A hybrid, bottom-up, structurally accurate, Gō-like coarse-grained protein model

Cite as: J. Chem. Phys. 151, 044111 (2019); doi: 10.1063/1.5108761

Submitted: 2 May 2019 • Accepted: 24 June 2019 •

Published Online: 29 July 2019

Tanmoy Sanyal,^{1,a)} Jeetain Mittal,² and M. Scott Shell^{1,b)}

AFFILIATIONS

¹Department of Chemical Engineering, University of California Santa Barbara, Santa Barbara, California 93106, USA

²Department of Chemical and Biomolecular Engineering, Lehigh University, Bethlehem, Pennsylvania 18015, USA

^{a)}Also at: Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, CA 94158, USA.

^{b)}Author to whom correspondence should be addressed: shell@engineering.ucsb.edu

ABSTRACT

Coarse-grained (CG) protein models in the structural biology literature have improved over the years from being simple tools to understand general folding and aggregation driving forces to capturing detailed structures achieved by actual folding sequences. Here, we ask whether such models can be developed systematically from recent advances in bottom-up coarse-graining methods without relying on bioinformatic data (e.g., protein data bank statistics). We use relative entropy coarse-graining to develop a hybrid CG but Gō-like CG peptide model, hypothesizing that the landscape of proteinlike folds is encoded by the backbone interactions, while the sidechain interactions define which of these structures globally minimizes the free energy in a unique native fold. To construct a model capable of capturing varied secondary structures, we use a new extended ensemble relative entropy method to coarse-grain based on multiple reference atomistic simulations of short polypeptides with varied α and β character. Subsequently, we assess the CG model as a putative protein backbone forcefield by combining it with sidechain interactions based on native contacts but not incorporating native distances explicitly, unlike standard Gō models. We test the model's ability to fold a range of proteins and find that it achieves high accuracy (~ 2 Å root mean square deviation resolution for both short sequences and large globular proteins), suggesting the strong role that backbone conformational preferences play in defining the fold landscape. This model can be systematically extended to non-natural amino acids and nonprotein polymers and sets the stage for extensions to non-Gō models with sequence-specific sidechain interactions.

Published under license by AIP Publishing. <https://doi.org/10.1063/1.5108761>

I. INTRODUCTION

It was established over half a century ago that folding of protein and other biomolecules forms the foundations of structural biology by linking structure to function for molecular-level biological processes.^{1,2} With the exception of intrinsically disordered proteins, the biological mechanism of a protein is determined by its three dimensional (3D) native structure, which is encoded in its particular sequence of amino acids. Due to the remarkable progress in experimental techniques for protein structure determination over the last two decades, such as crystallographic methods, nuclear magnetic resonance (NMR) spectroscopy, cryo electron microscopy, and other nonlinear optical techniques,^{3–8} we now have an extensive database of more than 130 000 3D protein structures

or their complexes.⁹ Computational models of protein structure are not only essential for supporting experimental methods of solving native structures from amino acid sequences, but they are also crucial for mechanistic investigations into the driving forces for folding and assembly and in biotechnological applications such as high throughput drug design.^{10–12} Molecular Dynamics (MD) simulations of protein folding using atomistically detailed models can in principle address these applications.¹³ However, all-atom (AA) MD is typically limited to time scales of tens of microseconds, while the characteristic folding time for even small (~ 50 amino acid residues) proteins is on the order of milliseconds and upwards. Except in very special cases, using custom-built hardware¹⁴ or globally distributed computing strategies,¹⁵ so far it has been nearly impossible to break the millisecond barrier. This fundamental challenge has

motivated the development of coarse-grained (CG) protein models which reduce the number and complexity of the degrees of freedom in MD simulations, thus allowing a faster sampling of conformational space in addition to reduced bare computational cost.

One of the principal aims of CG peptide models has been to study the driving forces behind protein folding. The first CG peptide models were developed more than three decades ago by Warshel and Levitt, achieving a resolution of 6.5 Å backbone root mean square deviation (RMSD) for folding the bovine pancreatic trypsin inhibitor protein PTI,¹⁶ and since then, there has been considerable progress in their development.^{17–20} Owing to the continued increase in computational resources, CG protein models have steadily become more sophisticated in both resolution and complexity of interactions. Here, we mention a few selected examples. Early efforts such as HP models focused on simple lattice chains described by a binary alphabet of amino acids: hydrophobic (H) and polar (P) to demonstrate that structure is uniquely modulated by sequence.^{21,22} However, lattice frameworks restrict orientational degrees of the peptide chain so that subsequently developed intermediate resolution models containing three or more CG backbone sites describing at least an α carbon and a CG carbonyl group were better able to capture proteinlike structures.^{23–33} For example, the four-site PRIME CG model by Hall and co-workers was successfully used to investigate large scale self-assembly behavior of polyalanine and polyglutamine oligomers.^{23,34,35} A similar resolution CG model developed by Deserno and co-workers was used to predict structure and kinetics in three-helix bundles and transmembrane helical peptides.^{25,36} The extremely popular MARTINI CG model developed by Marrink and co-workers provides a large set of amino acid specific parameters that can be strung together to create forcefields for different peptide sequences.^{27,37} MARTINI has been used extensively to probe protein-lipid interactions.^{38,39}

Most of the approaches enumerated above are top-down in that they utilize experimental data to inform CG models. In contrast, this work considers the potential for bottom-up methods to develop CG protein models with improved accuracy. Bottom-up strategies systematically remove degrees of freedom from detailed atomistic simulations and utilize statistical mechanical principles to *inverse* design simpler CG interactions that renormalize the entire set of atomistic intermolecular forces, in order to reproduce emergent properties in the original atomic system. A notable effort within this class of CG models is the Multiscale Coarse-Graining (MS-CG) method by Voth and co-workers, which parameterizes CG interactions by matching forces at CG sites between AA and CG resolutions.⁴⁰ The MS-CG approach was used to design four-site CG models of (helical) polyalanine and the (hairpin) sequence V₅PGV₅; this CG strategy preserved the native states as seen in the reference AA simulations within 1 Å backbone RMSD.⁴¹ Rudzinski and Noid used the MS-CG method to parameterize implicit water C- α models from atomistic polyalanine systems and found that while order parameters such as the radius of gyration and helix propensity were reproduced accurately, the low CG resolution and a simple basis consisting of three types of nonbonded CG potentials were insufficient to capture the AA helix-coil transitions.⁴² Mullinax and Noid used an extended-ensemble-coarse-graining method⁴³ to construct statistical potentials for a three-alphabet CG model (hydrophilic, hydrophobic, and neutral CG sites) from a model databank populated with

near native conformations of candidate α , β , and mixed α/β sequences.⁴⁴ Peter and co-workers used iterative Boltzmann inversion to develop an intermediate resolution CG model for the short amphiphilic EALA peptide to investigate its pH induced helix-coil transition.⁴⁵

Despite many successes of both top-down and bottom-up CG peptide models in the literature, a main intent has been to understand general features of the folding mechanism rather than accurate structure predictions. While top-down approaches furnish excellent insights into the folding free energy landscape, bottom-up models can leverage the entire underlying atomistic configurational landscape to systematically and automatically design effective CG interactions. However, inverse designing interactions that reproduce the very high dimensional all-atom configurational distribution admittedly makes pure bottom-up methods less efficient than enumerating structures that satisfy experimentally generated constraints, as often done in top-down approaches. In fact, some top-down efforts have resulted in robust models by supplementing physics-based Hamiltonians with bioinformatic terms derived from structure databases.^{24–26,31} However, a bottom-up approach is still attractive since independence from bioinformatically obtained constraints can enable the inclusion of non-natural and synthetic chemical constructs for which structure statistics would be either difficult or nearly impossible to generate.

This work explores the ability of a recent bottom-up coarse-graining strategy to develop a generic CG protein model that folds protein structures to high accuracy, at 1–2 Å resolution. Our investigation is motivated by the earlier effort of Carmichael and Shell that developed a bottom-up four-site CG model for polyalanine from a reference atomistic polyalanine simulation.⁴⁶ This earlier model demonstrated surprising accuracy and transferability, whereby both the free energy landscape and temperature-dependent folding behavior were quantitatively captured relative to atomistic simulations. Interestingly, though this model was developed from a mostly helical single polypeptide reference simulation, it also produced β -rich amyloid structures in self-assembly simulations, with correct sheet alignment, packing and twist.

Here, we investigate whether it is possible to extend Carmichael and Shell's approach to capture native structures of arbitrary globular proteins. Ideally, one would seek a sequence-flexible model with distinct sidechain parameters for all 20 amino acids, but as a necessary first step toward that grand goal, here, we consider the bottom-up development of native-structure-informed models. That is, the CG model includes favorable sidechain interactions between amino acids known to be in contact in the native structure. In this sense, this effort is a test of whether or not bottom-up methods can produce accurate backbone interactions and secondary structures even when the sidechain interactions are ideal—in that they are only attractive for residue contacts formed in the actual native structure—thus exploring the ability of the backbone interactions to capture a wide range of protein folds. Similar to the Carmichael and Shell model, we develop CG polypeptide forcefields from reference AA polypeptide simulations but subsequently parameterize sidechain potentials to favor native contact interactions. Understanding how to construct such Gō-like models sets the stage for non-Gō-like models that do not rely on native contact information by ultimately including distinct sidechain interactions for all 20 amino acids.

Long studied in the protein folding community since their introduction,⁴⁷ Gō models modulate sidechain interactions between residues depending on whether or not they are in contact (i.e., within a threshold proximity) in the experimental native structure; contact residues are given an attractive interaction which is most favorable at the distance of contact in the native structure, while all other non-native residue pairs merely experience excluded volume interactions. Such models are motivated by the principle of minimum frustration, which postulates that proteins evolved toward sequences with the ability to fold into low-energy conformations while selecting against local minima traps by avoiding misfolding and non-native contacts.⁴⁸ Gō models take this idea a step further by treating native contacts as the sole driving force for the folding process. Although such models are not *de novo* (i.e., they cannot be used to predict structure from sequence alone), they have proven useful for understanding functional protein conformational fluctuations and generic aspects of folding pathways and may also rapidly provide structural ensembles consistent with experimentally determined native contact information (e.g., via NMR) to improve structure prediction efforts.^{49–54} Over the last decade, several flavors of Gō-like and other native-structure based CG models have been proposed that utilize various physical and bioinformatic data, such as the distribution of hydrogen-bonds, homologous sequences, or solvation transfer free energies of amino acids, in addition to the native structure.^{55–58} In contrast, the present approach does not include any experimental data beyond the native structure. Specifically, we parameterize generic (i.e., amino acid unspecific) backbone interactions and Gō-like native sidechain interactions in a bottom-up fashion directly from small-scale atomistic simulations. The Gō-like interactions pursued here are meant to serve as simplistic protein-specific sidechain potentials that enforce minimal experimental information (list of native contacts) on the folding process to bring out the intrinsic capability of the backbone in sampling the correct conformational space.

Our premise is based on the observation that backbone interactions are key to the folding landscape: even simple sequences that do not fold to unique structures (e.g., homopolymeric peptides) can explore a wide range of realistic folds, involving both α and β secondary structure elements. An early study by Head-Gordon *et al.* found that polyalanine can mimic the secondary structures of several globular proteins such as PTI, crambin, ribonuclease A, and superoxide dismutase.⁵⁹ Specifically, the authors observed that the potential energy landscapes of alanine polymers of different lengths have local minima close to the global free energy minima at the native structures of different proteins of equivalent sequence length. More recent work by Laio and co-workers found that an atomistic 60-mer of polyvaline was able to produce practically *all* compact folds ($\sim 100\%$ of the CATH database⁶⁰ for 40–65 residue long structures) observed in nature for proteins of comparable length, including α , β , and mixed-content structures.⁶¹ A similar conclusion was reached through CG polyvaline simulations by Vendruscolo and co-workers, where they recovered 135 folds from the ensemble of structures generated by a valine 60-mer, out of the 265 folds reported in the CATH database for proteins containing between 40 and 75 residues.⁶² These studies suggest a picture whereby backbone interactions such as hydrogen bonding, dihedral energetics, and excluded volume interactions, together with

the inherent connectivity constraints of the backbone, define the conformational landscape for a protein chain, while the sidechains then adjust the energetic favorability of different structures so as to select a single fold.^{63–66} Thus, a fundamental challenge that we test here is the ability of a bottom-up strategy to generate computationally fast, generic CG backbone interactions that correctly reproduce the structural landscape. Favorable interactions between native contact pairs (i.e., Gō-like interactions) provide a natural context to establish an upper bound on the quality of such CG backbones since they, in effect, represent one of the simplest possible sidechain interactions.

II. METHODS

We use a four-site CG model motivated by the success of the Carmichael and Shell⁴⁶ four-site polyalanine model and by other work from the same authors showing that the four-site variant is more representable and transferable than other one-, two-, and three-site cases.^{33,67} Four-site models have also been successfully pursued by the Hall²³ and Deserno groups.²⁵ As shown in Fig. 1(a) with an example polyleucine 15-mer, the CG mapping ignores hydrogen and discretizes each amino-acid into CG sites corresponding to four heavy atom centers: nitrogen (N), α -carbon (C), carbonyl carbon and oxygen (O), and sidechain (S). Admittedly, larger amino acids, especially those with one or more aromatic rings (HIS, PHE, TYR, TRP, etc.) may require a higher sidechain resolution, but we leave such a refinement to the future.

In this paper, we develop and subsequently test CG peptide models by combining CG backbone forcefields with Gō-like sidechain interactions, both developed in a bottom-up manner from atomistic reference simulations of candidate peptides using the relative entropy minimization method.⁶⁸ The relative entropy between an AA system and its corresponding CG model is an information theoretic measure that encodes the overlap between AA and CG microstate probability distributions.⁶⁸ Minimizing the relative entropy with respect to forcefield parameters of the CG model guarantees maximal overlap of the AA and CG microstate probability

FIG. 1. Leucine is mapped to four heavy atom centers N, C, O, and S that sit at the centers-of-mass of the amino, α -carbon, carbonyl carbon and oxygen, and the side-chain groups, respectively. The CG model of a 15-mer leucine polypeptide is parameterized by minimizing the relative entropy using a reference atomistic simulation.

distributions, enabling the CG model to recapitulate important thermophysical properties of the atomistic system (e.g., secondary structure).⁶⁹ This technique requires an atomistic trajectory to calculate the relative entropy between an AA and a proposed CG model, which is then minimized in the space of CG forcefield parameters through an iterative procedure. The relative entropy between an AA polypeptide reference and its corresponding CG model is given by

$$S_{\text{rel}} = \int p_{\text{AA}}(\mathbf{r}) \ln \left(\frac{p_{\text{AA}}(\mathbf{r})}{p_{\text{CG}}(\mathbf{M}(\mathbf{r}))} \right) d\mathbf{r} + S_{\text{map}}, \quad (1)$$

where $\mathbf{r}$ and $\mathbf{R}$ represent AA and CG configurations, respectively, and $\mathbf{M}(\mathbf{r})$ is a mapping operator (typically a matrix) that replaces groups of atoms in the AA representation with center-of-mass sites in the CG model. p_x gives the equilibrium conformational probabilities for the ensembles with $X = \text{AA or CG}$. The integral in Eq. (2) proceeds over all AA conformations although it is possible to cast it in terms of the CG conformations.⁷⁰ S_{map} is a “mapping entropy” that accounts for the degeneracy in the $\text{AA} \rightarrow \text{CG}$ mapping, i.e., it measures the number of distinct AA configurations that map to the same CG one. However, this term is independent of the CG forcefield, and because we fix the CG resolution to four sites per amino acid, S_{map} can be treated as a constant that plays no role in this work. In the canonical ensemble, the derivative of the relative entropy with respect to the CG forcefield parameters λ can be written as

$$\frac{\partial S_{\text{rel}}}{\partial \lambda} = \beta \left[\left\langle \frac{\partial U_{\text{CG}}}{\partial \lambda} \right\rangle_{\text{AA}} - \left\langle \frac{\partial U_{\text{CG}}}{\partial \lambda} \right\rangle_{\text{CG}} \right], \quad (2)$$

where U_{CG} is the total CG interaction potential, β represents the inverse temperature ($1/k_B T$), and $\langle \rangle_X$ represents averaging over all conformations in the model with $X = \text{AA or CG}$. The parameterization of the CG model proceeds by using Eq. (2) to minimize S_{rel} , in practice using a combination of perturbation theory and conjugate gradient minimization as described in Ref. 33.

A. CG backbone forcefields

As in Ref. 46, we parameterize CG models from atomistic polypeptides and extract “backbone” forcefields from these models as the sum of intrabackbone and backbone-sidechain interactions. CG polypeptide forcefields ($U_{\text{CG}}^{\text{poly}}$) are represented using bonded (U_b), angular (U_θ), torsional ($U_{\phi,\psi}$), and nonbonded pair potentials (U_{pair}) applied to intrabackbone (BB), backbone-sidechain (BS), and intersidechain (SS) CG sites,

$$U_{\text{CG}}^{\text{poly}} = (U_b^{\text{BB}} + U_\theta^{\text{BB}} + U_{\phi,\psi}^{\text{BB}} + U_{\text{pair}}^{\text{BB}}) + (U_b^{\text{BS}} + U_\theta^{\text{BS}} + U_{\phi,\psi}^{\text{BS}} + U_{\text{pair}}^{\text{BS}}) + U_{\text{pair}}^{\text{SS}}. \quad (3)$$

The bond potentials are harmonic, while all other potentials are represented as cubic B-splines whose knot points are optimizable parameters of the CG model. The nonbonded pair potentials are cut off at 10 Å. Angle, torsion, and pair potentials use 40 knot points each with densities of 1.4°/knot, 0.22°/knot, and 0.25 Å/knot, respectively. Prior to developing Gō-like models, we first construct full CG models for a reference AA polypeptide using relative

entropy minimization. Later, in Sec. III A, we test the performance of these CG models by comparing the folding behavior between AA and CG models for the reference polypeptides. Subsequently, we then replace side-chain interactions with Gō-like potentials. The backbone forcefield ($U_{\text{CG}}^{\text{BB}}$) for a Gō-like model simply amounts to taking all CG interactions from the fully bottom-up CG model for polypeptides except for the intersidechain pair potentials,

$$U_{\text{CG}}^{\text{BB}} = (U_b^{\text{BB}} + U_\theta^{\text{BB}} + U_{\phi,\psi}^{\text{BB}} + U_{\text{pair}}^{\text{BB}}) + (U_b^{\text{BS}} + U_\theta^{\text{BS}} + U_{\phi,\psi}^{\text{BS}} + U_{\text{pair}}^{\text{BS}}). \quad (4)$$

It may be noted that the parameters of the interbackbone-sidechain potential $U_{\text{pair}}^{\text{BS}}$ are completely determined by the reference polypeptide and therefore are sidechain unspecific. As outlined earlier in Sec. I, this work strongly emphasizes the idea that a generic backbone is sufficient to determine the folding landscape regardless of specific amino acid secondary structure propensity. Thus, the bottom-up CG models in this paper are purely backbone models that are later (see Sec. II B) combined with native contact information.

Naturally, an important design criterion for a transferable CG backbone forcefield is an appropriate overall (i.e., amino-acid-unspecific, since the CG models are sequence independent) balance between α helical, β -sheet, and extended conformations such that additional protein- or sequence-specific sidechain interactions can then select among many proteinlike conformations to steer the peptide chain toward a well-defined native fold. As mentioned in Sec. I, proteins achieve this balance through geometrical constraints, steric repulsions, hydrogen bonds, electrostatic interactions, and hydrophobic effects.^{63–66} CG backbones developed in this work do not encode such interactions explicitly but instead critically depend on the sampling of conformation space by the reference AA polypeptides to “learn” effective coarse interactions within the functional form of the potentials in Eq. (4). Here, we focus on moderate-length polypeptides generated from hydrophobic amino acids as putative AA references. Furthermore, to retain computational efficiency, we do not incorporate either explicit solvent or explicit intrapeptide and peptide-solvent electrostatics in our CG models; all solvent-mediated effects are implicit, embedded in effective intramolecular interactions.

The choice of the reference AA system from which to develop the CG backbone model is critical because it influences the ultimate secondary structure balance. We examine four such references to explore and identify cases with a good balance of α , β , and flexible conformational propensities, as summarized in Table I. It is non-trivial to identify a system with such balance *a priori*, and ultimately, we characterize the success of a reference in terms of its later performance in folding known structures. We first develop CG backbones from two homogeneous polypeptides that capture either α or β conformation and subsequently discuss two strategies to design hybrid references that encode both these secondary structure elements. In all but one case (the extended ensemble model, discussed later in this section), we parameterize the CG models at the folding temperatures of the AA references to sample both folded and extended conformations.

We use polyisoleucine and polyisovaline peptides as our first two reference AA simulations. These systems provide excellent contrasts because isoleucine has a reasonably high helix propensity,⁷¹ while valine has very high stability for β -sheets.⁷² We parameterize CG

TABLE I. Nomenclature for different CG peptide backbone forcefields.

Forcefield ^a	Reference sequence/(s)	Parameterization temp.
LEU ₁₅	L ₁₅	407 K (folding temp.)
VAL ₁₅	V ₁₅	367 K (folding temp.)
LEU ₆ VAL ₉	LVVVVVVVLLLVLL	360 K (folding temp.)
LEU ₁₅ + VAL ₁₅	L ₁₅ , V ₁₅	300 K (ambient temp.)

^aGō-like models derived from these backbone forcefields are also named similarly.

models from atomistic poly-leucine (LEU₁₅, Fig. 1) and from poly-valine (VAL₁₅) 15-mers, at their folding temperatures, using relative entropy minimization. We expect the LEU₁₅ CG model to preferentially stabilize α helices and the VAL₁₅ model to be more biased toward β conformations.

In addition, we examine two additional reference cases that are designed to combine secondary structure preferences from the leucine and valine systems. The first strategy investigates AA leucine-valine copolymers. We tested four randomly generated 15-mer sequences with ~30%–60% leucine (L) and ~70%–40% valine (V) content and found that the sequence LVVVVVVVLLLVLL (LEU₆VAL₉) has the most balanced helical (cluster fraction ~42% at 270 K) and hairpin (cluster fraction ~36%) conformations in AA simulations. Clearly, there may be other sequence combinations that produce even closer α and β populations, but we proceed with this particular instructive case.

The second hybrid strategy and the fourth reference AA case uses a distinct relative entropy approach to directly combine information from multiple atomistic reference simulations into a single CG model. This method is motivated by the extended-ensemble coarse-graining concept of Mullinax and Noid,⁴³ in which a CG model is parameterized simultaneously against an ensemble of AA references. While originally developed in the context of force-matching, we adapt that approach to the relative entropy framework. Here, an extended ensemble can be formulated as follows: the optimal CG model that minimizes information loss simultaneously from N AA references actually minimizes the sum of *individual* relative entropies between the model and each AA member of the ensemble, i.e.,

$$S_{\text{rel}} = \sum_i^N S_{\text{rel}}^{(i)}, \quad (5)$$

where $S_{\text{rel}}^{(i)}$ is the relative entropy between the CG model and the i th AA reference. Motivation for the linear form of Eq. (5) can be found in the [supplementary material](#). The extended-ensemble strategy was originally introduced for developing models that are independent of the thermodynamic state point at parameterization and hence transferable across a range of states,⁴³ but more generally, the ensemble AA members can also be chemically distinct,^{43,73} including different sequences.⁴⁴ In the present work, we use the extended-ensemble relative entropy method to parameterize a hybrid CG polypeptide model (LEU₁₅ + VAL₁₅) simultaneously from the poly-leucine and polyvaline AA simulations. It is unclear how an overall folding temperature should be defined for the extended AA ensemble, so we choose ambient temperature as the relevant parameterization temperature. Table I outlines the four different CG backbone forcefields studied in this work.

B. CG sidechain forcefields (Gō-like models)

The final CG strategy utilizes a Gō-like model for the sidechains, which requires the native structure of the target protein sequence as input. Residues that are in contact in the native structure (i.e., within a certain threshold proximity) are supplied attractive (or “native”) interactions, while others are provided with purely excluded volume (or “non-native”) interactions. We use a cutoff of 8 Å between the centers-of-mass of entire residues to determine contacts. Residues within a contact order of 3 along the sequence are considered chain-adjacent, and no intersidechain interactions are applied to them. The nonbonded native interaction ($u_{\text{pair}}^{\text{SS, native}}$) is represented using a spline with a cutoff of 10 Å. Nonbonded non-native interactions ($u_{\text{WCA}}^{\text{SS, non-native}}$) are represented by a Weeks-Chandler-Andersen (WCA)⁷⁴ potential with $\varepsilon = 4.2k_B T$ and $\sigma = 3.8$ Å, which mimics the inner repulsive core region of the optimized native potentials (see Sec. III B and Fig. 7). Since the models presented in this work are sequence-independent, σ can be interpreted as an overall minimum sidechain size that provides pure intersidechain excluded volume interactions through the non-native WCA potential. The overall CG forcefield is given by

$$U_{\text{CG}} = \underbrace{(U_b^{\text{BB}} + U_\theta^{\text{BB}} + U_{\varphi, \psi}^{\text{BB}} + U_{\text{pair}}^{\text{BB}})}_{\text{min. } S_{\text{rel}} \text{ from AA polypeptide}} + \underbrace{(U_b^{\text{BS}} + U_\theta^{\text{BS}} + U_{\varphi, \psi}^{\text{BS}} + U_{\text{pair}}^{\text{BS}})}_{\text{min. } S_{\text{rel}} \text{ from AA polypeptide}} + \underbrace{U_{\text{pair}}^{\text{SS, native}}}_{\text{2nd round of min. } S_{\text{rel}} \text{ from AA protein}} + U_{\text{WCA}}^{\text{SS, non-native}}. \quad (6)$$

We parameterize the native interactions by minimizing the relative entropy between an AA simulation of a candidate small protein and the CG model represented by the forcefield U_{CG} in Eq. (6). Note that this constitutes a *second* round of relative entropy optimization, after CG backbone parameterization. In this case, the backbone potentials are held fixed at their optimized values obtained from the earlier optimization, and only the spline knots in $U_{\text{pair}}^{\text{SS, native}}$ are allowed to vary. Figure 2 provides a flowchart for the overall two-step parameterization of the CG forcefields.

Specifically, we use an AA trajectory of the 20-residue trp-cage miniprotein [protein data bank (PDB) code: 1L2Y] to parameterize the native interactions. 1L2Y is a synthetic protein that folds rapidly to a globular structure. Its small size (one of the smallest proteinlike constructs) and fast folding times make it an ideal model system for

Stage 1: Parameterize backbone interactions (U_{CG}^{BB})

experimental and computational studies of protein folding mechanisms.⁷⁵ 1L2Y folds extremely fast ($\sim 4.1 \mu\text{s}$ *in vitro*)⁷⁶ such that the iterative CG MD simulations during the relative entropy procedure need not be excessively long. The trp-cage protein has well defined helical and flexible regions but no β sheets. Thus, in our Gō-like model, the task of capturing β -sheets is largely shouldered by the CG backbone forcefield, consistent with our hypothesis that the backbone interactions define the landscape of putative folds. Alternatively, one might develop native interactions either from a more $\alpha - \beta$ balanced protein (e.g., protein G) or using an extended ensemble of two (or more) short peptides that selectively stabilize α and β conformations; due to added computational complexity, we leave these directions for future work. It is instructive to note that, beyond the set of native contacts, we do not incorporate any other bioinformatic-type information in the Gō-like model; this contrasts with other Gō models in the literature which include *actual* native distances⁵¹ or additional hydrogen bond potential functions parameterized using secondary structure databases or fragment libraries.^{26,49}

In the CG models, glycine must be treated specially because it lacks a sidechain site for native-contact interactions. We tested a preliminary model (not reported here) that omitted native-contact interactions for glycine and found poor folding performance for 1L2Y (RMSD $\geq 5 \text{ \AA}$) in the flexible regions. Instead, we supply glycine residues in the CG model with a pseudosidechain that permits native contact interactions with other amino acids. Such a treatment is reasonable since hydrogen bond lengths (between glycine and other residues along the backbone) are typically less than the $8 \text{ \AA}$ cutoff used to determine native contacts. Details for the pseudosidechain procedure for glycine in the model are given in the [supplementary material](#). It is also worthwhile noting that glycine's

lack of a sidechain should also affect the backbone dihedral and secondary structure preferences.⁷⁷ Proline is also expected to have distinct preferences. Thus, a key assumption in our current models is that a single, amino-acid-unspecific set of backbone parameters is sufficient for capturing proteinlike folds to high accuracy in the backbone RMSD.

C. Simulation details

All simulations employ Replica Exchange Molecular Dynamics (REMD) for enhanced sampling of conformational space.⁷⁸ Atomistic polypeptide simulations are carried out using an in-house Python REMD wrapper for the Amber-16 MD engine.⁷⁹ In each simulation, 16 replicas spaced exponentially in temperature between 270 and 500 K run for 60 ns each with 5 swap attempts between neighboring replicas every 20 ps. Most simulations employ the Amber ff96 forcefield⁸⁰ with a modified version of the igb5 implicit solvent model,⁸¹ which we found to correctly fold a variety of helical and sheetlike secondary structures.^{82–84} Atomistic simulations of 1L2Y (60 ns) at 300 K are taken from a previous study using the Amber ff14SBonlisc forcefield⁸⁵ with a modified version of the igb5 implicit water model,⁸¹ which was reported to produce the best top-cluster structure prediction among other Amber forcefields tested.⁸⁴ Data from the last 20 ns are used to generate statistics.

CG models are simulated using the LAMMPS MD package (15 May-2015 version).⁸⁶ REMD CG simulations of polypeptides during relative entropy minimization and subsequent structure prediction runs using Gō-like models are performed for 500 ns/replica, and the last 100 ns is used to collect statistics. Relative entropy runs use 8 replicas, and structure prediction runs are simulated with 10 replicas, both types using temperatures distributed

FIG. 2. CG backbone potentials (U_{CG}^{BB}) are first extracted from CG polypeptide models parameterized by minimizing relative entropy from AA polyleucine (red), polyvaline (dark yellow), and a mixed leucine-valine sequence LVVVVVVLL-LVVLL. A hybrid CG backbone embedding both α helical (from polyleucine) and β -sheet (from polyvaline) characteristics is also parameterized by simultaneously minimizing the total relative entropy from both AA leucine and valine polymer references. Gō-like models are developed for each of these backbone forcefields, by parameterizing native interactions $U_{pair}^{SS,native}$ through a second round of relative entropy minimization with a AA simulation of the trp-cage miniprotein (1L2Y). Lines across the 1L2Y structure connect the native contacts.

exponentially in the range 270–500 K swapping every 10 ps. Data from replicas are reweighted using the Multistate Bennett Acceptance Ratio (MBAR) as implemented using the pymbar package.⁸⁷ Top-cluster structures from both atomistic and CG simulations are determined using a hierarchical K-means-like clustering algorithm based on the RMSD⁸² and aligned with native structures using the well-known Kabsch algorithm.⁸⁸ Cartoon representations of protein secondary structures are rendered using both PyMOL⁸⁹ and VMD.⁹⁰

III. RESULTS AND DISCUSSION

A. Folding behavior of CG polypeptide models

Figure 3 presents a comparison of the folding curves between AA and CG simulations of polyleucine and polyvaline. These curves are calculated from the fractions of temperature reweighted trajectories that fold to within 3 Å RMSD of the atomistic top cluster structures at 270 K (polyleucine: 97% top cluster fraction for a near-ideal helix and polyvaline: 59% top cluster fraction for a hairpin with a slightly twisted loop). The AA folding temperatures for both polypeptides are quite high (407 K for polyleucine and 367 K for polyvaline) which indicates the high stability of their corresponding folds. However, this is expected since the reference atomistic simulations use implicit solvent. CG polyleucine reproduces the temperature dependent folding behavior for helix formation of its AA reference reasonably well with a ~ 5 K deviation from the AA folding temperature, while CG polyvaline incurs a ~ 27 K error in the folding temperature. CG polyvaline is likely more flexible than its AA counterpart and exhibits unstructured regions at both free ends of the hairpin whose lengths vary significantly over different temperatures resulting in register-shifted hairpin conformations (not shown).

FIG. 3. Comparison of folding curves between atomistic (blue) and CG (red) simulations of 15-mers of leucine (panel A) and valine (panel B). The folding fraction at a particular temperature is calculated as the fraction of the (reweighted) trajectory that is within 3 Å RMSD of the reference AA top-cluster structure at 270 K (helix for polyleucine and hairpin for polyvaline). Reference (AA top cluster) and CG top cluster structures are shown in blue and red, respectively, in the insets. Folding temperatures for AA polyleucine (407 K) and polyvaline (367 K) are marked with vertical dashed lines. The CG LEU₁₅ model has nearly similar (within 5 K error) folding temperature as its atomistic counterpart, while the VAL₁₅ model underestimates the atomistic folding temperature by ~ 27 K.

Figure 4 compares the free energy surface at 280 K between the AA and CG models for polyleucine and polyvaline, as a function of radius of gyration (R_g) and backbone RMSD from the atomistic top-cluster structures. Both AA and CG polyleucine models exclusively stabilize a near-ideal helix, while those for polyvaline are centered around two similar ($\Delta F \leq 2.5k_B T$) hairpin conformations that are register shifted from one another. Interestingly, both the helix and the hairpin have a similar $R_g \sim 7.2$ – 7.5 Å. Both the folding curves in Fig. 3 and the free energy surfaces in Fig. 4 show that the polyvaline hairpin is harder to stabilize than the polyleucine helix. However, the free energy surfaces are very similar between AA and CG resolutions for both leucine and valine polymers and the top cluster conformations are captured nearly quantitatively in the CG models. The only major difference is that CG polyvaline has a broader sampling of large R_g (up to 9 Å) and RMSD (up to 6 Å) structures relative to its AA counterpart. Furthermore, it should be noted that classifying folded vs unfolded states in the CG valine polymer is based on examining alignment with the particular 270 K top cluster atomistic hairpin. If one were to look at the second and third most populated clusters for CG polyvaline at 280 K, these have relative populations of $\sim 16\%$ and 6.5% , respectively, and are all hairpins that are register-shifted from the AA reference (such effects have also been observed in atomistic MD simulations⁹¹).

As mentioned in Sec. II B, we investigated two approaches to design a hybrid leucine-valine CG backbone balanced in α and β structural propensity. We first present the LEU₆VAL₉ model in which the CG polypeptide is parameterized directly from a reference atomistic copolymer LVVVVVVVLLLVLL. Figure 5(a) compares the AA and CG folding curves with respect to both the AA polyleucine and polyvaline reference structures at 270 K—a helix

FIG. 4. Comparison between AA and CG simulations of folding free energy surfaces (ΔF) at 280 K, as functions of radius of gyration (R_g) and backbone RMSD from the atomistic top-cluster structure at 270 K, for LEU₁₅ (top panel) and VAL₁₅ (bottom panel) CG models. While the LEU₁₅ CG model exclusively stabilizes an ideal helix similar to its AA reference, the VAL₁₅ AA and CG models have significant populations of two closely similar hairpins that are register-shifted from each other. In either case, the top cluster structures are reproduced nearly quantitatively ($\text{RMSD} \leq 1$ Å). AA structures are shown in blue, while CG structures are in red.

FIG. 5. The LEU_6VAL_9 CG polypeptide model constructed from the leucine-valine copolymer LVVVVVVVLLLVLL. The panels in (a) compare the folding curves between AA (blue) and CG (red) simulations of the copolymer. The folding fractions are calculated as the percent of the (reweighted) trajectory that is within 3 Å RMSD of a given reference structure as shown. The left panel in (a) uses the pure atomistic polyvaline helix at 270 K (see Fig. 3) as reference, thus showing an α -helix fraction, while the right panel uses the pure atomistic polyvaline hairpin at 270 K reference (again, Fig. 3), showing a β -fraction. The dashed lines mark the folding temperature of 360 K for the AA copolymer, where the total $\alpha + \beta$ (i.e., the structured fraction of configurations) is 50%. Both α and β fractions in the CG model are lower than their AA counterparts. The panels in (b) compare the AA (left) and CG (right) free energy surfaces at 280 K as a function of the radius of the gyration (R_g) and RMSD from the polyvaline helix reference. Shown in red are the actual top cluster structures at 280 K from the CG copolymer MD simulation. The top α clusters are similar, while the top β cluster is register shifted in the CG model.

and hairpin, respectively—thus giving effective α and β fractions. At the atomistic folding temperature of 360 K where the AA copolymer has a 50% population of structured (i.e., $\alpha + \beta$) states, the CG α fraction is nearly zero, while the CG β fraction is $\sim 5\%$. The disagreement between the AA and CG β fractions is larger than in Fig. 3 for the homopeptides and is related to the presence of a significant amount of flexible conformations that have small hairpinlike regions, similar to that observed in Figs. 3 and 4, which is also illustrated by the register shift between the AA reference and the CG top cluster hairpin [panels in Fig. 5(b)]. Furthermore, the CG model explores a much wider range of structures relative to the AA copolymer as exemplified by the prevalence of low free energy ($\leq 5k_B T$) regions in the CG potential of mean force (PMF). The relative α and β populations in the AA mixed polymer are 42% and 36%, while in the CG model, these fractions are 52% and 18%, respectively. Thus, it appears that the LEU_6VAL_9 CG model incorporates some elements of both α and β secondary structure such that it might be a reasonable candidate for a protein backbone forcefield.

In a distinct approach, we parameterize the $\text{LEU}_{15} + \text{VAL}_{15}$ extended ensemble CG model simultaneously from separate

polyvaline and polyvaline references, minimizing the sum of relative entropies between the model and each reference. Figure 6 presents the folding behavior of this extended ensemble CG polypeptide model. Panel (a) shows the free energy surface (at 280 K) as a function of R_g and RMSD from the AA polyvaline helical top cluster structure at 270 K. The free energy surface has two minima separated by a $7.5\text{--}10k_B T$ barrier, representing helical and hairpin states. A Ramachandran plot in panel (b) shows the free-energy landscape in the space of dihedral angles (ϕ , ψ) along the polymer backbone and shows the relatively higher stability of α -helical over β states in this CG model. Interestingly, there is a low fraction of highly unstructured conformations ($\geq 15k_B T$) in the $\phi \in [60^\circ, 90^\circ]$, $\psi \in [-90^\circ, -60^\circ]$ region which is likely contributed by the more flexible CG polyvaline since this region is also observed exclusively in the Ramachandran plot (not reported here) of polyvaline simulated using the VAL_{15} CG model. Folding curves in panel (c) relative to the AA top-cluster structures of polyvaline and polyvaline (similar to the procedure used in Fig. 3) reinforce the relative stability of helical states in this mixed CG model as the β folding fraction is consistently less than 5% across all temperatures.

Although it seems that the $\text{LEU}_{15} + \text{VAL}_{15}$ CG model prefers α -helices over β sheets, we reiterate that folding behavior in Fig. 6 is characterized with reference to a single conformation and may not necessarily reflect true β content. It should be noted that, using a single structure for characterizing β -sheet (or more generally

FIG. 6. The $\text{LEU}_{15} + \text{VAL}_{15}$ CG polypeptide model constructed using the extended-ensemble method, by combining data from AA leucine and valine 15-mers. Both the (a) free energy surface as a function of R_g and RMSD from the polyvaline top cluster structure (near perfect helix) and (b) Ramachandran plot at 280 K reveal the presence of basins dominated by both helices and hairpins separated by a $\sim 7.5k_B T$ barrier. (c) Folding curves (constructed by considering trajectory fractions within 3 Å of the top clusters of AA polyvaline and polyvaline) show that β -fractions remain consistently lower than 5%. Here, solid curves correspond to AA results, while dashed curves correspond to CG simulations. (d) However, when used in a self-assembly simulation of six polypeptide chains, the $\text{LEU}_{15} + \text{VAL}_{15}$ forcefield produces an antiparallel beta-zipper structure (inset) with $\sim 80\%$ β -content, also shown by dominant off-diagonal patterns in the interresidue contact map (A–F refers to the 6 peptide chains).

secondary structure) content in the CG polypeptide models is motivated by the observation that implicit water AA poly-leucine and poly-valine fold exclusively into an α -helix and a β -hairpin, respectively, thereby providing unique labels to classify structures obtained from the CG polypeptide models. A thorough characterization of secondary structure for atomistic proteins involves analyzing the strength and directionality of hydrogen bonds and sidechain contacts between regions of the backbone close to each other,⁹² while for CG models without hydrogen and/or an atomistically resolved carbonyl group, such a characterization might depend entirely on detecting known secondary structure patterns from CG contact maps which is nontrivial.⁹³ Instead, we present a simpler alternative to assess the β -propensity of the LEU₁₅ + VAL₁₅ CG model in panel (d) of Fig. 6, by deploying it in a self-assembly simulation of six polypeptide chains. In this case, the trajectory averaged interresidue “global” contact map across all polymer chains contains exclusively parallel and antiparallel β sheets characterized by contact patterns that are off-diagonal. Peptide assemblies have enhanced interstrand hydrogen bonded interactions relative to a single hairpin and as such are expected to stabilize β -rich structures, unless the underlying forcefield has an inordinate amount of helical or flexible character. In fact, a clustering analysis reveals that the dominant self-assembled oligomer is an antiparallel steric zipper, which is known to be one of the most common motifs that can result from an assembly of short peptide fragments.⁹⁴ The self-assembly behavior of this CG polypeptide thus provides good evidence that it embeds β propensity in addition to the α character seen in the single-peptide case.

B. Native and non-native Gō-like interactions

With the candidate backbone forcefields just developed, we proceed to determine sidechain native contact potentials using relative entropy minimization. Figure 7 shows the native Gō-like

FIG. 7. Native interactions optimized from the atomistic trp-cage miniprotein (1L2Y) at 300 K, for the candidate backbone forcefields in Table I. All native potentials have an inner repulsive core near ~ 3.8 Å and are cut off at 10 Å. The non-native interaction is fixed as a WCA potential with $\sigma = 3.8$ Å and $\epsilon = 4.2 k_B T$.

interactions optimized from the atomistic trp-cage miniprotein (1L2Y), after fixing the backbone parameters, for the different forcefields, as summarized in Table I. The four native contact potentials have similar forms, with broad minima near ~ 4 Å, followed by an approximately linear decay until the cutoff of 10 Å. The widths of the different minima (~ 1 Å) are close to the average radius of gyration of atomistic 1L2Y sidechains (~ 1.12 Å) which leads us to hypothesize that the broadness in the minima indirectly encodes the variation in sidechain sizes. The optimized native interaction using the LEU₁₅ + VAL₁₅ backbone has the largest inner-core depth of $\sim 9.5 k_B T$, which is interesting because it exceeds that of both the native interactions derived using the LEU₁₅ and VAL₁₅ backbone forcefields. Also interesting is that near the 10 Å cutoff, all the potentials exhibit a small repulsive bump. To assess if this small repulsion results from cutting off the potential prematurely, we re-optimized the LEU₁₅ + VAL₁₅ native interaction using larger cutoffs of 15 and 20 Å (not reported here) but still observed the same behavior. One possible explanation is that the optimal native interaction resulting from relative entropy minimization attempts to match the pair correlation of native contact distances within the cutoff since a spline potential is used.⁶⁹ Since by construction the correlation is truncated at the cutoff distance, the effective interaction develops a repulsive character there such that it produces a confinement effect for the native contacts similar to harmonic restraints in elastic network models.⁵²

For the repulsive non-native interactions, we use a WCA potential. Based on the location of the inner core for the native potentials, we set σ for this potential to 3.8 Å and ϵ to 4.2 $k_B T$, which provides a purely repulsive potential that is not overly stiff. Unlike the native interactions, the non-native potential is not parameterized from the 1L2Y reference. However, its exact form is likely unimportant for structure prediction in the Gō-like model since it is applied to residues that are not in contact.

C. Performance of Gō-like models

To test the Gō-like models in Table I, we perform folding simulations on two sets of proteins, starting all simulations from fully extended conformations (all dihedral angles set to 180°). The first set comprises a collection of short peptide fragments (~ 11 –20 residues) with both helical and hairpin structures used in the study in Ref. 84. The second set is a collection of moderately large globular proteins (26–73 residues) taken from Ref. 95. Table II provides more details on these sequences.

Generally, the bottom-up Gō-like models do an excellent job of capturing native structures in folding simulations. Figure 8 shows the fraction of sequences whose structures are predicted within a given RMSD threshold for α , β , and $\alpha + \beta$ motifs outlined in Table II, using REMD simulations of the CG Gō-like models of Table I. The LEU₁₅ + VAL₁₅ model is most consistent in producing faithful native structure alignments. It predicts 80% of the pure helical and pure β sequences, and 50% (one of the two) mixed $\alpha + \beta$ sequences within 2 Å RMSD (details follow in Figs. 9 and 10). The LEU₆VAL₉ copolymer model is a close second best with 60% helical, 40% β , and 40% mixed sequences predicted within 2 Å RMSD. Models derived from pure poly-leucine and pure poly-valine references, unsurprisingly, are biased toward α and β sequences, respectively: LEU₁₅ resolves 80% of helical sequences within 2 Å, and VAL₁₅ captures 90% of β

TABLE II. Target peptide sequences for validating the Go-models (shorter sequences are tabulated above the dividing lines).

Name ^a	PDB code	Length	Structure ^b	RMSD ^c (Å)	CG T_{fold} ^d
1CB3	1CB3	11	α	2.5	n.a.
1L2Y	1L2Y	20	α	1.7	n.a.
2I9M	2I9M	17	α	1.6	n.a.
C peptide	^e	13	α	0.9	n.a.
EK peptide	^e	12	α	0.7	n.a.
15- β	^e	15	β	2.1	n.a.
1GB1	1GB1	16	β	2.0	419 K
1E0Q	1E0Q	17	β	1.9	419 K
1LE1	1LE1	12	β	1.2	n.a.
1J4M	1J4M	14	β	1.6	419–457 K ^a
Protein A	1BDD	46	α	1.7	n.a.
Albumin binding domain	2FS1	49	α	1.5	n.a.
α 3D	2A3D	73	α	4.9	n.a.
YJQ8 WW (res 7–31)	1E0N	27	β	3.3	n.a.
FBP28 WW (res 6–31)	1E0L	26	β	1.5	n.a.
Ubiquitin (res 1–35)	1UBQ	35	$\alpha + \beta$	3.2	n.a.
Protein G	1PGB	56	$\alpha + \beta$	2.0	n.a.
α -spectrin SH3	1SHG	57	β	1.4	419–457 K ^a
src-SH3	1SRL	56	β	1.8	450–500 K ^a
Bacterial flavodoxin	1FUE	163	$\alpha + \beta$	1.8	331–345 K ^a
TIM barrel monomer	8TIM	247	α/β	2.3	305–318 K ^a

^aRange indicates the closest two replica temperatures that bound the temperature at which 50% of the protein is folded, i.e., has a RMSD ≤ 3 Å from the native structure.

^b α and β are simplified structure classifications that neglect finer details such as 3–10 helices, β bulges, turns, coils, and loops.

^cTrajectory averaged values at 280 K.

^dSimulated using the LEU₁₅ + VAL₁₅ CG model. “n.a.” indicates that sequences exhibit >50% folding fraction across all temperatures in REMD simulations up to 500 K.

^eDetails of the native structure can be found in Ref. 84.

sequences within 2–4 Å. Interestingly, the polyleucine model also allows for some β character since it correctly predicts one of the two mixed sequences (protein G) barely within 2 Å, while the polyvaline model has the worst overall accuracy, consistently straying from the desired ≤ 2 Å RMSD alignment even for β -rich sequences. However, it is remarkable that most of the models produce reasonably close alignment for the shorter peptides. The only information beyond the backbone forcefield in Gō-like models comes from protein-specific favorable interactions between native contacts, so the ability to accurately fold shorter peptides with fewer contacts is arguably an important performance metric for such models. Success with the shorter peptides is thus directly linked to the efficiency of the backbone (s) in exploring the relevant dihedral space and reflects considerable promise for the bottom-up development of a CG peptide model. Furthermore, it seems like the models derived here are not too sensitive to the lack of diversity of native contacts in the training set for optimizing the native interactions, especially considering that we used only the single 1L2Y reference.

We proceed by characterizing in more detail the CG model that performs best in folding simulations, namely, the LEU₁₅ + VAL₁₅ backbone optimized using the extended ensemble method plus the

native interaction optimized with this backbone. Figures 9 and 10 show the structures of target sequences for this model superposed on the corresponding native structures. All CG REMD simulations are initialized from fully extended states. It is remarkable that such a simple CG model parameterized from *single* polypeptide references and using a single native contact interaction for all sidechain types, without explicitly encoding the contact distances or (experimentally known) interaction strengths, is able to correctly predict 74% (14 out of 19) of all the target sequences within 2 Å, attesting to an overall high prediction accuracy. Note, however, that the folding temperatures reported in Table II for the CG models are very high and in some cases above the maximum replica temperature, preventing their exact calculation. While this certainly disagrees with experiment or atomistic simulations, it shows that the bottom-up procedure with a Gō-like strategy embeds highly stabilizing native contact interactions that are biased toward the native state. Typically, a Gō-like model provides faster access to the native structure by smoothening and exaggerating the energy landscape.⁵² The high folding temperatures may also partially result from the use of implicit water reference polypeptide simulations for parameterizing the backbone CG model, which themselves have high

FIG. 8. Assessment of the accuracy of Gō-like CG models in Table I for predicting the structure of α -helical (top panel), β -rich (middle panel), and $\alpha + \beta$ (bottom panel) sequences (detailed in Table II). In each panel, stacked bar charts show the fraction of sequences that fold to within 2 Å (green), 2–4 Å (teal), and >4 Å (brown) RMSD from the native structure. The LEU₁₅ + VAL₁₅ CG backbone at 300 K, derived from an extended ensemble of polyleucine and polyvaline AA references, best-predicts all trial proteins. All RMSDs are ensemble-averaged values from trajectories at 290 K.

folding temperatures (Fig. 3). Native-centric models in the literature have used experimentally observed folding temperatures to fit the native interactions and thus by design provide quantitative agreement with experimental folding temperature.⁴⁹ Similar constraints

to obtain accurate folding temperatures might be added to the relative entropy minimization procedure, but we leave these directions to future work.

Differences from native structures in Figs. 9 and 10 result mostly from mispredicting flexible regions such as turns and loops, with the exception of 1L2Y, α -spectrin SH3, and src-SH3. It is no surprise that the flexible region for 1L2Y is well-captured since the native interactions are optimized from a 1L2Y reference. The only sequence that does not fold well with this Gō-like model is α 3D (Fig. 10), where large misalignment in the interconnecting turn regions (residues 20–26, 46–53) leads to different major axis vectors for two of the helices (residues 1–19 and 54–73). The fraction of native contacts satisfied for this sequence is 87% (calculated from the top-cluster) as opposed to ~97% for the other sequences. Other major misalignments in turn regions include the (Fig. 10) 35-mer fragment of ubiquitin (residues 8–11), (Fig. 9) 1GB1 (residues 7–10), and 15- β (residues 7–10). The prevalence of specific amino acids in these four sequences reveals 23% glycine and 11% proline in the difficult-to-capture flexible regions as opposed to 3% glycine and no proline in regions with well-defined secondary structures. This clearly indicates that cases in which the model fails to correctly fold a particular protein region is due to the use of generic backbone interactions that do not account for the unique structural propensities of glycine and proline. It suggests that a more accurate CG model could be achieved by incorporating special backbone interactions (angular and dihedral) for these residues. Such interactions may be optimized from an extended ensemble of AA references that contain glycine and proline polymers.

The CG REMD simulations of the target sequences are ostensibly ~30 times faster than their corresponding implicit water atomistic simulations for folding the short fragments in Fig. 9. For folding the longer sequences in Fig. 10, simulations take ~10 h of total

FIG. 9. Top cluster structures for short sequences (11–20 residues) predicted by the Gō-like model derived from the extended-ensemble LEU₁₅ + VAL₁₅ backbone. Native structures are in blue, while simulated ones are in red. The RMSD from the native structure (averaged from the trajectory at 290 K) is reported beside the sequence name. The average standard error (standard deviation/mean) in calculating the RMSDs is ~6.5%.

FIG. 10. Top cluster structures for longer sequences (26–73 residues) predicted by the Gō-like model derived from the extended-ensemble LEU₁₅ + VAL₁₅ backbone. Native structures are in blue, while simulated ones are in red. The RMSD from the native structure (averaged from the trajectory at 290 K) is reported beside the sequence name. The average standard error (standard deviation/mean) in calculating the RMSDs is ~6.5%.

central processing unit (CPU) time to achieve equilibration, starting from fully extended states. It is worth noting that in implicit water MD simulations of atomistic protein G, folding does not occur until ~54 μ s of simulation time and approximately 54 h of real time with the Amber14 MD software on GPUs,⁹⁶ while in explicit-water MD simulations of protein G, folding can take as long as 1154 μ s of simulation time on the Anton supercomputer.⁹⁷ While efficiency in our simulations is expected since Gō-like forcefields are known to produce fast, unfrustrated folders,⁵² the relative speed-up observed in this work (e.g., 30 $\times$ for short sequences), nevertheless, provides a qualitative upper bound on the expected speed-up over AA forcefields using these four site CG models.

As a final set of tests, we use the LEU₁₅ + VAL₁₅ Gō-like model to fold two much longer sequences that presently would be impossible to tackle with all-atom simulations: the 163 residue flavodoxin encoded in the *H. pylori* genome (PDB code: 1FUE) and the 247 residue Triose-Phosphate-Isomerase (TIM) barrel monomer (PDB code for the dimer: 8TIM). Flavodoxin is characterized by a five-stranded parallel β sheet core sandwiched by α helices, while the TIM barrel is a highly common fold consisting of eight overlapping, alternating β – α – β supersecondary structures. Starting with fully extended conformations, the Gō-like model achieves

remarkable prediction accuracies in each case (as demonstrated in Fig. 11), scored by trajectory averaged RMSDs of 1.8 Å for flavodoxin and 2.3 Å for the TIM barrel, which are attained within ~100 ns of simulation time spanning ~2 days of total CPU time. To

FIG. 11. Top cluster structures for bacterial flavodoxin (163 residues) and the TIM barrel protein (247 residues) predicted by the Gō-like model derived from the extended-ensemble LEU₁₅ + VAL₁₅ backbone. Native structures are in blue, while simulated ones are in red. The RMSD from the native structure (averaged from the trajectory at 290 K) is reported beside the sequence name. The average standard error (standard deviation/mean) in calculating the RMSDs is ~6%.

FIG. 12. A test of robustness of the Gō-like model derived from the extended ensemble LEU₁₅ + VAL₁₅ backbone forcefield. This model is used in CG REMD simulations of protein G that progressively remove native-contact information by deleting zero to 20% of the native contacts. The RMSD with the native structure (ensemble averaged from the 290 K trajectory) varies between 2 and 5.6 Å and has a standard error (standard deviation/mean) of ~6.5%. The prediction quality does not decrease monotonically since contacts are removed randomly. Native and predicted structures are colored blue and red, respectively.

the best of our knowledge, CG Gō-like models have not been capable of this level of structurally accurate folding simulations for such large proteins.

D. Fault tolerance of the LEU₁₅ + VAL₁₅ Gō-like model

The predictive ability of Gō-like models depends on the accuracy of native contact information, which in practice might be provided by experiments or other computational strategies that are subject to errors. An important question is to what extent such models are robust to missing native contact information. As an initial assessment of such error, we simulate the folding of protein G with a randomly deleted fraction of the native contacts using the extended ensemble Gō-like potential (the LEU₁₅ + VAL₁₅ model). Figure 12 shows the prediction quality with an increasing fraction of protein G's contacts. We find that the prediction quality deteriorates from 2 Å with complete native contact information to 5.6 Å when 20% of native contacts are removed. The fraction of the reduced set of native contacts that are satisfied by the end of the simulation fluctuates between 76% and 80% (98% when all contacts are retained). The prediction quality does not decrease monotonically because the contacts are removed at random in each case, and the ones that are relatively more important, such as those closer to the folding core, may be better retained in some cases even when a larger overall number of contacts are removed. (Only a single case is considered for each contact-removed model.) It is interesting to note, however, that the simulated structures retain the helix in all cases with relatively high native alignment, while the β character fluctuates. Moreover, even when 20% of the contacts are removed, the Gō-like model still gives a structural collapse that achieves the correct spatial proximity of the helical and hairpin regions. We do not probe further into which contacts are more important to retain over others, or what the critical fraction of retained contacts needs to be for a desired accuracy, but such directions may be interesting to study in the future. Nevertheless, Fig. 12 shows that at least for protein G, the extended ensemble CG backbone produces Gō-like models that can reasonably tolerate low to moderate levels of missing information.

IV. CONCLUSIONS

In this work, we explored whether or not bottom-up coarse graining techniques could construct CG protein-backbone

forcefields, which when combined with Gō-like sidechain interactions produce models with structurally accurate folding behavior. We developed four-site CG polypeptide models using relative entropy minimization with atomistic, single peptide references. We then augmented backbone interactions developed from these cases with Gō-like sidechain potentials for native contact interactions; we further optimized the functional form of these native-contact potentials with relative entropy minimization using a miniprotein reference. While Gō-like sidechain interactions require the native structure as an input and thus do not enable *de novo* structure predictions, they provide an excellent upper bound for the capability of CG backbone interactions in sampling the relevant landscape of proteinlike folds and set the stage for subsequent models that incorporate chemistry-specific sidechain interactions not dependent on native contact information.

We found that the choice of the AA reference system is important to provide a broadly applicable CG model. CG models derived from pure polyleucine and polyvaline are biased toward the secondary structure propensities of the references. While combining leucine and valine secondary structure behaviors using a 40% leucine–60% valine copolymer improved structure prediction to some extent, a better strategy proved to be an extended-ensemble approach to parameterize a CG polypeptide forcefield simultaneously to the two reference leucine and valine simulations. Our best and final model proved to be the LEU₁₅ + VAL₁₅ extended-ensemble model, which could ultimately resolve simulated structures of a variety of small to large proteins within 2 Å of the native conformations, when used in conjunction with Gō-like sidechain interactions.

These tests highlight the potential of bottom-up CG methods toward the development of accurate protein models. Indeed, it seems rather remarkable that a simple bottom-up CG model with only native contact information can capture the structures of up to ~247 residue globular proteins to within 2.5 Å. This work also underscores the importance of the backbone in defining the fold landscape and of the sidechain interactions in filtering for specific conformations;^{63–66} in other words, the backbone interactions play a significant role in sampling the relevant dihedral space and secondary structures. The results here show that sidechain interactions can be described at a lower level of detail than that required for the backbone, in agreement with several earlier works showing the same result.^{23,25,34,98} It may be noted, however, that even after ignoring sidechain size and packing, the conformational

diversity afforded by the LEU₁₅ + VAL₁₅ backbone forcefield enables the Go-like model to reconstruct the cores of large globular proteins correctly.

The relative-entropy-optimized, bottom-up, four-site CG models of peptides developed here hold considerable promise as putative CG protein forcefields that can offer predictive insight into folding and oligomerization processes. Moreover, while native-aware sidechain potentials do not offer *de novo* capabilities, the Gō-like model may be highly useful for quickly enumerating structurally realistic conformations that are compatible with a reduced set of constraints. For instance, it might facilitate sampling of conformational fluctuations in large proteins or enable a preliminary screening of allowed conformations of flexible peptide ligands docking onto large macromolecules. The Gō-like model would be particularly useful to structure prediction scenarios where contact and/or secondary structure constraints are available, e.g., through experiments or bioinformatic techniques. This could be extremely helpful to quickly narrow down NMR-compatible ensembles of structures, given nuclear Overhauser enhancement (NOE) constraints.

Despite its success, there are several obvious routes to improve the capabilities of the present Gō-like model. One possibility would be to replace the detailed spline interaction for native contacts (consisting of 40 knots) with simpler functional forms, such as Lennard-Jones potentials with variable exponents. This may avoid overfitting to the training data (in this case, the reference 1L2Y simulation) and in the future allow amino-acid specific sidechain interactions that can be combined with each other using simple mixing rules. Another immediate area of improvement includes accounting for glycine and proline through special backbone potentials, which is expected to improve predictions of turn and loop regions in which these amino acids are overrepresented. Finally, one might include in the model optimization experimental data such as folding temperatures using constraints in the relative entropy minimization algorithm, perhaps by using Bayesian inference based approaches.^{19,99}

This work provides a step and proof of concept toward developing completely bottom-up and physics based CG models of proteins. Here, we used Gō-like sidechain interactions to systematically evaluate the quality of backbone forcefields developed from short polypeptide references. An attractive next step would be to use the relative entropy framework, or other coarse-graining technique, to parameterize non-Gō models with residue specific sidechain interactions. An instructive strategy would be to first examine reduced alphabet models wherein amino acids are broadly classified as hydrophobic, hydrophilic, cationic, anionic, etc., before extending to full alphabet models with distinct parameters for all 20 amino acids. In this context, the results presented in this work are encouraging as a starting point for the systematic, bottom-up development of sequence/chemistry based CG protein models with a purely physical origin and that readily adapts to non-native sequences, synthetic amino acids, and even nonpolypeptide heteropolymer platforms.

SUPPLEMENTARY MATERIAL

See [supplementary material](#) for a derivation of the extended-ensemble strategy in the relative entropy framework as well as

LAMMPS compatible forcefield files for the extended-ensemble CG model.

ACKNOWLEDGMENTS

T.S. and M.S.S. acknowledge funding from the National Science Foundation through Award Nos. CHEM-1300770 and CHEM-1800344 and the Materials World Network Project Award No. 1312548. J.M. acknowledges funding from the National Institutes of Health through Award No. R01GM120537. Computational support was provided by the Center for Scientific Computing supported by the California NanoSystems Institute and the Materials Research Science and Engineering Center (MRSEC) at UC Santa Barbara through Grant Nos. NSF DMR 1720256 and NSF CNS 1725797.

REFERENCES

- 1 J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips, "A three-dimensional model of the myoglobin molecule obtained by x-ray analysis," *Nature* **181**, 662–666 (1958).
- 2 M. F. Perutz, M. G. Rossmann, A. F. Cullis, H. Muirhead, G. Will, and A. C. T. North, "Structure of haemoglobin: A three-dimensional Fourier synthesis at 5.5-Å. Resolution, obtained by x-ray analysis," *Nature* **185**, 416–422 (1960).
- 3 J. A. Brito and M. Archer, "X-ray crystallography," in *Practical Approaches to Biological Inorganic Chemistry*, edited by R. R. Crichton and R. O. Louro (Elsevier, Oxford, 2013), Chap. 9, pp. 217–255.
- 4 C. Nitsche and G. Otting, "NMR studies of ligand binding," *Curr. Opin. Struct. Biol.* **48**, 16–22 (2018).
- 5 C. M. Davis, M. Gruebele, and S. Sukenik, "How does solvation in the cell affect protein folding and binding?," *Curr. Opin. Struct. Biol.* **48**, 23–29 (2018).
- 6 K. Murata and M. Wolf, "Cryo-electron microscopy for structural analysis of dynamic biological macromolecules," *Biochim. Biophys. Acta, Gen. Subj.* **1862**, 324–334 (2018).
- 7 B. Moree, K. Connell, R. B. Mortensen, C. T. Liu, S. J. Benkovic, and J. Salafsky, "Protein conformational changes are detected and resolved site specifically by second-harmonic generation," *Biophys. J.* **109**, 806–815 (2015).
- 8 I. Sorokina and A. Mushegian, "Modeling protein folding *in vivo*," *Biol. Direct* **13**, 13 (2018).
- 9 H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic Acids Res.* **28**, 235–242 (2000).
- 10 T. Schwede, A. Sali, B. Honig, M. Levitt, H. M. Berman, D. Jones, S. E. Brenner, S. K. Burley, R. Das, N. V. Dokholyan, R. L. Dunbrack, K. Fidelis, A. Fiser, A. Godzik, Y. J. Huang, C. Humblet, M. P. Jacobson, A. Joachimiak, S. R. Krystek, T. Kortemme, A. Kryshtafovych, G. T. Montelione, J. Mout, D. Murray, R. Sanchez, T. R. Sosnick, D. M. Standley, T. Stouch, S. Vajda, M. Vaszquez, J. D. Westbrook, and I. A. Wilson, "Outcome of a workshop on applications of protein models in biomedical research," *Structure* **17**, 151–159 (2009).
- 11 G. A. Khoury, J. Smadbeck, C. A. Kieslich, and C. A. Floudas, "Protein folding and de novo protein design for biotechnological applications," *Trends Biotechnol.* **32**, 99–109 (2014).
- 12 P. Sledz and A. Caflisch, "Protein structure-based drug design: From docking to molecular dynamics," *Curr. Opin. Struct. Biol.* **48**, 93–102 (2018).
- 13 J. Mittal and R. B. Best, "Dependence of protein folding stability and dynamics on the density and composition of macromolecular crowders," *Biophys. J.* **98**, 315–320 (2010).
- 14 D. E. Shaw, R. O. Dror, J. K. Salmon, J. P. Grossman, K. M. Mackenzie, J. A. Bank, C. Young, M. M. Deneroff, B. Batson, K. J. Bowers, E. Chow, M. P. Eastwood, D. Ierardi, J. L. Klepeis, J. Kuskin, R. H. Larson, K. Lindorff-Larsen, P. Maragakis, M. A. Moraes, S. Piana, Y. Shan, and B. Towles,

"Millisecond-scale molecular dynamics simulations on anton," in *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis* (IEEE, 2009), pp. 1–11.

- ¹⁵M. Shirts and V. S. Pande, "Screen savers of the world unite!," *Science* **290**, 1903–1904 (2000).
- ¹⁶M. Levitt and A. Warshel, "Computer simulation of protein folding," *Nature* **253**, 694–698 (1975).
- ¹⁷V. Tozzini, "Coarse-grained models for proteins," *Curr. Opin. Struct. Biol.* **15**, 144–150 (2005).
- ¹⁸A. Morriss-Andrews and J.-E. Shea, "Simulations of protein aggregation: Insights from atomistic and coarse-grained models," *J. Phys. Chem. Lett.* **5**, 1899–1908 (2014).
- ¹⁹B. Webb, K. Lasker, J. Velázquez-Muriel, D. Schneidman-Duhovny, R. Pellarin, M. Bonomi, C. Greenberg, B. Raveh, E. Tjioe, D. Russel, and A. Sali, "Modeling of proteins and their assemblies with the integrative modeling platform," in *Structural Genomics: General Applications, Methods in Molecular Biology*, edited by Y. W. Chen (Humana Press, Totowa, NJ, 2014), pp. 277–295.
- ²⁰S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A. E. Dawid, and A. Kolinski, "Coarse-grained protein models and their applications," *Chem. Rev.* **116**, 7898–7936 (2016).
- ²¹K. A. Dill, S. Bromberg, K. Yue, H. S. Chan, K. M. Fiebig, D. P. Yee, and P. D. Thomas, "Principles of protein folding: A perspective from simple exact models," *Protein Sci.* **4**, 561–602 (1995).
- ²²K. A. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl, "The protein folding problem," *Annu. Rev. Biophys.* **37**, 289–316 (2008).
- ²³A. Voegler Smith and C. K. Hall, "α-helix formation: Discontinuous molecular dynamics on an intermediate-resolution protein model," *Proteins: Struct., Funct., Bioinf.* **44**, 344–360 (2001).
- ²⁴A. Kolinski, "Protein modeling and structure prediction with a reduced representation," *Acta Biochim. Pol.* **51**, 349–371 (2004).
- ²⁵T. Bereau and M. Deserno, "Generic coarse-grained model for protein folding and aggregation," *J. Chem. Phys.* **130**, 235106 (2009).
- ²⁶A. Davtyan, N. P. Schafer, W. Zheng, C. Clementi, P. G. Wolynes, and G. A. Papoian, "AWSEM-MD: Protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing," *J. Phys. Chem. B* **116**, 8494–8503 (2012).
- ²⁷L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S.-J. Marrink, "The martini coarse-grained force field: Extension to proteins," *J. Chem. Theory Comput.* **4**, 819–834 (2008).
- ²⁸M. Pasi, R. Lavery, and N. Ceres, "Palace: A coarse-grain protein model for studying mechanical properties," *J. Chem. Theory Comput.* **9**, 785–793 (2013).
- ²⁹P. Kar, S. M. Gopal, Y.-M. Cheng, A. Predeus, and M. Feig, "PRIMO: A transferable coarse-grained force field for proteins," *J. Chem. Theory Comput.* **9**, 3769–3788 (2013).
- ³⁰N. Basdevant, D. Borgis, and T. Ha-Duong, "Modeling protein–protein recognition in solution using the coarse-grained force field scorpion," *J. Chem. Theory Comput.* **9**, 803–813 (2013).
- ³¹F. Sterpone, S. Melchionna, P. Tuffery, S. Pasquali, N. Mousseau, T. Cragno-lini, Y. Chebaro, J.-F. St-Pierre, M. Kalimeri, A. Barducci, Y. Laurin, A. Tek, M. Baaden, P. H. Nguyen, and P. Derreumaux, "The OPEP protein model: From single molecules, amyloid formation, crowding and hydrodynamics to DNA/RNA systems," *Chem. Soc. Rev.* **43**, 4871–4893 (2014).
- ³²A. Liwo, M. Baranowski, C. Czaplinski, E. Gołaś, Y. He, D. Jagiela, P. Krupa, M. Maciejczyk, M. Makowski, M. A. Mozolewska, A. Niadzvedtski, S. Ołdziej, H. A. Scheraga, A. K. Sieradzian, R. Ślusarz, T. Wirecki, Y. Yin, and B. Zaborowski, "A unified coarse-grained model of biological macromolecules based on mean-field multipole–multipole interactions," *J. Mol. Model.* **20**, 2306 (2014).
- ³³S. P. Carmichael and M. S. Shell, "A new multiscale algorithm and its application to coarse-grained peptide models for self-assembly," *J. Phys. Chem. B* **116**, 8383–8393 (2012).
- ³⁴H. D. Nguyen and C. K. Hall, "Spontaneous fibril formation by polyalanine; discontinuous molecular dynamics simulations," *J. Am. Chem. Soc.* **128**, 1890–1901 (2006).
- ³⁵A. J. Marchut and C. K. Hall, "Side-chain interactions determine amyloid formation by model polyglutamine peptides in molecular dynamics simulations," *Biophys. J.* **90**, 4574–4584 (2006).
- ³⁶T. Bereau and M. Deserno, "Enhanced sampling of coarse-grained transmembrane-peptide structure formation from hydrogen-bond replica exchange," *J. Membr. Biol.* **248**, 395–405 (2015).
- ³⁷S. J. Marrink, A. H. de Vries, and A. E. Mark, "Coarse grained model for semiquantitative lipid simulations," *J. Phys. Chem. B* **108**, 750–760 (2004).
- ³⁸Y. C. Kim and G. Hummer, "Coarse-grained models for simulations of multi-protein complexes: Application to ubiquitin binding," *J. Mol. Biol.* **375**, 1416–1433 (2008).
- ³⁹S. J. Marrink and D. P. Tieleman, "Perspective on the martini model," *Chem. Soc. Rev.* **42**, 6801 (2013).
- ⁴⁰S. Izvekov and G. A. Voth, "A multiscale coarse-graining method for biomolecular systems," *J. Phys. Chem. B* **109**, 2469–2473 (2005).
- ⁴¹J. Zhou, I. F. Thorpe, S. Izvekov, and G. A. Voth, "Coarse-grained peptide modeling using a systematic multiscale approach," *Biophys. J.* **92**, 4289–4303 (2007).
- ⁴²J. F. Rudzinski and W. G. Noid, "Bottom-up coarse-graining of peptide ensembles and helix-coil transitions," *J. Chem. Theory Comput.* **11**, 1278–1291 (2015).
- ⁴³J. W. Mullinax and W. G. Noid, "Extended ensemble approach for deriving transferable coarse-grained potentials," *J. Chem. Phys.* **131**, 104110 (2009).
- ⁴⁴J. W. Mullinax and W. G. Noid, "Recovering physical potentials from a model protein databank," *Proc. Natl. Acad. Sci. U. S. A.* **107**, 19867–19872 (2010).
- ⁴⁵C. Dalgicdir, C. Globisch, M. Sayar, and C. Peter, "Representing environment-induced helix-coil transitions in a coarse grained peptide model," *Eur. Phys. J.: Spec. Top.* **225**, 1463–1481 (2016).
- ⁴⁶S. P. Carmichael and M. S. Shell, "Entropic (de)stabilization of surface-bound peptides conjugated with polymers," *J. Chem. Phys.* **143**(24), 243103 (2015).
- ⁴⁷H. Taketomi, Y. Ueda, and N. Gö, "Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effect of specific amino acid sequence represented by specific inter-unit interactions," *Int. J. Pept. Protein Res.* **7**, 445–459 (1975).
- ⁴⁸J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, "Funnel, pathways, and the energy landscape of protein folding: A synthesis," *Proteins* **21**, 167–195 (1995).
- ⁴⁹D. K. Klimov and D. Thirumalai, "Mechanisms and kinetics of β-hairpin formation," *Proc. Natl. Acad. Sci. U. S. A.* **97**, 2544–2549 (2000).
- ⁵⁰C. Clementi, H. Nymeyer, and J. N. Onuchic, "Topological and energetic factors: What determines the structural details of the transition state ensemble and 'en-route' intermediates for protein folding? An investigation for small globular proteins," *J. Mol. Biol.* **298**, 937–953 (2000).
- ⁵¹J. Karanicolas and C. L. Brooks, "The origins of asymmetry in the folding transition states of protein L and protein G," *Protein Sci.* **11**, 2351–2361 (2002).
- ⁵²R. Hills, C. Brooks, R. D. Hills, and C. L. Brooks, "Insights from coarse-grained gō models for protein folding and dynamics," *Int. J. Mol. Sci.* **10**, 889–905 (2009).
- ⁵³J. Mittal and R. B. Best, "Thermodynamics and kinetics of protein folding under confinement," *Proc. Natl. Acad. Sci. U. S. A.* **105**, 20233–20238 (2008).
- ⁵⁴P. Tian and R. B. Best, "Structural determinants of misfolding in multidomain proteins," *PLoS Comput. Biol.* **12**, e1004933 (2016).
- ⁵⁵J. K. Noel, P. C. Whitford, K. Y. Sanbonmatsu, and J. N. Onuchic, "SMOG@ctbp: Simplified deployment of structure-based models in GROMACS," *Nucleic Acids Res.* **38**, W657–W661 (2010).
- ⁵⁶K. Chen, J. Eargle, J. Lai, H. Kim, S. Abeyirigunawardena, M. Mayerle, S. Woodson, T. Ha, and Z. Luthey-Schulten, "Assembly of the five-way junction in the ribosomal small subunit using hybrid MD–gō simulations," *J. Phys. Chem. B* **116**, 6819–6831 (2012).
- ⁵⁷B. Lutz, C. Sinner, G. Heuermann, A. Verma, and A. Schug, "eSBMTools 1.0: enhanced native structure-based modeling tools," *Bioinformatics* **29**, 2795–2796 (2013).
- ⁵⁸J. F. Rudzinski and T. Bereau, "Structural-kinetic-thermodynamic relationships identified from physics-based molecular simulation models," *J. Chem. Phys.* **148**, 204111 (2018).

- ⁵⁹T. Head-Gordon, F. H. Stillinger, M. H. Wright, and D. M. Gay, "Poly(l)-alanine as a universal reference material for understanding protein energies and structures," *Proc. Natl. Acad. Sci. U. S. A.* **89**, 11513–11517 (1992).
- ⁶⁰N. L. Dawson, T. E. Lewis, S. Das, J. G. Lees, D. Lee, P. Ashford, C. A. Orengo, and I. Sillitoe, "CATH: An expanded resource to predict protein function through structure and sequence," *Nucleic Acids Res.* **45**, D289–D295 (2017).
- ⁶¹P. Cossio, A. Trovato, F. Pietrucci, F. Seno, A. Maritan, and A. Laio, "Exploring the universe of protein structures beyond the protein data bank," *PLoS Comput. Biol.* **6**, e1000957 (2010).
- ⁶²P. Kukic, A. Kannan, M. J. J. Dijkstra, S. Abeln, C. Camilloni, and M. Vendruscolo, "Mapping the protein fold universe using the camtub force field in molecular dynamics simulations," *PLoS Comput. Biol.* **11**, e1004435 (2015).
- ⁶³P. G. Wolynes, "Symmetry and the energy landscapes of biomolecules," *Proc. Natl. Acad. Sci. U. S. A.* **93**, 14249–14255 (1996).
- ⁶⁴D. Baker, "A surprising simplicity to protein folding," *Nature* **405**, 39–42 (2000).
- ⁶⁵B. Fain and M. Levitt, "Funnel sculpting for *in silico* assembly of secondary structure elements of proteins," *Proc. Natl. Acad. Sci. U. S. A.* **100**, 10700–10705 (2003).
- ⁶⁶J. R. Banavar and A. Maritan, "Physics of proteins," *Annu. Rev. Biophys. Biomol. Struct.* **36**, 261–280 (2007).
- ⁶⁷M. S. Shell, "Coarse graining with the relative entropy," in *Advances in Chemical Physics* (John Wiley & Sons, Ltd., 2016), pp. 395–441.
- ⁶⁸M. S. Shell, "The relative entropy is fundamental to multiscale and inverse thermodynamic problems," *J. Chem. Phys.* **129**, 144108 (2008).
- ⁶⁹A. Chaimovich and M. S. Shell, "Coarse-graining errors and numerical optimization using a relative entropy framework," *J. Chem. Phys.* **134**, 094112 (2011).
- ⁷⁰J. F. Rudzinski and W. G. Noid, "Coarse-graining entropy, forces, and structures," *J. Chem. Phys.* **135**, 214101 (2011).
- ⁷¹C. Pace and J. Scholtz, "A helix propensity scale based on experimental studies of peptides and proteins," *Biophys. J.* **75**, 422–427 (1998).
- ⁷²J. S. Nowick and S. Insaf, "The propensities of amino acids to form parallel β -sheets," *J. Am. Chem. Soc.* **119**, 10903–10908 (1997).
- ⁷³J. F. Rudzinski, K. Lu, S. T. Milner, J. K. Maranas, and W. G. Noid, "Extended ensemble approach to transferable potentials for low-resolution coarse-grained models of ionomers," *J. Chem. Theory Comput.* **13**, 2185–2201 (2017).
- ⁷⁴J. D. Weeks, D. Chandler, and H. C. Andersen, "Role of repulsive forces in determining the equilibrium structure of simple liquids," *J. Chem. Phys.* **54**, 5237–5247 (1971).
- ⁷⁵J. W. Neidigh, R. M. Fesinmeyer, and N. H. Andersen, "Designing a 20-residue protein," *Nat. Struct. Mol. Biol.* **9**, 425 (2002).
- ⁷⁶L. Qiu, S. A. Pabit, A. E. Roitberg, and S. J. Hagen, "Smaller and faster: The 20-residue trp-cage protein folds in 4 μ s," *J. Am. Chem. Soc.* **124**, 12952–12953 (2002).
- ⁷⁷B. K. Ho and R. Brasseur, "The ramachandran plots of glycine and pre-proline," *BMC Struct. Biol.* **5**, 14 (2005).
- ⁷⁸Y. Sugita and Y. Okamoto, "Replica-exchange molecular dynamics method for protein folding," *Chem. Phys. Lett.* **314**, 141–151 (1999).
- ⁷⁹D. Case, R. Betz, D. Cerutti, T. Cheatham III, T. Darden, R. Duke, T. Giese, H. Gohlke, A. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T. Lee, S. LeGrand, P. Li, C. Lin, T. Luchko, R. Luo, B. Madej, D. Mermelstein, K. Merz, G. Monard, H. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, D. Roe, A. Roitberg, C. Sagui, C. Simmerling, W. Botello-Smith, J. Swails, R. Walker, J. Wang, R. Wolf, X. Wu, L. Xiao, and P. Kollman, *Amber 2016* (University of California, San Francisco, 2016).
- ⁸⁰P. A. Kollman, R. Dixon, W. Cornell, T. Fox, C. Chipot, and A. Pohorille, "The development/application of a 'minimalist' organic/biochemical molecular mechanic force field using a combination of *ab initio* calculations and experimental data," in *Computer Simulation of Biomolecular Systems*, edited by A. Wilkinson, P. Weiner, and W. F. van Gunsteren (Elsevier, 1997), Vol. 3, pp. 83–96.
- ⁸¹R. Geney, M. Layten, R. Gomperts, V. Hornak, and C. Simmerling, "Investigation of salt bridge stability in a generalized born solvent model," *J. Chem. Theory Comput.* **2**, 115–127 (2006).
- ⁸²M. S. Shell, R. Ritterson, and K. A. Dill, "A test on peptide stability of amber force fields with implicit solvation," *J. Phys. Chem. B* **112**, 6878–6886 (2008).
- ⁸³E. Lin and M. S. Shell, "Convergence and heterogeneity in peptide folding with replica exchange molecular dynamics," *J. Chem. Theory Comput.* **5**, 2062–2073 (2009).
- ⁸⁴M. K. Robinson, J. I. Monroe, and M. S. Shell, "Are amber force fields and implicit solvation models additive? A folding study with a balanced peptide test set," *J. Chem. Theory Comput.* **12**, 5631–5642 (2016).
- ⁸⁵J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, "ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB," *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).
- ⁸⁶S. Plimpton, "Fast parallel algorithms for short-range molecular dynamics," *J. Comput. Phys.* **117**, 1–19 (1995).
- ⁸⁷M. R. Shirts and J. D. Chodera, "Statistically optimal analysis of samples from multiple equilibrium states," *J. Chem. Phys.* **129**, 124105 (2008).
- ⁸⁸W. Kabsch, "A solution for the best rotation to relate two sets of vectors," *Acta Crystallogr. Sect. A* **32**, 922–923 (1976).
- ⁸⁹Schrödinger, LLC, The PyMOL molecular graphics system, version 1.8, 2015.
- ⁹⁰W. Humphrey, A. Dalke, and K. Schulten, "VMD—visual molecular dynamics," *J. Mol. Graphics* **14**, 33–38 (1996).
- ⁹¹P. Kůhrová, A. De Simone, M. Otyepka, and R. B. Best, "Force-field dependence of chignolin folding and misfolding: Comparison with experiment and redesign," *Biophys. J.* **102**, 1897–1906 (2012).
- ⁹²D. Frishman and P. Argos, "Knowledge-based protein secondary structure assignment," *Proteins: Struct., Funct., Bioinf.* **23**, 566–579 (1995).
- ⁹³S. P. Carmichael, "Multiscale methodologies and simulations of emergent self-assembly phenomena," Ph.D. thesis, copyright—Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated June 05, 2016.
- ⁹⁴M. R. Sawaya, S. Sambashivan, R. Nelson, M. I. Ivanova, S. A. Sievers, M. I. Apostol, M. J. Thompson, M. Balbirnie, J. J. W. Wiltzius, H. T. McFarlane, A. O. Madsen, C. Riekel, and D. Eisenberg, "Atomic structures of amyloid cross- β spines reveal varied steric zippers," *Nature* **447**, 453–457 (2007).
- ⁹⁵S. B. Ozkan, G. A. Wu, J. D. Chodera, and K. A. Dill, "Protein folding by zipping and assembly," *Proc. Natl. Acad. Sci. U. S. A.* **104**, 11987–11992 (2007).
- ⁹⁶H. Nguyen, J. Maier, H. Huang, V. Perrone, and C. Simmerling, "Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent," *J. Am. Chem. Soc.* **136**, 13959–13962 (2014).
- ⁹⁷K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, "How fast-folding proteins fold," *Science* **334**, 517–520 (2011).
- ⁹⁸W. G. Noid, "Perspective: Coarse-grained models for biomolecular systems," *J. Chem. Phys.* **139**, 090901 (2013).
- ⁹⁹A. Perez, J. L. MacCallum, and K. A. Dill, "Accelerating molecular simulations of proteins using Bayesian inference on weak information," *Proc. Natl. Acad. Sci. U. S. A.* **112**, 11846–11851 (2015).