⁷

By performing the replica trick, we have effectively averaged over quenched disorder in polymer sequences. Interactions in replica space, thus, represent couplings between homopolymers, folded in their own spaces into possibly different conformations. Given the absence of sequence-based disorder, we invoke the general Lifshitz condition in globules and use a generalized density distribution as an order parameter for the system.²⁴ The matrices $\hat{Q}$ and $\hat{W}$ represent such a density distribution and contain the only conformation dependence in the partition function.

To perform the sum over conformations indicated in Eq. (A21), we can therefore transform the problem to an integration over $\hat{Q}$ and $\hat{W}$. Instead of the expression shown in Eq. (A21),

$$\langle Z^n \rangle = \frac{1}{2} \sum_{\{\mathbf{r}_j^\kappa\}} \exp[-\mathcal{K}_{\text{eff}}], \quad (\text{A22})$$

we now have the general expression

$$\begin{aligned} \langle Z^n \rangle &= \int \mathcal{D}\{\hat{Q}\} \mathcal{D}\{\hat{W}\} \exp[-\mathcal{F}\{\hat{Q}, \hat{W}\}], \\ \mathcal{F}\{\hat{Q}, \hat{W}\} &= \mathcal{K}_{\text{eff}}\{\hat{Q}, \hat{W}\} - S\{\hat{Q}, \hat{W}\}. \end{aligned} \quad (\text{A23})$$

The nature of this entropy term is complicated, and we can avoid calculating it explicitly with a simple set of scaling arguments, detailed below.

To perform the integral in Eq. (A23), we make a mean-field approximation in the partition function which is equivalent to a saddle-point approximation for the integral over $\hat{Q}$ and $\hat{W}$. As an oddity in performing analytical continuation in the $n \rightarrow 0$ limit, the saddle-point approximation is made by taking the minimal, rather than the maximal, value of integrand. Accordingly, we estimate the partition function by maximizing the replica free energy functional.^{17,18}

$$\langle Z^n \rangle = \int \mathcal{D}\{\hat{Q}\} \mathcal{D}\{\hat{W}\} \exp \left[- \max_{\hat{Q}, \hat{W}} \mathcal{F}\{\hat{Q}, \hat{W}\} \right]. \quad (\text{A24})$$

We carry out this maximization using a variational method, as performed in previous work.¹⁷ We introduce two normalized trial functions, $\omega(\mathbf{x})$ and $\nu(\mathbf{x})$, taken to be

Gaussians for simplicity. Assuming spatial uniformity in our polymer globules, we can then write

$$Q^{\kappa\gamma}(\mathbf{R}_1 - \mathbf{R}_2) = \frac{\rho}{(R_t^{\kappa\gamma})^d} \omega \left(\frac{\mathbf{R}_1 - \mathbf{R}_2}{R_t^{\kappa\gamma}} \right), \quad (A25)$$

$$W^{\kappa\gamma}(\mathbf{T}_1 - \mathbf{T}_2) = \frac{\Upsilon}{(T_t^{\kappa\gamma})^d} \nu \left(\frac{\mathbf{T}_1 - \mathbf{T}_2}{T_t^{\kappa\gamma}} \right),$$

where the normalization conditions determine the nature of $R_t^{\kappa\gamma}$ and $T_t^{\kappa\gamma}$.¹⁷ The parameter $R_t^{\kappa\gamma}$ (and, analogously, $T_t^{\kappa\gamma}$) can be interpreted through the lens of Shakhnovich and Gutin as the diameter of a generalized “tube” in which replicas κ and γ are confined.⁵ If the diameter is very small, the replicas κ and γ overlap nearly completely and represent similar conformations; if $R_t^{\kappa\gamma}$ (or $T_t^{\kappa\gamma}$) is on the order of the globule size, κ and γ have very little overlap and represent completely different conformations. We present our analysis below for the parameter $R_t^{\kappa\gamma}$; the interpretation of $T_t^{\kappa\gamma}$, however, is parallel to that given in the next few paragraphs.

To estimate equilibrium values for these overlap functions, we again borrow from the arguments of Shakhnovich and Gutin to examine replica free energies as a function of $R_t^{\kappa\gamma}$.⁵ Assuming Gaussian statistics hold over small scales in the polymer chain, the entropic cost of confining a polymer in a tube scales as $1/R^2$.^{21,25} Invoking analytical continuation in the $n \rightarrow 0$ limit, confinement of replicas κ and γ should provide an entropic gain that scales as $-(R_t^{\kappa\gamma})^{-2}$. Furthermore, based on the scaling of $\hat{Q}$ and $\hat{W}$, the energy must also scale as $(R_t^{\kappa\gamma})^{-d}$. For replicas $\kappa \neq \gamma$, the free energy thus takes a radically simplified form:

$$\mathcal{F}^{\kappa\gamma} = -\frac{C_1}{(R_t^{\kappa\gamma})^2} + \frac{C_2}{(R_t^{\kappa\gamma})^d}, \quad (A26)$$

where C_1 and C_2 are positive constants.⁵ Since we restrict ourselves to the $d = 3$ case, this free energy has two maxima: $R_t^{\kappa\gamma} = 0$ and $R_t^{\kappa\gamma} = \infty$. Accordingly, in the thermodynamic limit, under a mean-field approximation, and with $n \rightarrow 0$, replicas either overlap completely ($R_t^{\kappa\gamma} = 0$), or do not overlap at all ($R_t^{\kappa\gamma} = \infty$). We see here a simple example of replica symmetry breaking, where replicas collapse into identical or completely disparate states under analytical continuation. This assumption of “zero or total overlap” in the thermodynamic limit is tantamount to approximating the system under the random energy model, which states that different polymer configurations are uncorrelated.¹⁷

Based on these simple scaling arguments, we can rewrite our (trial) overlap parameters as

$$Q^{\kappa\gamma}(\mathbf{R}_1, \mathbf{R}_2) = \rho q_{\kappa\gamma} \delta(\mathbf{R}_1 - \mathbf{R}_2), \quad (A27)$$

$$W^{\kappa\gamma}(\mathbf{T}_1, \mathbf{T}_2) = \Upsilon w_{\kappa\gamma} \delta(\mathbf{T}_1 - \mathbf{T}_2),$$

where the off-diagonal elements of the new matrices $q_{\kappa\gamma}$ and $w_{\kappa\gamma}$ are either 0 or 1, and the replicas on-diagonal, of course, overlap completely.¹⁷

Before we proceed, we will make one additional simplifying approximation. Though the replica overlap parameters $\hat{Q}$ and $\hat{W}$ are measured in differing full and tangent vector spaces, their overlap matrices should coincide in the thermo-

dynamic limit: i.e., if $q_{\kappa\gamma} = 1$, then $w_{\kappa\gamma} = 1$, and if $q_{\kappa\gamma} = 0$, then $w_{\kappa\gamma} = 0$. We formalize this notion by defining the new overlap matrix $o_{\kappa\gamma}$:

$$o_{\kappa\gamma} = q_{\kappa\gamma} \times w_{\kappa\gamma},$$

$$Q^{\kappa\gamma}(\mathbf{R}_1, \mathbf{R}_2) = \rho o_{\kappa\gamma} \delta(\mathbf{R}_1 - \mathbf{R}_2), \quad (A28)$$

$$W^{\kappa\gamma}(\mathbf{T}_1, \mathbf{T}_2) = \Upsilon o_{\kappa\gamma} \delta(\mathbf{T}_1 - \mathbf{T}_2).$$

Qualitatively, $\hat{Q}$ and $\hat{W}$ are simply different measures of overlap between the same microscopic conformations, meaning replicas under these different measures should be expected to coincide.^{17,26}

We have now reduced our problem to the optimization of the new matrix $o_{\kappa\gamma}$, which is independent of $\mathbf{R}$ and $\mathbf{T}$ and contains only ones and zeros. To maximize the n -replica free energy, we now simply need to find the optimal grouping of replicas. Grouping replicas that overlap completely, we perform row and column operations on $o_{\kappa\gamma}$ to yield a matrix in block diagonal form. This matrix is called the one-step Parisi matrix, and is used in the simplest cases of replica symmetry breaking.^{17,26}

One block in the $(n + 1) \times (n + 1)$ -dimensional Parisi matrix contains $y + 1$ replicas that overlap with the “target replica” (in the target conformation, $\star$). The remaining $n - y$ replicas are split into $(n - y)/x$ groups, each containing x replicas. In the optimization process, y replicas have been “adsorbed” onto the target conformation; the other groupings of x replicas result from spontaneous symmetry breaking of replicas.¹⁷

4. Replica thermodynamics and phase transitions

Upon organization of the Parisi matrix, we have reduced our heteropolymer freezing order parameter to two numbers, x and y , that describe how replicas group into spontaneous and target groups, respectively. With this simplification in hand, we can approximate the entropy in replica space as the entropic price to grouping replicas. In each replica group, one conformation is chosen arbitrarily, while the remaining $x - 1$ or y conformations are fixed. The loss of entropy due to grouping is thus given by

$$S = -N s \left[\frac{n - y}{x} (x - 1) + y \right], \quad (A29)$$

where we have noted that the entropy should scale with the number of monomers N and a polymer rigidity parameter, s .^{17,18,21} The nature of this s parameter is important, as noted in the main body of this article.

Given that replicas only interact within their respective groupings, the expression for the energy given in Eq. (A21) can also be transformed into a relation dependent on x and y . While this process is somewhat laborious, it is relatively simple to remove the real and replica spaces and write $\mathcal{K}_{\text{eff}}$ in terms of replica groupings.¹⁷ The resulting expression takes a form similar to that of the entropy, containing the energy of the target replica grouping and the $(n - y)/x$ spontaneously

formed groups. The expression for $\mathcal{K}_{\text{eff}}$ now reads

$$\begin{aligned} \mathcal{K}_{\text{eff}} = & \frac{N}{2} \left(\frac{n-y}{x} \right) \{ \ln(\det[\hat{I} + x \hat{\Delta} \hat{\Omega}] \det[\hat{I} + x \hat{\Gamma} \hat{A}]) \\ & + \langle \vec{p} | x \hat{\Omega} [\hat{I} + \hat{\Delta} \hat{\Omega}]^{-1} | \vec{p} \rangle + \langle \vec{u} | x \hat{A} [\hat{I} + \hat{\Gamma} \hat{A}]^{-1} | \vec{u} \rangle \} \\ & + \frac{N}{2} \ln(\det[\hat{I} + \hat{\Delta} \hat{\Omega}_{\text{des}} + y \hat{\Delta} \hat{\Omega}]) \\ & \times \det[\hat{I} + \hat{\Gamma} \hat{A}_{\text{des}} + y \hat{\Gamma} \hat{A}]) \\ & + \langle \vec{p} | \left[\frac{1}{2} \hat{\Omega}_{\text{des}} + y \hat{\Omega} \right] [\hat{I} + \hat{\Delta} \hat{\Omega}_{\text{des}} + y \hat{\Delta} \hat{\Omega}]^{-1} | \vec{p} \rangle \\ & + \langle \vec{u} | \left[\frac{1}{2} (\hat{A}_{\text{des}}) + y \hat{A} \right] [\hat{I} + \hat{\Gamma} \hat{A}_{\text{des}} + y \hat{\Gamma} \hat{A}]^{-1} | \vec{u} \rangle, \end{aligned} \quad (\text{A30})$$

where the fields $\vec{p}$ and $\vec{u}$ have been transformed into simple probability composition vectors, $\vec{p}$ and $\vec{u}$.^{17,27}

Given that our free energy has retained the two order parameters x and y from classical freezing models, we still expect to see three phases in the present theory: folded (target), glassy (non-target), and random. To investigate the nature of these phases, we need to study the free energy as a function of these order parameters x and y .^{17,26}

Noting the approximations we have made to the model thus far, we can make a number of simplifications to Eqs. (A29) and (A30). In particular, to facilitate freezing transitions and accommodate our high temperature expansion, we truncate the energy to quadratic order in $\hat{\Omega}$ and $\hat{A}$ and consider only relatively small values of s . Furthermore, we note that, since $0 \leq y \leq n$, and we use analytical continuation in the $n \rightarrow 0$ limit, terms at quadratic order and above in y will be ineffectual. Truncating the energy to quadratic order and linearizing the free energy in y , we are left with

$$\begin{aligned} \frac{F}{NT} = & \text{Tr} \left[(n-y) \left\{ \frac{\hat{\Delta} \hat{\Omega}}{2T} - x \frac{\hat{\Delta} \hat{\Omega} \hat{\Delta} \hat{\Omega}}{4T^2} + \frac{\hat{P}_{\Omega} \hat{\Omega}}{2T} - x \frac{\hat{P}_{\Omega} \hat{\Omega} \hat{\Delta} \hat{\Omega}}{2T^2} \right. \right. \\ & + \frac{\hat{\Gamma} \hat{A}}{2T} - x \frac{\hat{\Gamma} \hat{A} \hat{\Gamma} \hat{A}}{4T^2} + \frac{\hat{P}_A \hat{A}}{2T} - x \frac{\hat{P}_A \hat{A} \hat{\Gamma} \hat{A}}{2T^2} \left. \right\} \\ & + \frac{\hat{\Delta} \hat{\Omega}_{\text{des}}}{2T_{\text{des}}} + y \frac{\hat{\Delta} \hat{\Omega}}{2T} - y \frac{\hat{\Delta} \hat{\Omega}_{\text{des}} \hat{\Delta} \hat{\Omega}}{2T T_{\text{des}}} - \frac{\hat{\Delta} \hat{\Omega}_{\text{des}} \hat{\Delta} \hat{\Omega}_{\text{des}}}{4T_{\text{des}}^2} \\ & + \frac{\hat{P}_{\Omega} \hat{\Omega}_{\text{des}}}{2T_{\text{des}}} + y \frac{\hat{P}_{\Omega} \hat{\Omega}}{2T} - \frac{\hat{P}_{\Omega} \hat{\Omega}_{\text{des}} \hat{\Delta} \hat{\Omega}_{\text{des}}}{2T_{\text{des}}^2} \\ & - y \frac{\hat{P}_{\Omega} \hat{\Omega} \hat{\Delta} \hat{\Omega}_{\text{des}}}{2T T_{\text{des}}} - \frac{\hat{P}_{\Omega} \hat{\Omega}_{\text{des}} \hat{\Delta} \hat{\Omega}}{2T T_{\text{des}}} \\ & + \frac{\hat{\Gamma} \hat{A}_{\text{des}}}{2T_{\text{des}}} + y \frac{\hat{\Gamma} \hat{A}}{2T} - y \frac{\hat{\Gamma} \hat{A}_{\text{des}} \hat{\Gamma} \hat{A}}{2T T_{\text{des}}} - \frac{\hat{\Gamma} \hat{A}_{\text{des}} \hat{\Gamma} \hat{A}_{\text{des}}}{4T_{\text{des}}^2} \\ & + \frac{\hat{P}_A \hat{A}_{\text{des}}}{2T_{\text{des}}} + y \frac{\hat{P}_A \hat{A}}{2T} - \frac{\hat{P}_A \hat{A}_{\text{des}} \hat{\Gamma} \hat{A}_{\text{des}}}{2T_{\text{des}}^2} \\ & - y \frac{\hat{P}_A \hat{A} \hat{\Gamma} \hat{A}_{\text{des}}}{2T T_{\text{des}}} - \frac{\hat{P}_A \hat{A}_{\text{des}} \hat{\Gamma} \hat{A}}{2T T_{\text{des}}} \left. \right] \\ & + s \left[y + (n-y) \frac{x-1}{x} \right], \end{aligned} \quad (\text{A31})$$

where we have defined two new probability matrices with $(\hat{P}_{\Omega})_{ij} = p_i p_j$ and $(\hat{P}_A)_{yz} = p_y p_z$ corresponding to products of sequence and secondary structural identity probabilities, respectively.¹⁷

We are now ready to study the phase transitions induced by the order parameters x and y . The first type of phase transition, related to non-specific heteropolymer freezing, corresponds to the spontaneous symmetry breaking of replicas into groups of size x . By optimizing the free energy in Eq. (A31) with respect to x (i.e., taking $\partial F / \partial x = 0$), we find the equation that determines x :

$$s = \frac{x^2}{4T^2} \text{Tr} [\hat{\Delta} \hat{\Omega} \hat{\Delta} \hat{\Omega} + 2 \hat{P}_{\Omega} \hat{\Omega} \hat{\Delta} \hat{\Omega} + \hat{\Gamma} \hat{A} \hat{\Gamma} \hat{A} + 2 \hat{P}_A \hat{A} \hat{\Gamma} \hat{A}]. \quad (\text{A32})$$

As in traditional models of heteropolymer freezing, the freezing transition occurs at $x = 1$, where replicas begin to group spontaneously. This freezing transition represents a phase transition from random globule conformations to a glassy ground state that is very likely not the target conformation.¹⁷

Freezing to the target conformation is controlled by the order parameter y . Given that we have used analytical continuation techniques in the $n \rightarrow 0$ limit, the optima in F will correspond to either $y = 0$ (frozen into the non-target phase) or $y = n$ (frozen into the target phase). To determine the crossover between these optima, we take $\partial F / \partial y = 0$ and find the relationship

$$\begin{aligned} s = & \text{Tr} \left[\frac{x}{4} \left(\frac{\hat{\Delta} \hat{\Omega} \hat{\Delta} \hat{\Omega}_{\text{des}}}{T T_{\text{des}}} + 2 \frac{\hat{P}_{\Omega} \hat{\Omega} \hat{\Delta} \hat{\Omega}_{\text{des}}}{T T_{\text{des}}} + \frac{\hat{\Delta} \hat{\Omega}_{\text{des}} \hat{\Delta} \hat{\Omega}}{T_{\text{des}} T} \right. \right. \\ & + 2 \frac{\hat{P}_{\Omega} \hat{\Omega}_{\text{des}} \hat{\Delta} \hat{\Omega}}{T T_{\text{des}}} \\ & + \frac{\hat{\Gamma} \hat{A} \hat{\Gamma} \hat{A}_{\text{des}}}{T T_{\text{des}}} + 2 \frac{\hat{P}_A \hat{A} \hat{\Gamma} \hat{A}_{\text{des}}}{T T_{\text{des}}} + \frac{\hat{\Gamma} \hat{A}_{\text{des}} \hat{\Gamma} \hat{A}}{T_{\text{des}} T} + 2 \frac{\hat{P}_A \hat{A}_{\text{des}} \hat{\Gamma} \hat{A}}{T T_{\text{des}}} \left. \right) \\ & - \frac{x^2}{4T^2} (\hat{\Delta} \hat{\Omega} \hat{\Delta} \hat{\Omega} + 2 \hat{P}_{\Omega} \hat{\Omega} \hat{\Delta} \hat{\Omega} + \hat{\Gamma} \hat{A} \hat{\Gamma} \hat{A} + 2 \hat{P}_A \hat{A} \hat{\Gamma} \hat{A}) \left. \right]. \end{aligned} \quad (\text{A33})$$

The phase transition described by this expression corresponds to freezing to the designed ground state conformation for the system. The nature and extent of both of these phase transitions are discussed in the “Results and Discussion” section of this article.¹⁷

5. Geometric parameterization

Prior to deriving the model’s phase diagram, however, the expressions taken from x and y can be simplified somewhat with a geometric parameterization of the interaction matrices. We break each trace operator into sums over the interaction and probability matrix products, and we define the operators $\delta \hat{\Omega}$ and $\delta \hat{A}$ such that Eqs. (A34) and (A35) defined here

$$s = \frac{x^2}{4T^2} (\delta \Omega^2 + \delta A^2), \quad (\text{A34})$$

$$s = \frac{x}{2T T_{\text{des}}} (\delta \Omega \delta \Omega_{\text{des}} + \delta A \delta A_{\text{des}}) - \frac{x^2}{4T^2} (\delta \Omega^2 + \delta A^2), \quad (\text{A35})$$

correspond exactly to Eqs. (A32) and (A33), respectively.¹⁷ Accordingly, for an arbitrary operator $\hat{G}$ (here representing $\hat{\Omega}$ or $\hat{A}$), the operator $\delta\hat{G}$ is defined by

$$(\delta\hat{G})_{kl} = G_{kl} - \bar{G}, \quad (\text{A36})$$

where the mean of the operator, $\bar{G}$, is given by

$$\bar{G} = \sum_{k,l} p_k G_{kl} p_l. \quad (\text{A37})$$

Interpreting the objects $\delta\hat{G}$ and $\delta\hat{G}'$ as vectors in a generalized vector space, we can define their inner product by

$$\delta\hat{G} \cdot \delta\hat{G}' = \sum_{k,l} p_k \delta G_{kl} \delta G'_{kl} p_l, \quad (\text{A38})$$

meaning the square-norm of a vector $\delta\hat{G}$ is given by

$$\delta G^2 = \sum_{k,l} p_k \delta G_{kl} \delta G_{kl} p_l = \delta\hat{G} \cdot \delta\hat{G}. \quad (\text{A39})$$

We can thus interpret the terms in Eqs. (A32) and (A33) through these geometric definitions, where operations are defined in a generalized vector space with elements determined by the model's standard interaction kernels.¹⁷ Colloquially, the quantity δG can be referred to as the “variance” of the kernel G .

With this geometric parameterization in hand, we thus arrive at the conditions needed to describe the phase behavior in our model, originally presented in Eqs. (9) and (10) in this article.

¹K. A. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl, *Annu. Rev. Biophys.* **37**, 289 (2008).

²G. R. Bowman and V. S. Pande, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 10890 (2010).

³F. Chiti and C. M. Dobson, *Nat. Chem. Biol.* **5**, 15 (2009).

⁴J. D. Bryngelson and P. G. Wolynes, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 7524 (1987).

⁵E. I. Shakhnovich and A. M. Gutin, *Biophys. Chem.* **34**, 187 (1989).

⁶H. S. Chan and K. A. Dill, *Phys. Today* **46**, 24 (1993).

⁷H. Taketomi, Y. Ueda, and N. Go, *Int. J. Pept. Protein Res.* **7**, 445 (1975).

⁸C. J. Camacho and D. Thirumalai, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 6369 (1993).

⁹V. S. Pande, C. Joerg, A. Y. Grosberg, and T. Tanaka, *J. Phys. A* **27**, 6231 (1994).

¹⁰V. A. Voelz, G. R. Bowman, K. Beauchamp, and V. S. Pande, *J. Am. Chem. Soc.* **132**, 1526 (2010).

¹¹K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, *Science* **334**, 517 (2011).

¹²T. J. Lane, G. R. Bowman, K. Beauchamp, V. A. Voelz, and V. S. Pande, *J. Am. Chem. Soc.* **133**, 18413 (2011).

¹³K. A. Beauchamp, R. McGibbon, Y. S. Lin, and V. S. Pande, *Proc. Natl. Acad. Sci. U.S.A.* **109**, 17807 (2012).

¹⁴V. S. Pande, *Phys. Rev. Lett.* **105**, 198101 (2010).

¹⁵M. B. Prigozhin and M. Gruebele, *J. Am. Chem. Soc.* **133**, 19338 (2011).

¹⁶J. K. Weber, R. L. Jack, and V. S. Pande, *J. Am. Chem. Soc.* **135**, 5501–5504 (2013).

¹⁷V. S. Pande, A. Y. Grosberg, and T. Tanaka, *Rev. Mod. Phys.* **72**, 259 (2000).

¹⁸M. Mezard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore/Teaneck, NJ, 1987).

¹⁹V. S. Pande, A. Y. Grosberg, and T. Tanaka, *Proc. Nat. Acad. Sci. U.S.A.* **91**, 12976 (1994).

²⁰E. I. Shakhnovich and A. Gutin, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 7195 (1993).

²¹A. Y. Grosberg and A. R. Khokhlov, *Statistical Physics of Macromolecules* (AIP, New York, 1994).

²²S. Doniach, T. Garel, and H. Orland, *J. Chem. Phys.* **105**, 1601 (1996).

²³J. Yan and J. F. Marko, *Phys. Rev. Lett.* **93**, 108108 (2004).

²⁴I. M. Lifshitz, A. Y. Grosberg, and A. R. Khokhlov, *Rev. Mod. Phys.* **50**, 683 (1978).

²⁵P. G. de Gennes, *Scaling Concepts in Polymer Physics* (Cornell University, Ithaca, NY, 1979).

²⁶G. Parisi, *Physics of Complex Systems: Proceedings of the Enrico Fermi School of Physics, Course CXXXIV* (Soc. It. Fisica, Bologna, 1997).

²⁷V. S. Pande, A. Y. Grosberg, and T. Tanaka, *Phys. Rev. E* **51**, 3381 (1995).

²⁸S. Miyazawa and R. Jernigan, *Macromolecules* **18**, 534 (1985).

²⁹Z. Qin and M. J. Buehler, *Phys. Rev. Lett.* **104**, 198304 (2010).