Editors' Suggestion

Effect of stereochemical constraints on the structural properties of folded proteins

Jack A. Logan,[1,*] Jacob Sumner,[2,3,*] Alex T. Grigas,[2,3] Mark D. Shattuck,[4] and Corey S. O'Hern[5,2,3,6,7]

[1]*Department of Mechanical Engineering, Yale University, New Haven, Connecticut 06520, USA*
[2]*Graduate Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut 06520, USA*
[3]*Integrated Graduate Program in Physical and Engineering Biology, Yale University, New Haven, Connecticut 06520, USA*
[4]*Benjamin Levich Institute and Physics Department, The City College of New York, New York, New York 10031, USA*
[5]*Department of Mechanical Engineering, Yale University, New Haven, Connecticut 06520, USA*
[6]*Department of Physics, Yale University, New Haven, Connecticut 06520, USA*
[7]*Department of Applied Physics, Yale University, New Haven, Connecticut 06520, USA*

(Received 21 July 2025; accepted 9 October 2025; published 6 November 2025)

Proteins are composed of chains of amino acids that fold into complex three-dimensional structures. Several key features, such as the radius of gyration, fraction of core amino acids $f_{\rm core}$, packing fraction $\langle\phi\rangle$ of core amino acids, and structure factor $S(q)$ define the structure of folded proteins. It is well-known that folded proteins are compact with a radius of gyration $R_g(N) \sim N^\nu$ that obeys power-law scaling with the number of amino acids N and $\nu \sim 1/3$, $f_{\rm core} \approx 0.09$, and $\langle\phi\rangle \approx 0.55$. We also investigate the *internal* scaling of the radius of gyration $R_g(n)$ versus the chemical separation n between amino acids for subchains of length n and show that it does not obey simple power-law scaling with $\nu \sim 1/3$. Instead, $R_g(n) \sim n^{\nu_{1,2}}$ with a larger exponent $\nu_1 > 1/3$ for small n and a smaller exponent $\nu_2 < 1/3$ for large n. To develop a minimal model for proteins that recapitulates these defining structural features, we carry out collapse simulations for a series of coarse-grained models with increasing complexity. We show that a model, which coarse-grains amino acids into a single spherical backbone bead and several variable-sized side-chain beads and enforces bend- and dihedral-angle constraints for the backbone, recapitulates $R_g(n)$, $f_{\rm core}$, $\langle\phi\rangle$, and $S(q)$ for more than 2500 x-ray crystal structures of proteins.

DOI: 10.1103/9wf9-ywhw

I. INTRODUCTION

Proteins are polypeptide chains containing tens to thousands of amino acids that carry out important cellular and extracellular functions. While breakthroughs in machine learning have improved our ability to predict the x-ray crystal structures of proteins from their amino acid sequences [1–3] and to design new protein sequences [4], modeling the physical and dynamic process of protein folding remains a challenge. In particular, experimental studies of protein folding have revealed intermediate kinetic traps, fold switching, mechanisms of misfolding and aggregation, allostery, and structural changes in response to mutations [5–10], all of which still require theoretical and computational modeling.

Globular proteins fold into complex three-dimensional conformations with compact interiors, or core regions, that determine their thermal stability [11]. Previous studies have shown that the fraction of amino acids in protein cores $f_{\rm core} \approx 0.09$, and the average packing fraction of core amino acids (without nonbonded atomic overlaps) $\langle\phi\rangle \approx 0.55$ [12–16]. The overall structure of folded proteins can be characterized by the structure factor $S(\vec{q}) = N^{-1}\sum_{k=1}^{N}\sum_{l=1}^{N} e^{i\vec{q}\cdot(\vec{r}_k-\vec{r}_l)}$ and radius of gyration of the protein backbone,

$$R_g(N) = \sqrt{\frac{1}{N}\sum_{k=1}^{N}|\vec{\mathbf{r}}_k - \vec{\mathbf{r}}_{\rm com}|^2}, \qquad (1)$$

where $\vec{q}$ is the wavevector, $\vec{\mathbf{r}}_k$ is the position of the N C_α atoms in the protein, and $\vec{\mathbf{r}}_{\rm com}$ is its center of mass [17]. Both $S(q)$ and $R_g(N)$ have been employed as reaction coordinates for the folding process [18] and used to identify intrinsically disordered proteins (IDPs), which do not adopt a single compact structure, but contain both open and compact regions [19].

The radius of gyration for simple polymers follows power-law scaling relations, $R_g(N) \propto N^\nu$, where $\nu = 1$ for fully extended polymers, 0.5 for random-walk polymers, and 1/3 for collapsed polymers. Recent studies of x-ray crystal structures of globular proteins have shown that $R_g(N) \sim N^{\nu^*}$ with exponent $\nu^* \sim 0.33$–0.4 [20], similar to the behavior for collapsed polymers [see the inset to Fig. 1(a)]. Deviations from the power-law scaling behavior with exponent ν^* are found for proteins with small ratios of hydrophobicity to electric charge [21–26]. However, proteins with similar N can possess strongly differing conformations.

To gain additional insight into the internal structure of proteins, we can define $\langle R_g(n)\rangle$ as the average radius of gyration over all subchains of length $n \leqslant N$,

$$\langle R_g(n)\rangle = \frac{1}{N-n+1}\sum_{i=1}^{N-n+1} R_g(i, i+n-1), \qquad (2)$$

where

$$R_g(i,j) = \left[\frac{1}{j-i+1}\sum_{k=i}^{j}(\vec{\mathbf{r}}_k - \langle\vec{\mathbf{r}}_k\rangle)^2\right]^{1/2} \qquad (3)$$

[*] These authors contributed equally to this work.

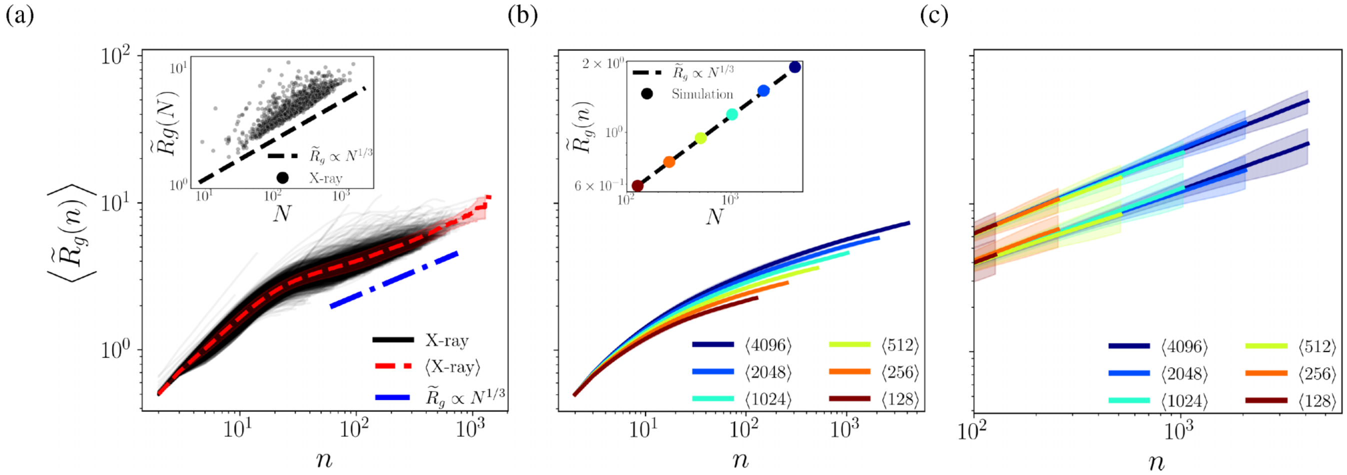

FIG. 1. Average normalized radius of gyration $\langle \widetilde{R}_g(n) \rangle$ plotted as a function of subchain length n. (a) The anomalous scaling of $\langle \widetilde{R}_g(n) \rangle$ for 2531 x-ray crystal structures of single-chain proteins with variable numbers of amino acids N (thin black lines). The dashed red line gives the average over all proteins. The dot-dashed blue line has a slope of 1/3. In the inset, we show $\langle \widetilde{R}_g(N) \rangle$ for the same x-ray crystal structures (filled black circles). The dashed black line has a slope of 1/3. (b) For collapsed, excluded-volume bead-spring polymers as for folded proteins, $\langle \widetilde{R}_g(n) \rangle$ does not obey power-law scaling behavior with a *single* exponent. However, in the inset, we show that the endpoints obey $\widetilde{R}_g(N) \propto N^{1/3}$ for $N = 128$ (black line) to 4096 (violet line) spherical monomers. (c) $\langle \widetilde{R}_g(n) \rangle \propto n^{\nu}$ with $\nu \sim 0.59$ for excluded-volume random-walk polymers (upper curves) compared to $\nu \sim 0.50$ for ideal random-walk polymers (lower curves).

and

$$\langle \vec{\mathbf{r}}_k \rangle = \frac{1}{j - i + 1} \sum_{k=i}^{j} \vec{\mathbf{r}}_k. \tag{4}$$

The advantage of focusing on $\langle R_g(n) \rangle$ is that it is possible to investigate the scaling of $\langle R_g(n) \rangle$ with sequence separation n for $n < N$. For excluded volume random walks, Flory scaling is self-similar with $\nu \sim 0.59$. However, $\langle R_g(n) \rangle$ is not self-similar for collapsed polymers and folded proteins. In Fig. 1(a), we show that while the $R_g(N)$ scaling for folded proteins obeys $R_g(N) \propto N^{\nu^*}$ with $\nu^* \sim 0.33$–0.4, the internal scaling $R_g(n)$ is more complex. $R_g(n)$ possesses two characteristic power-law scaling regions: $R_g(n) \propto n^{\nu_{1,2}}$ with $\nu_1 \sim 0.7 > 1/3$ for small n and $\nu_2 \sim 0.2 < 1/3$ for large n, which differs significantly from $R_g(n)$ for collapsed bead-spring polymers [Fig. 1(b)], as well as excluded-volume and ideal random-walk polymers [Fig. 1(c)]. Thus, because of its unique scaling properties, $\langle R_g(n) \rangle$ for $n < N$ can be used to validate coarse-grained models of proteins against x-ray crystal structures of folded proteins.

We seek to develop a minimal model for proteins that captures the key structural properties observed in high-resolution x-ray crystal structures of proteins. All-atom models have been used to fold proteins computationally [27–36], yet they have only folded proteins with $N \lesssim 100$ using physics-based force-fields within molecular dynamics (MD) simulations, and these simulations typically capture only μs to ms time scales [37–41]. More recently, slightly larger proteins with $N \approx 250$ have been accurately folded using a combination of machine learning-based structure prediction methods and MD simulations [42]. Folding proteins with $N > 10^2$ using MD simulation methods remains an important challenge, especially for proteins with little structural homology in the Protein Data Bank. Coarse-grained models can potentially be used to fold larger proteins by reducing the geometric complexity of the amino acids. Coarse-grained models for proteins range from one spherical bead per amino acid [43–46] to one spherical bead for the backbone and one or multiple beads for the side chains [47–52]. The "tube model" has also been used to coarse-grain proteins, where the polypeptide chain is represented as a tube with finite thickness. The tube model limits the protein conformational space to only compact folds that resemble realistic protein structures [53,54]. Prior coarse-grained models for proteins are typically calibrated by matching the radius of gyration $R_g(N)$ to within 10% of the x-ray crystal structure or achieving root-mean-square deviation (RMSD) of the C_α atoms RMSD $\lesssim 3$ Å from the x-ray crystal structure. However, matching only these two metrics to the x-ray crystal structures does not ensure that the model captures key features of protein cores of the x-ray crystal structure.

Thus, in this work, we investigate a range of coarse-grained models of proteins to determine the minimal model that recapitulates four key properties that generically define the structure of folded proteins: $\langle R_g(n) \rangle$, $\langle \phi \rangle$, $f_{\rm core}$, and $S(q)$. We focus on six coarse-grained protein models with increasing complexity: a collapsed excluded-volume bead-spring random-walk polymer model, the previous polymer model with effective bend- and dihedral-angle constraints, the previous polymer model with an additional side-chain spherical bead attached to each backbone spherical bead, the previous polymer model except the sizes of each side-chain spherical beads are selected to mimic the side chains of amino acids in the protein, the previous polymer model with the same side chain representations as those employed in Martini3 [50], and the previous polymer model with the single side-chain spherical beads of leucine and valine replaced with multiple side chain beads. To simplify the polymer models, we do not include explicit attractive interactions between amino acids. Instead, to induce hydrophobic collapse of the coarse-grained protein models, we employ an external compressive central force with damped MD simulations. Previous studies have shown that the structural properties of bead-spring polymers collapsed using attractive interactions are similar to those for purely repulsive bead-spring polymers compressed using a central force [55]. In addition, static packings of purely repulsive, rigid, amino acid-shaped particles compressed to jamming onset (i.e., the maximum packing fraction that does not give rise to overlaps between amino acids) achieve a similar average packing fraction as that found in the cores of x-ray crystal structures of globular proteins [13,56].

Below, we describe the results for simulations of chain collapse for all six coarse-grained models for more than 2500 individual proteins (with $N = 100$–1500). We show that models with sufficiently complex side-chain representations accurately reproduce $\langle R_g(n)\rangle$, $\langle \phi \rangle$, $f_{\rm core}$, and $S(q)$ over the full data set of proteins. In future studies, the accurate coarse-grained models described here can potentially be used for folding proteins of unknown structure, docking protein monomers to determine protein-protein interactions, and other protein structure prediction applications.

This article is organized into three additional sections and four Appendixes. In Sec. II, we describe the six coarse-grained protein models and the simulation protocol for studying protein chain collapse. In Sec. III, we describe the results for $R_g(n)$, $S(q)$, $\langle \phi \rangle$, and $f_{\rm core}$ for each coarse-grained protein model and compare the results to those for the x-ray crystal structures of proteins. In Sec. IV, we emphasize the importance of developing coarse-grained protein models that can accurately capture the structure of protein cores in x-ray crystal structures. We also outline future coarse-grained simulations that can recapitulate protein folding dynamics with small root-mean-square deviations from x-ray crystal structures of proteins. In Appendix A, we describe the constraints that we used to obtain the dataset of ∼2500 x-ray crystal structures of proteins from the Protein Data Bank. In Appendix B, we provide the procedure for generating the initial conformations for each coarse-grained protein model. In Appendix C, we describe the dihedral angle potential energy function for the coarse-grained protein models. Finally, in Appendix D, we illustrate our method to identify the core residues and calculate $f_{\rm core}$ and $\langle \phi \rangle$ in both the x-ray crystal structures and coarse-grained protein models.

II. METHODS

In Fig. 2, we illustrate six coarse-grained models of proteins [51,57–61]. Each model has a connected backbone including one spherical bead per amino acid backbone with the same average separation between successive C$_\alpha$ atoms in proteins, i.e., $\sigma_{bb} \approx 3.8$ Å. In order of increasing complexity, the models are: 1) a collapsed freely-jointed excluded-volume random-walk (CRW) polymer model, 2) the previous polymer model with constrained effective bend and dihedral angles (BADA) among the backbone spherical beads, 3) the previous polymer model with an additional spherical bead with a randomly chosen diameter that is freely-jointed to each backbone monomer to represent the side chain for each amino acid (FJSC), 4) an "in-sequence" freely-jointed side chain polymer model (In Seq), where the diameter of the side chain bead mimics the size of the side chain of the protein's amino acid

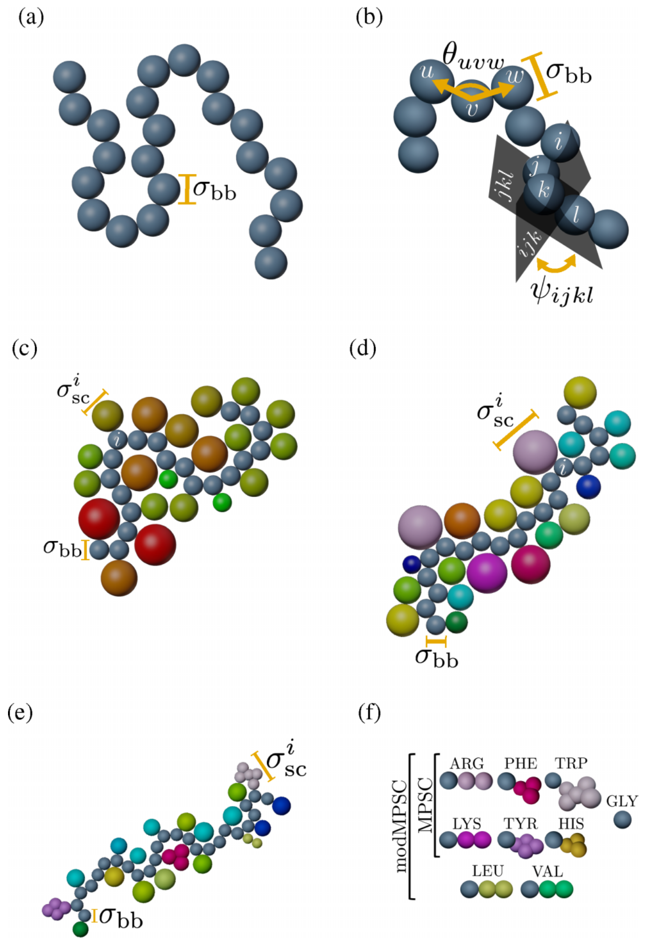

FIG. 2. (a)–(e) Snapshots of the six coarse-grained models of proteins, shown as 2D projections. When moving from (a)–(e), the successive models include all features of the previous models. $\sigma_{\rm bb}$ indicates the diameter of the spherical bead that represents the backbone of each amino acid. (a) A collapsed freely-jointed excluded-volume random walk (CRW) polymer chain with inter-amino acid separation $\sigma_{\rm bb}$. (b) For the bend- and dihedral-angle potential (BADA) polymer model, the effective bend angles θ_{uvw} between three consecutive amino acids are constrained to values determined by x-ray crystal structures of proteins by a harmonic potential $U_{\rm bend}$, and the effective dihedral angles ψ_{ijkl} between four consecutive amino acids are constrained to values determined by x-ray crystal structures of proteins by the dihedral angle potential $U_{\rm dh}$. (c) The freely jointed side-chain polymer model (FJSC) includes an additional spherical bead with diameter $\widetilde{\sigma}^i_{\rm sc}$ (colored by size) chosen randomly from a distribution of amino acid side chain diameters from x-ray crystal structures of proteins that are freely-jointed to each backbone monomer i. (d) For the "in-sequence" FJSC (In Seq) polymer model, the diameter of the side chain bead (colored by amino acid) is determined by the amino acid sequence that it is modeling. (e) The multi-particle side chain (MPSC) and modified MPSC (modMPSC) models use the geometry of the Martini3 side chains for seven types of amino acids. The modMPSC model differs from the MPSC model in using two spherical beads with a bend angle of 180° for the side chains of LEU and VAL. (f) A summary of the amino acid side chain representations for the MPSC and modMPSC models. All other amino acids in these models have a single bead representation for the side chain, as for the In Seq model. The examples in (d)–(e) are sections of the protein, PDBID: 3ZZO.

sequence, 5) a multi-particle side chain (MPSC) model similar to that for Martini3 [50], where six of the amino acids contain more than one spherical side-chain bead and glycine does not have a side-chain bead, and 6) a modified MPSC model (modMPSC), where leucine and valine have two spherical side-chain beads. For each model, we perform more than 2500 independent simulations, one for each protein in a dataset of high-resolution x-ray crystal structures of single-chain proteins [62]. See Appendix A for more information about how we constructed the dataset used in this study.

In Fig. 2(a), we illustrate the CRW polymer model, where each of the N spherical beads represents an amino acid with diameter $\sigma_{\rm bb}$. Neighboring amino acids i and $j = i + 1$ are connected using the harmonic bond length potential,

$$U_{\rm bond}(r_{ij}) = \frac{U_{\rm bb}}{2}\left(1 - \frac{r_{ij}}{\sigma_{ij}}\right)^2, \qquad (5)$$

where r_{ij} is the separation between amino acids i and j, $U_{\rm bb}$ is the strength of the bond length potential, and σ_{ij} is the sum of the radii of the bonded monomers i and j, $\sigma_{ij} = (\sigma_i + \sigma_j)/2$. Nonbonded amino acids interact via the purely repulsive linear spring potential,

$$U_{\rm rep}(r_{ij}) = \frac{\epsilon_{\rm rep}}{2}\left(1 - \frac{r_{ij}}{\sigma_{ij}}\right)^2 \Theta\left(1 - \frac{r_{ij}}{\sigma_{ij}}\right), \qquad (6)$$

where $\Theta(\cdot)$ is the Heaviside step function and $\epsilon_{\rm rep}$ is the strength of the nonbonded repulisve interactions between amino acids. Physical quantities will be made dimensionless using the energy scale $\epsilon_{\rm rep}$, the mass m of an amino acid backbone bead, and the lengthscale $\sigma_{\rm bb}$. Throughout this work a tilde over a given symbol is used to denote dimensionless quantities, e.g. $\widetilde{U}_{\rm bb} = U_{\rm bb}/\epsilon_{\rm rep}$. All dimensionless simulation parameters are defined in Appendixes B and C.

In Fig. 2(b), we show that the BADA polymer model also includes constraints on the bend and dihedral angles between amino acids. The bend angles θ_{ijk} between three sequential amino acids i, $j = i + 1$, and $k = i + 2$ are constrained by

$$U_{\rm bend}(\theta_{ijk}) = \frac{U_{\rm ba}}{2}\left(1 - \frac{\theta_{ijk}}{\theta^0_{ijk}}\right)^2, \qquad (7)$$

where the average bend angle θ^0_{ijk} is obtained from the x-ray crystal structure dataset. The dihedral-angle potential energy constrains the angle ψ_{ijkl} between planes formed by the three beads i, j, and k and three beads j, k, and l among the four consecutive backbone beads i, j, k, and l:

$$U_{\rm dh}(\psi_{ijkl}) = U_{\rm da}\sum_{\langle ijkl\rangle}\sum_{s=1}^{4}[A_s\cos(s\,\psi_{ijkl}) + B_s\sin(s\,\psi_{ijkl})], \qquad (8)$$

where $U_{\rm da}$ is the strength of the dihedral-angle potential, and the dimensionless coefficients A_s and B_s are determined by the x-ray crystal structure dataset. (See Appendix C.)

In Fig. 3(a), we show the distribution $\mathcal{P}(\theta_{ijk})$ of bend angles between each set of three successive C_α atoms from the x-ray crystal structure dataset. The distribution has a strong peak around $\theta_{ijk} \approx 90^\circ$ and a secondary peak near 120°. For each coarse-grained model that we simulate, we sample the bend angles randomly from $\mathcal{P}(\theta_{ijk})$, and then they are constrained using $U_{\rm bend}$ in Eq. (7). The dihedral-angle potential energy $U_{\rm dh}(\psi_{ijkl})$ [19,63] has a global minimum at $\psi_{ijkl} = \pm 180^\circ$, a peak near 60°, and a plateau extending over the range $0^\circ \leqslant \psi_{ijkl} \leqslant 120^\circ$. Calculating the Boltzmann weight for $U_{\rm dh}$ yields $\mathcal{P}(\psi_{ijkl})$ for the x-ray crystal structure dataset, which is shown in Fig. 3(b). The key features in $\mathcal{P}(\theta_{ijk})$ and $\mathcal{P}(\psi_{ijkl})$ are attributed to protein secondary structure. The peak around $\theta_{ijk} \approx 90^\circ$ in $\mathcal{P}(\theta_{ijk})$ and the plateau in $\mathcal{P}(\psi_{ijkl})$ originate from α-helical structures. The secondary peak near $\theta_{ijk} \approx 120^\circ$ and low-energy tails at $\psi_{ijkl} = \pm 180^\circ$ stem from β-sheet structures. Note that α-helices are not favored by the coarse-grained dihedral-angle potential energy $U_{\rm dh}$.

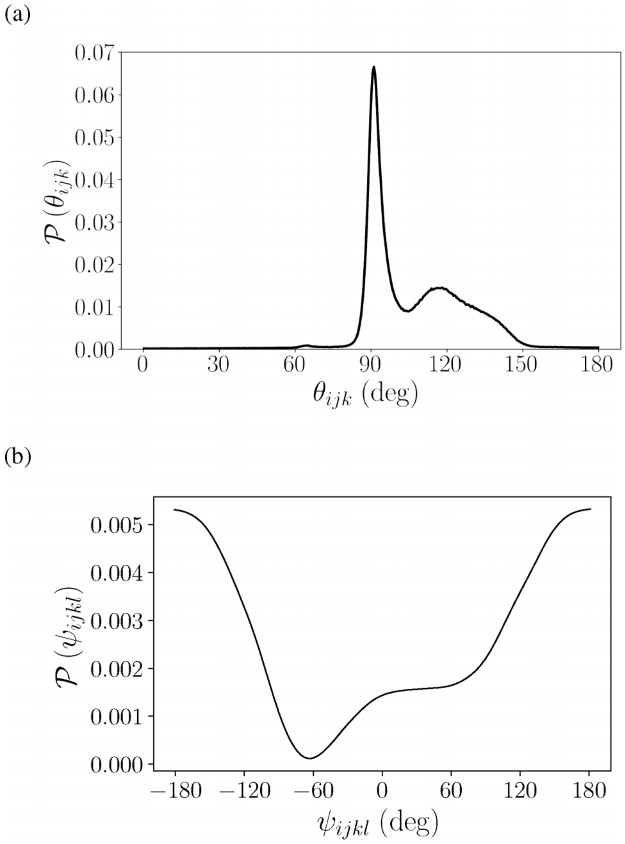

FIG. 3. (a) Distribution $\mathcal{P}(\theta_{ijk})$ of the effective bend angles between three consecutive C_α atoms from the dataset of x-ray crystal structures of proteins. (b) The distribution $\mathcal{P}(\psi_{ijkl})$ of effective dihedral angles ψ_{ijkl} between four consecutive C_α atoms observed in the x-ray crystal structure dataset when Boltzmann-weighting $U_{\rm dh}(\psi_{ijkl})$ [19].

The coarse-grained protein models FJSC, In Seq, MPSC, and modMPSC in Fig. 2(c)–2(f) incorporate side chain degrees of freedom, by freely-joining a spherical bead to each backbone bead [using Eq. (5)]. To approximate the effective diameter of each side chain, we calculate the maximum distance between all pairs of atoms in a side chain and add the average of the radii of the two atoms that are the farthest apart. The selected atomic radii have been used previously to calculate the average packing fraction of amino acids in protein cores [12,56,64] and are provided in Appendix D. Amino acid side chains can take on many conformations, so each amino

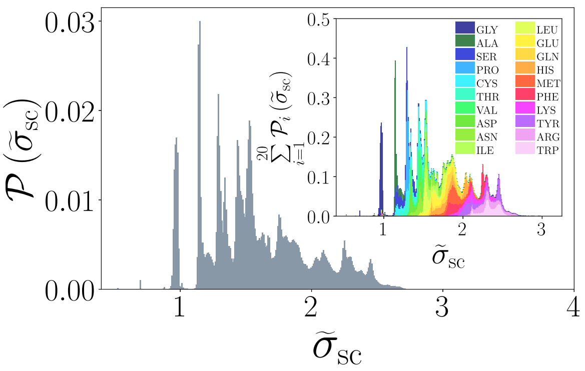

FIG. 4. Distribution $\mathcal{P}(\widetilde{\sigma}_{\rm sc})$ of the effective side chain diameters (normalized by $\sigma_{\rm bb}$) binned over all amino acid types. The inset shows the sum of the distributions $\mathcal{P}_i(\widetilde{\sigma}_{\rm sc})$ for each amino acid type i indicated by different colors.

acid possesses a distribution of effective side chain diameters. These distributions can either be calculated independently for each amino acid type or binned together to obtain an overall distribution of side chain diameters as shown in Fig. 4. For the FJSC polymer model in Fig. 2(c), the diameter $\sigma_{i,\rm sc}$ of the side chain bead bonded to backbone bead i is chosen randomly from the overall distribution of effective amino acid side chain diameters $\mathcal{P}(\sigma_{\rm sc})$ in the main panel of Fig. 4. In contrast, for the In Seq polymer model in Fig. 2(d), we select the diameter of each side chain bead according to the amino acid sequence of each protein in the x-ray crystal structure dataset. In particular, the diameters of the side chain beads are randomly sampled from the individual amino acid side chain diameter distributions $\mathcal{P}_i(\sigma_{\rm sc})$ illustrated in the inset of Fig. 4, where $\mathcal{P}(\widetilde{\sigma}_{\rm sc}) = A\sum_{i=1}^{20}\mathcal{P}_i(\widetilde{\sigma}_{\rm sc})/A_i$, A is the normalization constant determined by $\int\mathcal{P}(\widetilde{\sigma}_{\rm sc})d\widetilde{\sigma}_{\rm sc} = 1$, $A_i = 1/(\Delta\widetilde{\sigma}_{\rm sc}N_c^i)$ is the normalization constant for the diameter distribution of amino acid i with N_c^i total counts and bin width $\Delta\widetilde{\sigma}_{\rm sc}$.

In Fig. 2(e), we show the MPSC model, which includes a single backbone spherical bead and side chains made up of 0-5 spherical beads. The geometrical representations of the side chains are similar to those used in Martini3. Glycine is now only represented by a backbone spherical bead. Each amino acid with a single side-chain sphere is unchanged from the In Seq model. Six amino acids in the MPSC model are represented by multiple side-chain spherical beads: arginine, phenylalanine, tryptophan, lysine, tyrosine, and histidine. The maximum dimension of the side chains with multiple spherical beads is the same as the diameter of the single side-chain bead representation in the In Seq model. To achieve this, the multiple side-chain spherical beads are the same size and rescaled so that the sum of the diameters matches the single side-chain bead diameter in the In Seq model. The modMPSC model is similar to the MPSC model, except the side chains for leucine and valine are represented by two spherical beads that form a 180° bend angle with the backbone bead. In Fig. 2(f), we summarize the side-chain representations for the MPSC and modMPSC models.

When generating the initial coarse-grained protein conformations, the total potential energy contributions, $U_{\rm rep}^{\rm tot} = \sum_{\langle i,j\rangle}U_{\rm rep}(r_{ij}) \approx 0$ and $U_{\rm bond}^{\rm tot} = \sum_{\langle i,j\rangle}U_{\rm bond}(r_{ij}) \approx 0$ for all models, and $U_{\rm bend}^{\rm tot} = \sum_{\langle i,j,k\rangle}U_{\rm bend}(\theta_{ijk}) \approx 0$ for the BADA, FJSC, and In Seq models. We employ damped molecular dynamics (MD) simulations with an additional attractive central force on each bead to generate a collapsed conformation for each model and target protein. We employ a dimensionless damping parameter $\widetilde{\gamma} = 0.1$ in the overdamped limit, and run the damped MD simulations until the maximum magnitude of the net force on any bead i satisfies $\max_i\widetilde{F}_i < \widetilde{F}_{\rm tol}$, where $F_i = |\vec{F}_i| = |\vec{\nabla}_{\vec{r}_i}U|$, U is the total potential energy for a given model, and $\widetilde{F}_{\rm tol} = 5\times 10^{-13}$. We include an extra factor of the ratio of the bead diameter σ_i to the maximum bead diameter $\sigma_{\max}$ raised to a power in the expression for the central force to ensure that the coarse-grained models do not form clusters of similar-sized beads during collapse when the beads are polydisperse [55]:

$$\vec{\mathbf{F}}_{\rm cent} = -F_{\rm cent}\left(\frac{\sigma_i}{\sigma_{\max}}\right)^{9/4}\hat{\mathbf{r}}_i. \tag{9}$$

The strength of the central force $\widetilde{F}_{\rm cent} = 10^{-4}$ compared to the constraint forces is such that the stereochemical constraints remain satisfied during collapse, e.g., the bend and dihedral angle distributions $\mathcal{P}(\theta_{ijk})$ and $\mathcal{P}(\psi_{ijkl})$ are nearly identical in the collapsed and initial states, and the results do not depend on $\widetilde{F}_{\rm cent}$.

For each coarse-grained model, we generate one conformation per protein, matching the chain length of its x-ray crystal structure. All measured quantities are averaged over the ~2500 proteins in the Dunbrack database. We verified that ensemble-averaging over ~100 random initial conformations per protein does not significantly change the results. Although the individual conformations differ, $\langle R_g(n)\rangle$ averaged over initial conditions give standard errors that are at least one order of magnitude smaller than those from averaging over the ~2500 x-ray crystal structures. Thus, our results for $\langle R_g(n)\rangle$ will not change significantly if we average over initial conformations in addition to averaging over the ~2500 proteins. We then calculate $\langle\phi\rangle$, $f_{\rm core}$, and $S(q)$, in addition to $\langle R_g(n)\rangle$, in the collapsed conformations for each coarse-grained protein model and protein target.

III. RESULTS

The results for the normalized radius of gyration $\langle\widetilde{R}_g(n)\rangle$ as a function of subchain length n for the six coarse-grained protein models and the dataset of x-ray crystal structures are shown in Fig. 5(a). To quantify differences in the radius of gyration between each model and the x-ray crystal structure dataset, we compute the normalized mean-squared error (MSE) in $R_g(n)$:

$$\mathrm{MSE}(\langle\widetilde{\mathrm{R}}_{\rm g}\rangle) = \frac{\sum_{n=2}^{N}(\Delta\langle\widetilde{R}_g(n)\rangle)^2}{\sum_{n=2}^{N}\left(\langle\widetilde{R}_g^{\text{x-ray}}(n)\rangle\right)^2}, \tag{10}$$

where $\Delta\langle\widetilde{R}_g(n)\rangle = \langle\widetilde{R}_g^{\rm model}(n)\rangle - \langle\widetilde{R}_g^{\text{x-ray}}(n)\rangle$.

As shown in Fig. 5(a), the simplest coarse-grained model (CRW) does not recapitulate $\langle R_g(n)\rangle$ for folded proteins. $\langle R_g(n)\rangle$ for the CRW model is highly curved on a log - log plot (i.e., does not possess a kink) at small n and is a factor of ~1.5 smaller than $\langle R_g(n)\rangle$ for the x-ray crystal structure data at large n. The CRW model has the largest normalized mean-squared

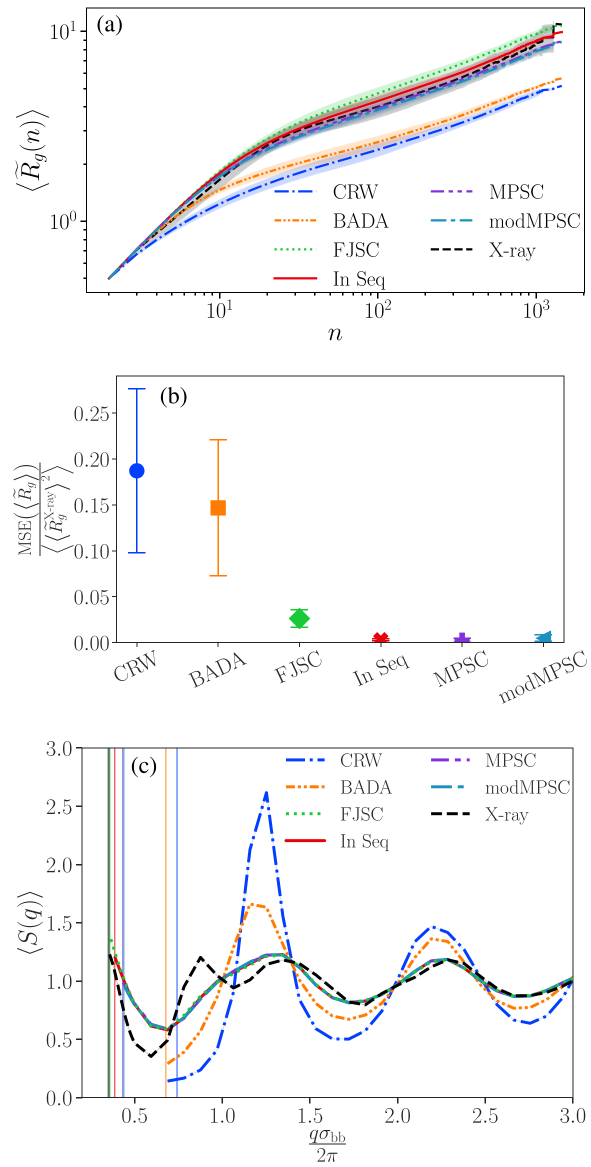

FIG. 5. (a) The average radius of gyration $\langle \widetilde{R}_g(n) \rangle$ plotted versus subchain length n for the x-ray crystal structures (black dashed line) and coarse-grained protein models with corresponding colors and line styles in the legend. The shading indicates the standard deviation about $\langle \widetilde{R}_g \rangle$ for each dataset. (b) Normalized mean-squared error in Eq. (10) between $\langle \widetilde{R}_g(n) \rangle$ for each model and the average over the x-ray crystal structures. (c) The average structure factor $\langle S(q) \rangle$ plotted versus the wavenumber q scaled by the diameter of the coarse-grained backbone size σ_{bb}. The vertical lines indicate the wavenumbers $q = 2\pi/\max\{\langle \widetilde{R}_g(N) \rangle\}$ for each polymer model.

error relative to the x-ray crystal structure data of the six models we considered, as shown in Fig. 5(b). Similarly, $S(q)$ for the CRW model is not similar to that for the x-ray crystal structures as shown in Fig. 5(c).

Introducing effective bend- and dihedral-angle potentials leads to a small, but important change in $\langle R_g(n) \rangle$ for the BADA polymer model, i.e., the appearance of a kink near $n^* \sim 10$ that separates the small- and large-n regions. $\langle R_g(n) \rangle \sim n^{\nu_{1,2}}$, where $\nu_1 \sim 0.7$ for $n \lesssim n^*$ and $\nu_2 \sim 0.2$ for $n \gtrsim n^*$, which is similar to the results for the x-ray crystal structure data. However, n^* for the BADA polymer model is smaller than that for the x-ray crystal structure data, and the normalized MSE in $R(n)$ for the BADA model is still quite large ($\sim$0.15). The large MSE is caused by the fact that the persistence length of subchains in the BADA model is shorter than that for the x-ray crystal structures, and the effective bend- and dihedral-angle constraints are not sufficient to keep the subchains from over-collapsing at small n.

When the amino acids are coarse-grained to include *single* spherical beads for both the backbone and side chain degrees of freedom (i.e., the FJSC and In Seq models), the backbone can no longer collapse as densely as found for the CRW and BADA models. For the FJSC and In Seq models, the kink location increases to $n^* \sim 30$ and their $\langle R_g(n) \rangle$ are similar to that for the x-ray crystal structure data [Fig. 5(a)]. The normalized MSE is $\lesssim$0.02 for both the FJSC and In Seq polymer models [Fig. 5(b)]. In addition, $S(q)$ for the FJSC and In Seq models match the x-ray crystal structures much better than $S(q)$ for the BADA and CRW models [Fig. 5(c)]. Thus, coarse-grained protein models require at least a single side-chain bead with backbone bend- and dihedral-angle restraints to recapitulate $\langle R_g(n) \rangle$ and $S(q)$ of folded proteins. However, can the FJSC and In Seq models capture the core packing properties of proteins, such as $\langle \phi \rangle$ and f_{core}?

Protein cores are dense packings of amino acids in the solvent-inaccessible interior of proteins, whose size and structure have been directly correlated with the stability of the protein [11,65]. Previous studies have shown that the average core packing fraction in x-ray crystal structures of proteins is $\langle \phi \rangle \approx 0.55$ [15,16,56,64]. To identify core amino acids, we implement the software `FreeSASA` [66] to compute the relative solvent accessible surface area (rSASA) using the Lee-Richards algorithm [67]. This method employs a probe sphere to represent a solvent molecule of diameter $\widetilde{\sigma}_{\mathrm{probe}}$ that rolls over the folded protein to determine how much surface area of each amino acid it can make contact with relative to the total surface area of the fully solvated amino acid. In this work, we consider an amino acid to be in the core if rSASA $\leqslant 10^{-3}$, which has been previously used as an effective rSASA cutoff for identifying core amino acids [12,14–16,55,64,68]. Smaller diameter probes can access amino acids that are buried deeper in the protein because they can fit through smaller void spaces. Thus, we expect that as the probe shrinks, the number of amino acids found in the core will decrease and when $\widetilde{\sigma}_{\mathrm{probe}} \to 0$ the entire protein will be labeled as "surface," with $\langle f_{\mathrm{core}} \rangle = 0$. Because proteins typically reside in water, we used a probe sphere with a diameter given by the size of a water molecule, $\sigma_{\mathrm{H_2O}} \approx 0.73\sigma_{\mathrm{bb}}$. Core amino acids in x-ray crystal structures are often not all nearest neighbors and instead occur in separate clusters. Motivated by this, we calculate the average *local* packing fraction $\langle \phi \rangle$ for each coarse-grained model conformation or x-ray crystal structure. To calculate $\langle \phi \rangle$, we perform a Voronoi tessellation and find the ratio of the volume V_i of amino acid i to the local Voronoi cell volume V_i^{voro} of amino acid i, averaged over all N_{core} core amino acids:

$$\langle \phi \rangle = \frac{1}{N_{\mathrm{core}}} \sum_{i=1}^{N_{\mathrm{core}}} \frac{V_i}{V_i^{\mathrm{voro}}}. \tag{11}$$

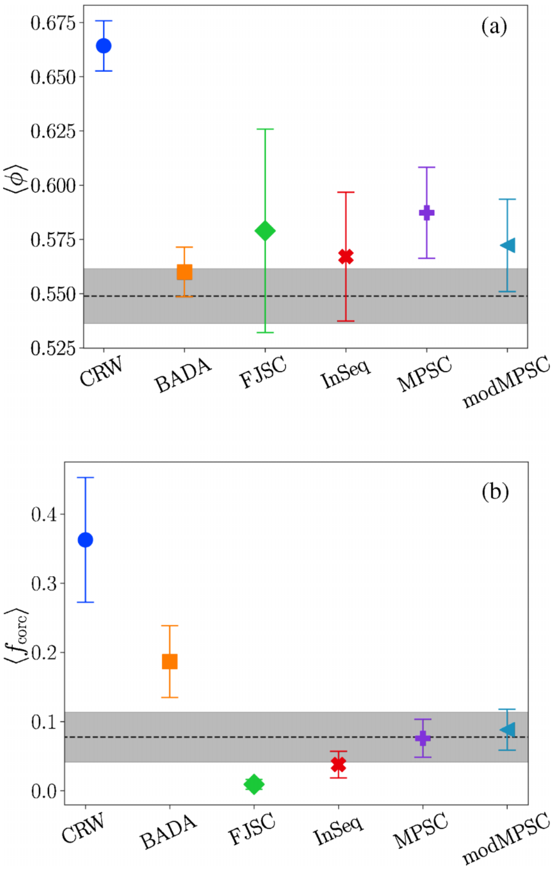

FIG. 6. For each coarse-grained protein model indicated by the colors and line types in the legend, we compare (a) the average local packing fraction $\langle\phi\rangle$, and (b) the average fraction of amino acids in the core $\langle f_{\mathrm{core}}\rangle$. In panels (a) and (b), the horizontal dashed black line marks the average values for the x-ray crystal structures with ±1 standard deviation shaded in gray. The error bars for the data points in (a) and (b) represent the standard deviation of the distributions for each model.

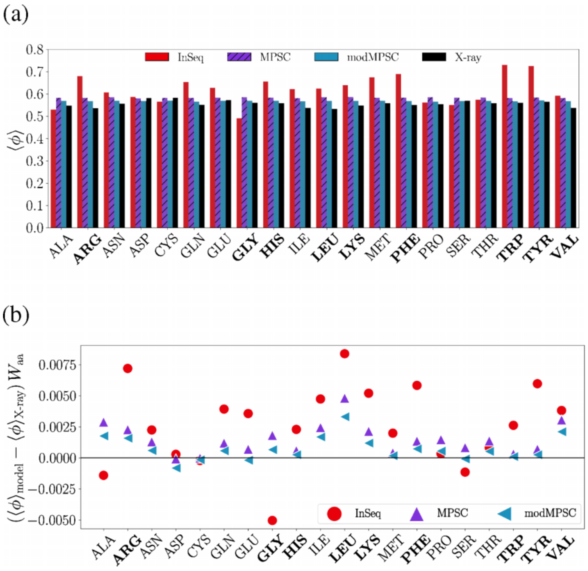

FIG. 7. (a) The average packing fraction for each amino acid using the InSeq, MPSC, and modMPSC models. (b) The difference in average packing fraction between the models and the x-ray crystal structures for each protein, weighted by the abundance of each amino acid type in the dataset. As the coarse-grained models become more detailed, the packing fraction approaches the values for the x-ray crystal structures. The amino acids with more than one side chain atom in the MPSC and modMPSC models are labeled in bold.

The fraction of core amino acids is given by $f_{\mathrm{core}} = N_{\mathrm{core}}/N$. $\langle\phi\rangle$ and $\langle f_{\mathrm{core}}\rangle$ for the x-ray crystal structures are found using Voronoi tessellation, as described above, but with atomic radii used in previous studies [12,56,64]. Appendix D includes additional details concerning calculations of the local packing fraction, fraction of core amino acids, and rSASA.

We find that the core packing properties of the FJSC and In Seq models do not strongly agree with those for x-ray structures of proteins. $\langle\phi\rangle \approx 0.57$–$0.58$ for the FJSC and In Seq models, which is similar to $\langle\phi\rangle \approx 0.55$ for the x-ray crystal structures [Fig. 6(a)]. However, the FJSC and In Seq models are not able to match the average packing fraction for each amino acid individually in x-ray crystal structures as shown in Figs. 7(a) and 7(b). Further, $\langle f_{\mathrm{core}}\rangle \approx 0.02$ and 0.05 for the FJSC and In Seq models, respectively, which indicates that the cores for the FJSC and In Seq models are significantly smaller than the cores with $\langle f_{\mathrm{core}}\rangle \approx 0.09$ for x-ray crystal structures of proteins [Fig. 6(b)]. Thus, we also included simulations for the MPSC and modMSPC models to show that adding multiple side chain beads can improve the coarse-grained description of the core packing properties.

We next investigate whether the slightly higher packing fraction for the MPSC and modMPSC models is the result of poorly modeling specific amino acids. In Fig. 7, we show the average packing fraction for each amino acid type for the In Seq, MPSC, and modMPSC models compared to the x-ray crystal structures. From Fig. 7(a), we observe that as the side chain representations become more complex, the average packing fraction for each amino acid begins to converge to the values for the x-ray crystal structures. The packing fractions for arginine, glycine, histidine, leucine, lysine, phenylalanine, tryptophan, tyrosine, and valine are all notably higher in the InSeq model than the x-ray crystal structure data. Relative to InSeq, the side-chain geometries in the MPSC and modMPSC models yield better agreement with the amino acid packing fractions found in the x-ray crystal structures. The reason for the remaining packing fraction error for the MPSC and modMPSC models can be seen in Fig. 7(b), which shows the difference between the average packing fraction by residue for the models and the x-ray crystal structures weighted by the relative abundance of each amino acid in the dataset. The amino acids for the In Seq model have average packing fractions that are both greater than and less than the values for the x-ray crystal structures, while the models with more detailed side chain representations have average packing fractions that are closer to (but slightly larger than) the x-ray crystal structures. Thus, the average packing fractions for each amino acid in the modMPSC model are greater than those for the In Seq model, even though the average values for the individual amino acids for the modMPSC model are converging to the values for the x-ray crystal structures.

In summary, we find that, as expected, $\langle R_g(n)\rangle$ and $S(q)$ for the MPSC and modMPSC models are both similar to those for the x-ray crystal structures [see Figs. 5(a) and 5(c)] $\langle f_{\rm core}\rangle$ for the MPSC model improved relative to that for the In Seq model and falls within the range for the x-ray crystal structures ($\langle f_{\rm core}\rangle \approx 0.09$). However, the core packing fraction $\langle\phi\rangle$ for the MPSC model did not move toward the value for the x-ray crystal structures ($\langle\phi\rangle \approx 0.59$). However, $\langle\phi\rangle \approx 0.57$ improves for the modMPSC model (and the average packing fractions for each amino acid individually approach those for the x-ray crystal structures), while $\langle f_{\rm core}\rangle \approx 0.09$ remains essentially unchanged and identical to the x-ray crystal structure value. These results emphasize that the side-chain representation in the modMPSC model can recapitulate the overall structural and core packing properties in x-ray crystal structures of proteins.

IV. CONCLUSIONS AND OUTLOOK

Using a series of coarse-grained protein models with increasing complexity, we identified the minimal coarse-grained models that can recapitulate several key structural properties that define folded proteins, obtained from a large dataset of more than 2500 high-resolution x-ray crystal structures of single-chain proteins. We show that coarse-grained models with only a single backbone spherical bead cannot capture the structural properties of folded proteins. Coarse-grained protein models with a single side-chain bead (plus a single backbone bead) can recapitulate $\langle R_g(n)\rangle$ and $S(q)$, but are not able to accurately describe the core packing properties of folded proteins. Using a new coarse-grained model (modMPSC) with multiple side-chain beads, we obtain $\langle\phi\rangle$ and $f_{\rm core}$ (as well as $\langle R_g(n)\rangle$ and $S(q)$) to <4% of the values for the x-ray crystal structures of proteins.

An important goal of this work was to identify the minimal coarse-grained protein model that can capture important generic structural properties of folded proteins, including the scaling of the subchain radius of gyration $R_g(n)$, the structure factor $S(q)$, and the core packing fraction $\phi \approx 0.55$ and fraction of core amino acids $f_{\rm core} \approx 0.09$. Our results show that a purely repulsive bead–spring backbone, plus stereochemical constraints, and a minimal side–chain representation can be collapsed into compact structures whose ensemble–averaged properties are statistically similar to those of x-ray crystal structures of proteins.

We now compare the root-mean-square deviations (RMSD) of the C_α positions between the modMPSC models for each protein and the corresponding x-ray crystal structures. In the limit of both weak radial and damping forces, we have shown that the model proteins can find densely packed conformations that are similar to the corresponding x-ray crystal structures. To estimate the RMSD in this limit, we placed the modMPSC model beads in positions that approximate the atomic positions in the corresponding x-ray crystal structures and then minimized the total energy of the protein in the presence of the radial force. We find that the resulting average C_α RMSD of the core amino acids in the modMPSC models is ~3.5 Å.

In previous studies, we have shown that it is possible to achieve $\lesssim$1 Å core RMSD for all-atom models with bend angle, backbone and side chain dihedral angle restraints using radial compressive and damping forces [16]. In future studies, we will develop *coarse-grained* protein models that can also achieve core RMSD $\lesssim$1 Å. One method to lower the core RMSD is to add improper dihedral angle restraints to the side chains in the modMPSC model. Without dihedral angle restraints, the coarse-grained side chains can potentially take on nonphysical conformations. In the current modMPSC model, we increased the geometric complexity of the side chain representations for only leucine and valine. Our findings also highlight the importance of "directionality" or amino acid anisotropy in accurately modeling proteins [69–71]. Without explicitly modeling anisotropic side chains in the MPSC and modMPSC models, the predicted protein packing fractions became less accurate, particularly when compared to the average packing fraction of each amino acid observed in x-ray crystal structures [Fig. 7(a)]. In future studies, we can increase the number of spherical beads to represent the side chains (with corresponding dihedral angle restraints) of the other amino acids. We will also determine the optimal ratio of the radial compressive and damping forces that yields core C_α RMSD $\lesssim$1 Å, as well as develop short-range attractive interactions for hydrophobic amino acids in the modMPSC models that can achieve protein-specific folded states upon decreases in temperature.

Thus, we can potentially use the modMPSC model to fold all protein sequences in the human proteome. These studies would complement existing *de novo* protein structure prediction methods, such as AlphaFold3 [1]. For amino acid sequences in the human proteome without experimentally determined structures, we can compare and contrast the results for folding simulations of the modMPSC model to those for AlphaFold.

ACKNOWLEDGMENTS

The authors acknowledge support from NIH Training Grant No. T32GM145452 (A.T.G., J.S., and C.S.O.), NIH Training Grant No. T15LM007056-37 (J.A.L.), and the High Performance Computing facilities operated by Yale's Center for Research Computing.

DATA AVAILABILITY

The data that support the findings of this article are openly available [72].

APPENDIX A: X-RAY CRYSTAL STRUCTURE DATASET

In this work, we compare the results from the coarse-grained protein models to a subset of the Dunbrack 1.8 PISCES Protein Database of high-resolution x-ray crystal structures [62,73]. The complete dataset consists of more than 5000 proteins ranging in length from fewer than 100 residues to greater than 8000 residues, with less than 50% sequence similarity between structures and resolution $\leqslant$1.8 Å. Hydrogen atoms have been added to each protein x-ray crystal structure using the Reduce software [74].

We cull the Dunbrack 1.8 dataset based on two criteria. First, we exclude any proteins that have unknown

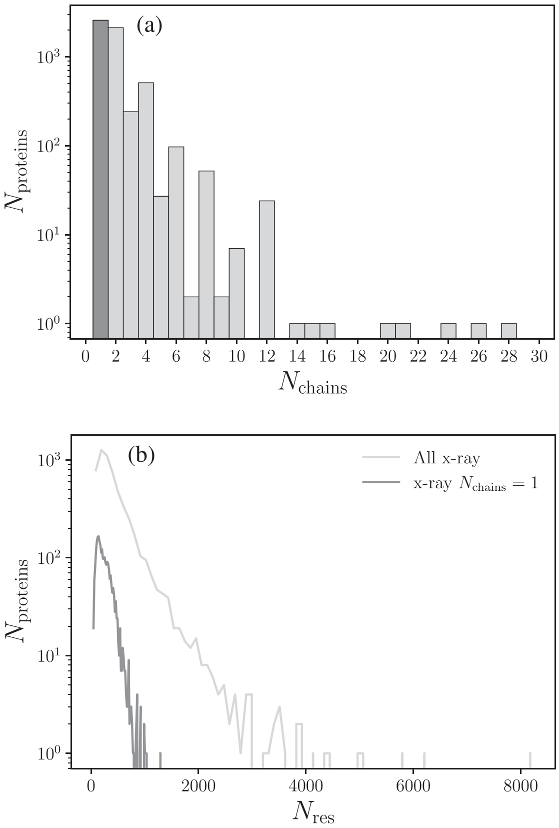

FIG. 8. (a) The frequency distribution $N_{\rm proteins}$ of the number of chains $N_{\rm chains}$ in each x-ray crystal structure. The dark gray bar for $N_{\rm chains} = 1$ represents about half of the entries in the x-ray crystal structure dataset. (b) Frequency distributions $N_{\rm proteins}$ of the number of residues $N_{\rm res}$ in each x-ray crystal structure in the full dataset (black line) and the subset with $N_{\rm chains} = 1$ (dark gray line).

residues or noncanonical amino acids, such as selenocysteine (Sec) [75–77]. For the second criterion, we only keep proteins that consist of one chain (i.e., monomeric proteins). As shown in Fig. 8(a), single-chain proteins make up nearly half of the full Dunbrack 1.8 dataset (with 2531 single-chain proteins). Figure 8(b) shows the frequency distribution of the number of residues $N_{\rm res}$ in each x-ray crystal structure for the full dataset in black and for the single-chain proteins in pink.

Among the excluded proteins are those that are composed of only a single secondary structure, e.g., only α-helices or β-sheets. The proteins included in the analyses in the main text have an average $\langle R_g(n)\rangle$ that is similar to that for most entries in the full dataset. Thus, the subset of proteins is an accurate representation of the full dataset.

APPENDIX B: GENERATING INITIAL CONFORMATIONS

For each x-ray crystal structure and coarse-grained protein model, the initial conformation is constructed one residue at a time. We generate an initial conformation that minimizes the bond, bend-angle, and dihedral-angle potential energies, as well as the nonbonded potential energy. We begin with two backbone atoms in contact and, if the model includes explicit side chains, randomly choose a position around the respective backbone atom to attach a sidechain bead with a diameter sampled from the appropriate distribution in Fig. 4 in the main text. For the FJSC model, the diameter is chosen from a normalized distribution of all sidechain sizes regardless of the amino acid, and for the In Seq model, the diameter is chosen from the size distribution for the correct amino acid type. The sidechain bead sizes are obtained from the x-ray crystal structures in the full Dunbrack 1.8 dataset. The sidechain bead diameter is defined as the largest center-to-center distance between any two atoms plus the average of their atomic radii. We use the atomic radii from previous studies of core packing in folded proteins [12,56,64], and are listed in Table I.

TABLE I. Atomic radii used to describe the all-atom x-ray crystal structures. HX denotes a hydrogen atom bound to a carbon atom; H denotes a hydrogen atom bound to any other atom.

Atom type	Radius (Å)
C	1.50
CO	1.30
N	1.30
O	1.40
H	1.00
S	1.75
HX	1.10

We check for overlaps each time a new bead is added to the model. Hence, we do not bias the sampling of sidechain diameters to smaller values, we allow for small overlaps between sidechain beads when initially building the model conformation. Starting with the third backbone bead, a new backbone position is chosen randomly in the subspace that minimizes the bond length and bend angle potential energies, without overlapping another backbone bead. Depending on the model, subsequent backbone beads are also selected so that the dihedral angles sample $\mathcal{P}(\psi_{ijkl}) \propto e^{-\tilde{U}_{\rm dh}(\psi_{ijkl})}$. After all beads have been added for a given coarse-grained model and x-ray crystal structure, damped molecular dynamics simulations are carried out to remove bead overlaps.

The MPSC model is initialized, starting with conformations from the In Seq model. The side chain representations for the In Seq model are modified to have 0-5 spherical beads as used for the Martini3 model [50]. Compared to the In Seq model, the MPSC model changes the side chain representations for glycine, histidine, arginine, lysine, phenylalanine, tryptophan, and tyrosine. The spherical beads for the MPSC side chain representations of each amino acid are the same size, and the spherical beads are scaled so that the sum of their diameters matches the diameter of the single side chain bead for the In Seq model. The modMPSC model is initialized in the same manner as the MPSC model. However, the modMPSC model modifies the side chain representations for valine and leucine to include two spherical beads instead of one. After changing the side chains for the MPSC and modMPSC models, we run a short NVE simulation for each of the coarse-grained proteins to ensure a novel conformation before the production simulation.

TABLE II. Coefficients of the backbone dihedral angle potential energy $\boldsymbol{U}_{\mathbf{dh}}$.

s	1	2	3	4
A_s (5×10^{-7})	70.5	−31.3	−7.9	4.1
B_s (5×10^{-7})	−17.5	−9.3	3.0	3.0

APPENDIX C: DIHEDRAL ANGLE POTENTIAL ENERGY

The dihedral angle potential energy in Fig. 3(b) in the main text was constructed from the probability distribution $\mathcal{P}(\psi_{ijkl})$ of dihedral angles between four consecutive C_α atoms using Langevin dynamics of the united atom (UA) model for proteins in previous work [19]. The dihedral angle potential energy is obtained from $U_{\rm dh} \propto -k_bT\langle \ln \mathcal{P}(\psi_{ijkl})\rangle$ and fit with a fourth-order Fourier series, with the coefficients listed in Table II. The peak near $\psi_{ijkl} = -60^\circ$ and the plateau in the range $0^\circ < \psi_{ijkl} < 120^\circ$ can be attributed to secondary structure in proteins.

$$U_{\rm dh}(\psi_{ijkl}) = U_{\rm da}\sum_{\langle ijkl\rangle}\sum_{s=1}^{4}[A_s\cos(s\,\psi_{ijkl}) + B_s\sin(s\,\psi_{ijkl})]. \tag{C1}$$

All physical quantities, including the dihedral angle potential energy, are made dimensionless using the energy, mass, and length units: $\epsilon_{\rm rep}$, m, and $\sigma_{\rm bb}$. The values used in the simulations are listed in Table III.

APPENDIX D: CALCULATING CORE PROPERTIES

1. Fraction core, $f_{\rm core}$

The first step in calculating the fraction of core amino acids ($f_{\rm core}$) is to identify which residues are in the core. In this work, we define a residue to be in the protein core if the relative solvent accessible surface area (rSASA) rSASA $\leqslant 10^{-3}$. A probe particle of diameter $\sigma_{\rm probe}$ ($\sigma_{\rm probe} = 0.73\sigma_{\rm bb}$ in this work) represents the solvent and moves on the surface and through the geometrically accessible void space of the protein. rSASA is the ratio of the surface area of the residue that the probe can make contact with to the surface area of the full residue, removed from the protein and fully solvated. An example of the all-atom structure for 5TKW is shown in Fig. 9(a), with its identified core shown in Fig 9(b). Once the core residues are identified, $f_{\rm core}$ is calculated by dividing the number of core amino acids by the total number of amino acids in the protein.

TABLE III. Dimensionless simulation parameters.

Parameter	Value
$\widetilde{U}_{\rm bb}$	1.0
$\widetilde{U}_{\rm ba}$	1.0
$\widetilde{U}_{\rm da}$	1.0
$\widetilde{F}_{\rm cent}$	10^{-4}

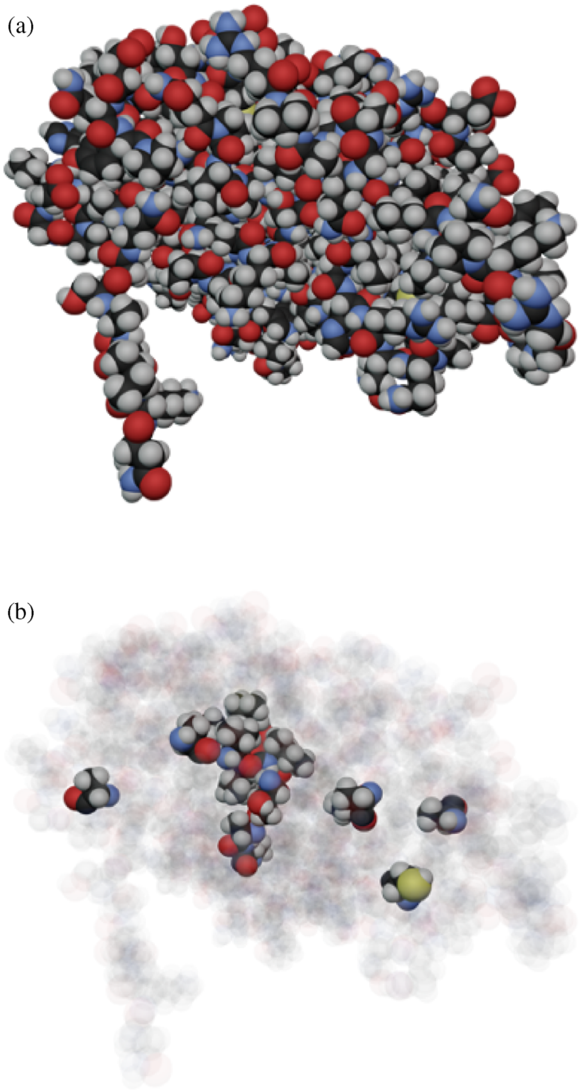

FIG. 9. (a) All-atom representation of PDBID: 5TKW. (b) The core of protein 5TKW is highlighted by making all noncore atoms transparent. This protein has a core that is not contiguous, which is common for folded proteins.

2. Packing fraction, ϕ

Calculating the packing fraction (ϕ) starts with identifying the core residues using the same method as that used for $f_{\rm core}$. In addition to rSASA, we check for surface residues using a radical Voronoi tessellation. In a Voronoi tessellation, if a point lies on the collapsed polymer surface, the volume of its cell will depend on the bounding box. Any Voronoi cell volume that remains constant after scaling the boundary is labeled as being in the interior of the polymer. The complement of the union of surface beads from the rSASA and Voronoi methods defines the core residues (and their Voronoi cells) used for calculating the packing fraction. Any residues identified in the core are then used to compute the packing fraction of the core.

For all collapsed models and x-ray crystal structures, we treat each sphere individually, whether it represents a side chain atom or backbone atom. If a pair of atoms has negligible overlap, each Voronoi cell will contain the entire sphere. If there are overlaps between two spheres, the Voronoi plane will intersect the spheres. The collapsed models have negligible overlaps, but the all-atom x-ray crystal structures possess both inter-residue and intra-residue atomic overlaps. There are several ways to compute the net volume of spheres in a residue while accounting for overlaps, such as the inclusion-exclusion principle, which requires adding and subtracting the overlap volumes of pairs, triples, etc.. We chose to use a Monte Carlo method for calculating volumes of amino acids. We place the core residues in a box and randomly sample points in the box. In the case of a point that falls in the overlap region between a core atom and a surface atom, we check on which side of the Voronoi plane the point falls. If the point is on the core side of the Voronoi plane, it is counted as in the core; otherwise, it is outside of the core. Given N points sampled in the bounding box of volume $V_{\rm box}$, if $N_{\rm in}$ are found to fall in the amino acid on the core side of any atoms near the core-surface boundary, the volume of the amino acid is

$$V_{\rm res} \approx \frac{N_{\rm in}}{N} V_{\rm box}.$$

After calculating $V_{\rm res}$ for all core residues in the protein, we can determine the packing fraction ϕ by averaging the local packing fractions of each core residue,

$$\phi_{\rm res} \approx \frac{V_{\rm res}}{V_{\rm res}^{\rm voro}},$$

and $\phi = \langle \phi_{\rm res} \rangle$. As the number of test points grows, the volume computation can be made as accurate as desired, and the packing fraction converges. In addition to using a high density of sampled points ($20\,000$ points/$\sigma_{\rm bb}^3$), we also compute the final local packing fraction for each residue by averaging over 50 independent realizations.

[1] J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie *et al.*, Accurate structure prediction of biomolecular interactions with AlphaFold 3, Nature (London) **630**, 493 (2024).

[2] M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A. C. Ebrecht, D. J. Opperman *et al.*, Accurate prediction of protein structures and interactions using a three-track neural network, Science **373**, 871 (2021).

[3] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, and A. Rives, Evolutionary-scale prediction of atomic-level protein structure with a language model, Science **379**, 1123 (2023).

[4] J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera *et al.*, Robust deep learning–based protein sequence design using ProteinMPNN, Science **378**, 49 (2022).

[5] C. B. Anfinsen, Principles that govern the folding of protein chains, Science **181**, 223 (1973).

[6] L. L. Porter and L. L. Looger, Extant fold-switching proteins are widespread, Proc. Natl. Acad. Sci. USA **115**, 5968 (2018).

[7] C. M. Dobson, Protein misfolding, evolution and disease, Trends Biochem. Sci. **24**, 329 (1999).

[8] J. Monod, J. Wyman, and J. P. Changeux, On the nature of allosteric transitions: A plausible model, J. Mol. Biol. **12**, 88 (1965).

[9] A. R. Fersht, A. Matouschek, and L. Serrano, The folding of an enzyme: I. Theory of protein engineering analysis of stability and pathway of protein folding, J. Mol. Biol. **224**, 771 (1992).

[10] M. Gruebele, K. Dave, and S. Sukenik, Globular protein folding in vitro and in vivo, Annu. Rev. Biophys. **45**, 233 (2016).

[11] K. A. Dill, Dominant forces in protein folding, Biochemistry **29**, 7133 (1990).

[12] J. D. Treado, Z. Mei, L. Regan, and C. S. O'Hern, Void distributions reveal structural link between jammed packings and protein cores, Phys. Rev. E **99**, 022416 (2019).

[13] Z. Mei, J. D. Treado, A. T. Grigas, Z. A. Levine, L. Regan, and C. S. O'Hern, Analyses of protein cores reveal fundamental differences between solution and crystal structures, Proteins Struct. Funct. Bioinf. **88**, 1154 (2020).

[14] A. T. Grigas, Z. Mei, J. D. Treado, Z. A. Levine, L. Regan, and C. S. O'Hern, Using physical features of protein core packing to distinguish real proteins from decoys, Protein Sci. **29**, 1931 (2020).

[15] A. T. Grigas, Z. Liu, L. Regan, and C. S. O'Hern, Core packing of well-defined X-ray and NMR structures is the same, Protein Sci. **31**, e4373 (2022).

[16] A. T. Grigas, Z. Liu, J. A. Logan, M. D. Shattuck, and C. S. O'Hern, Protein folding as a jamming transition, PRX Life **3**, 013018 (2025).

[17] M. Rubinstein and R. H. Colby, *Polymer Physics* (Oxford University Press, Oxford, 2003).

[18] E. Yamamoto, T. Akimoto, A. Mitsutake, and R. Metzler, Universal relation between instantaneous diffusivity and radius of gyration of proteins in aqueous solution, Phys. Rev. Lett. **126**, 128101 (2021).

[19] W. W. Smith, P.-Y. Ho, and C. S. O'Hern, Calibrated Langevin-dynamics simulations of intrinsically disordered proteins, Phys. Rev. E **90**, 042709 (2014).

[20] L. Hong and J. Lei, Scaling law for the radius of gyration of proteins and its dependence on hydrophobicity, J. Polym. Sci. B Polym. Phys. **47**, 207 (2009).

[21] G. Damaschun, H. Damaschun, K. Gast, D. Gerlach, R. Misselwitz, H. Welfle, and D. Zirwer, Streptokinase is a flexible multi-domain protein, Eur. Biophys. J. **20**, 355 (1992).

[22] E. E. Lattman, Small-angle scattering studies of protein folding, Curr. Opin. Struct. Biol. **4**, 87 (1994).

[23] M. Kataoka and Y. Goto, X-ray solution scattering studies of protein folding, Folding Des. **1**, R107 (1996).

[24] H. Durchschlag, P. Zipper, G. Purr, and R. Jaenicke, Comparative studies of structural properties and conformational changes of proteins by analytical ultracentrifugation and other techniques, Colloid Polym. Sci. **274**, 117 (1996).

[25] Y. Men, J. Rieger, P. Lindner, H.-F. Enderle, D. Lilge, M. O. Kristen, S. Mihan, and S. Jiang, Structural changes and chain radius of gyration in cold-drawn polyethylene after annealing: small-and wide-angle x-ray scattering and small-angle neutron scattering studies, J. Phys. Chem. B **109**, 16650 (2005).

[26] R. Biehl, M. Monkenbusch, and D. Richter, Exploring internal protein dynamics by neutron spin echo spectroscopy, Soft Matter **7**, 1299 (2011).

[27] J. A. McCammon, B. R. Gelin, and M. Karplus, Dynamics of folded proteins, Nature (London) **267**, 585 (1977).

[28] M. Karplus and J. A. Mccammon, Protein structural fluctuations during a period of 100 ps, Nature (London) **277**, 578 (1979).

[29] A. T. Brünger, J. Kuriyan, and M. Karplus, Crystallographic *R* factor refinement by molecular dynamics, Science **235**, 458 (1987).

[30] M. Karplus and J. A. McCammon, Molecular dynamics simulations of biomolecules, Nat. Struct. Biol. **9**, 646 (2002).

[31] K. Lindorff-Larsen, P. Maragakis, S. Piana, M. P. Eastwood, R. O. Dror, and D. E. Shaw, Systematic validation of protein force fields against experimental data, PLoS ONE **7**, e32131 (2012).

[32] Y. Duan and P. A. Kollman, Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution, Science **282**, 740 (1998).

[33] R. Zhou, B. J. Berne, and R. Germain, The free energy landscape for beta hairpin folding in explicit water, Proc. Natl. Acad. Sci. USA **98**, 14931 (2001).

[34] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, How fast-folding proteins fold, Science **334**, 517 (2011).

[35] S. Piana, K. Lindorff-Larsen, and D. E. Shaw, Protein folding kinetics and thermodynamics from atomistic simulation, Proc. Natl. Acad. Sci. USA **109**, 17845 (2012).

[36] V. A. Voelz, G. R. Bowman, K. Beauchamp, and V. S. Pande, Molecular simulation of *ab initio* protein folding for a millisecond folder NTL9(1–39), J. Am. Chem. Soc. **132**, 1526 (2010).

[37] R. E. Burton, G. S. Huang, M. A. Daugherty, P. W. Fullbright, and T. G. Oas, Microsecond protein folding through a compact transition state, J. Mol. Biol. **263**, 311 (1996).

[38] D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. Shan, and W. Wriggers, Atomic-level characterization of the structural dynamics of proteins, Science **330**, 341 (2010).

[39] L. Freddolino, C. B. Harrison, Y. Liu, and K. Schulten, Challenges in protein-folding simulations, Nat. Phys. **6**, 751 (2010).

[40] Y. Liu, J. Strümpfer, L. Freddolino, M. Gruebele, and K. Schulten, Structural characterization of repressor folding from all-Atom molecular dynamics simulations, J. Phys. Chem. Lett. **3**, 1117 (2012).

[41] F. Noé, Beating the millisecond barrier in molecular dynamics simulations, Biophys. J. **108**, 228 (2015).

[42] R. Nassar, E. Brini, S. Parui, C. Liu, G. L. Dignon, and K. A. Dill, Accelerating protein folding molecular dynamics using inter-residue distances from machine learning servers, J. Chem. Theory Comput. **18**, 1929 (2022).

[43] M. Levitt and A. Warshel, Computer simulation of protein folding, Nature (London) **253**, 694 (1975).

[44] B. Berger and T. Leighton, Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete, J. Comput. Biol. **5**, 27 (1998).

[45] D. Thirumalai and D. K. Klimov, Deciphering the timescales and mechanisms of protein folding using minimal off-lattice models, Curr. Opin. Struct. Biol. **9**, 197 (1999).

[46] C. 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 (2000).

[47] A. Liwo, M. Khalili, and H. A. Scheraga, *Ab initio* simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains, Proc. Natl. Acad. Sci. USA **102**, 2362 (2005).

[48] F. Sterpone, S. Melchionna, P. Tuffery, S. Pasquali, N. Mousseau, T. Cragnolini, 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 (2014).

[49] S. Kmiecik and A. Kolinski, One-dimensional structural properties of proteins in the coarse-grained CABS model, in *Prediction of Protein Secondary Structure*, edited by Y. Zhou, A. Kloczkowski, E. Faraggi, and Y. Yang (Humana Press, New York, 2017), Vol. 1484, pp 83–113.

[50] P. C. T. Souza, R. Alessandri, J. Barnoud, S. Thallmair, I. Faustino, F. Grünewald, I. Patmanidis, H. Abdizadeh, B. M. H. Bruininks, T. A. Wassenaar, P. C. Kroon, J. Melcr, V. Nieto, V. Corradi, H. M. Khan, J. Domański, M. Javanainen, H. Martinez-Seara, N. Reuter, R. B. Best *et al.*, Martini 3: a general purpose force field for coarse-grained molecular dynamics, Nat. Methods **18**, 382 (2021).

[51] A. 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 (2012).

[52] P. Kar, S. M. Gopal, Y.-M. Cheng, A. Panahi, and M. Feig, Transferring the PRIMO coarse-grained force field to the membrane environment: simulations of membrane proteins and Helix–Helix association, J. Chem. Theory Comput. **10**, 3459 (2014).

[53] A. Maritan, C. Micheletti, A. Trovato, and J. R. Banavar, Optimal shapes of compact strings, Nature (London) **406**, 287 (2000).

[54] T. X. Hoang, A. Trovato, F. Seno, J. R. Banavar, and A. Maritan, Geometry and symmetry presculpt the free-energy landscape of proteins, Proc. Natl. Acad. Sci. USA **101**, 7960 (2004).

[55] A. T. Grigas, A. Fisher, M. D. Shattuck, and C. S. O’Hern, Connecting polymer collapse and the onset of jamming, Phys. Rev. E **109**, 034406 (2024).

[56] J. C. Gaines, W. W. Smith, L. Regan, and C. S. O’Hern, Random close packing in protein cores, Phys. Rev. E **93**, 032415 (2016).

[57] H. Taketomi, Y. Ueda, and N. Gō, Studies on protein folding, unfolding and fluctuations by computer simulation, Int. J. Pept. Protein Res. **7**, 445 (1975).

[58] I. Bahar, A. R. Atilgan, and B. Erman, Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential, Folding and Design **2**, 173 (1997).

[59] P. Derreumaux, From polypeptide sequences to structures using Monte Carlo simulations and an optimized potential, J. Chem. Phys. **111**, 2301 (1999).

[60] A. Kolinski, Protein modeling and structure prediction with a reduced representation, Acta Biochim. Pol. **51**, 349 (2004).

[61] P. 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 (2013).

[62] G. Wang and R. L. Dunbrack Jr., PISCES: a protein sequence culling server, Bioinformatics **19**, 1589 (2003).

[63] C. Chen, P. Depa, V. G. Sakai, J. K. Maranas, J. W. Lynn, I. Peral, and J. R. Copley, A comparison of united atom, explicit atom, and coarse-grained simulation models for poly (ethylene oxide), J. Chem. Phys. **124**, 234901 (2006).

[64] J. C. Gaines, A. H. Clark, L. Regan, and C. S. O'Hern, Packing in protein cores, J. Phys.: Condens. Matter **29**, 293001 (2017).

[65] J. Liang and K. A. Dill, Are proteins well-packed? Biophys. J. **81**, 751 (2001).

[66] S. Mitternacht, FreeSASA: An open source C library for solvent accessible surface area calculations, F1000Research **5**, 189 (2016).

[67] B. Lee and F. M. Richards, The interpretation of protein structures: estimation of static accessibility, J. Mol. Biol. **55**, 379 (1971).

[68] J. Gaines, S. Acebes, A. Virrueta, M. Butler, L. Regan, and C. O'Hern, Comparing side chain packing in soluble proteins, protein-protein interfaces, and transmembrane proteins, Proteins Struct. Funct. Bioinf. **86**, 581 (2018).

[69] G. Magi Meconi, I. R. Sasselli, V. Bianco, J. N. Onuchic, and I. Coluzza, Key aspects of the past 30 years of protein design, Rep. Prog. Phys. **85**, 086601 (2022).

[70] C. Cardelli, V. Bianco, L. Rovigatti, F. Nerattini, L. Tubiana, C. Dellago, and I. Coluzza, The role of directional interactions in the designability of generalized heteropolymers, Sci. Rep. **7**, 4986 (2017).

[71] I. Coluzza, Transferable coarse-grained potential for *De Novo* protein folding and design, PLoS ONE **9**, e112852 (2014).

[72] https://github.com/jalogan/Constrained-Polymer-Collapse.

[73] G. Wang and R. L. Dunbrack Jr, Pisces: recent improvements to a PDB sequence culling server, Nucleic Acids Res. **33**, W94 (2005).

[74] J. M. Word, S. C. Lovell, J. S. Richardson, and D. C. Richardson, Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation, J. Mol. Biol. **285**, 1735 (1999).

[75] A. Böck, K. Forchhammer, J. Heider, W. Leinfelder, G. Sawers, B. Veprek, and F. Zinoni, Selenocysteine: the 21st amino acid, Mol. Microbiol. **5**, 515 (1991).

[76] R. Longtin, A forgotten debate: is selenocysteine the 21st amino acid? J. Natl. Cancer Inst. **96**, 504 (2004).

[77] V. H. B. Serrão and J. F. Scortecci, Why selenocysteine is unique? Front. Mol. Biosci. **7**, 2 (2020).