JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

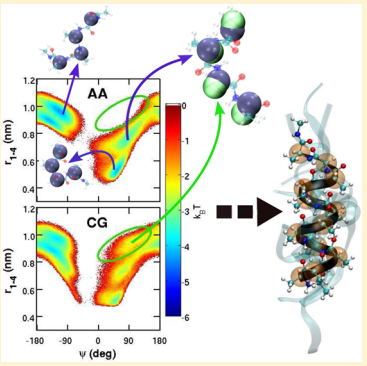
Bottom-Up Coarse-Graining of Peptide Ensembles and Helix−Coil Transitions

Joseph F. Rudzinski and William G. Noid*

Department of Chemistry, The Pennsylvania State University, University Park, Pennsylvania 16802, United States

Supporting Information

ABSTRACT: This work investigates the capability of bottom-up methods for parametrizing minimal coarse-grained (CG) models of disordered and helical peptides. We consider four high-resolution peptide ensembles that demonstrate varying degrees of complexity. For each high-resolution ensemble, we parametrize a CG model via the multiscale coarse-graining (MS-CG) method, which employs a generalized Yvon−Born−Green (g-YBG) relation to determine potentials directly (i.e., without iteration) from the high-resolution ensemble. The MS-CG method accurately describes high-resolution models that fluctuate about a single conformation. However, given the minimal resolution and simple molecular mechanics potential, the MS-CG method provides a less accurate description for a high-resolution peptide model that samples a disordered ensemble with multiple distinct conformations. We employ an iterative g-YBG method to develop a CG model that more accurately describes the relevant distribution functions and free energy surfaces for this disordered ensemble. Nevertheless, this more accurate model does not reproduce the cooperative helix−coil transition that is sampled by the high resolution model. By comparing the different models, we demonstrate that the errors in the MS-CG model primarily stem from the lack of cooperative interactions afforded by the minimal representation and molecular mechanics potential. This work demonstrates the potential of the MS-CG method for accurately modeling complex biomolecular structures, but also highlights the importance of more complex potentials for modeling cooperative transitions with a minimal CG representation.



INTRODUCTION

Atomically detailed molecular dynamics (MD) simulations provide tremendous insight into protein structure and fluctuations on nanosecond time scales.[1] Nevertheless, despite great strides in computational methods and resources, atomically detailed models remain prohibitively inefficient for investigating many complex biological processes, such as peptide aggregation, that evolve on much longer time scales.[2] Consequently, lower resolution coarse-grained (CG) models continue to enjoy surging popularity.[3−6]

In particular, minimal CG models have proven to be particularly useful for modeling protein folding and interactions. By eliminating an explicit treatment of solvent and by representing each amino acid with a single site, which is usually associated with the corresponding α-carbon, these models provide tremendous efficiency.[7] Moreover, because of the regularity of the protein backbone geometry,[8,9] this remarkably sparse minimal representation still allows for an accurate atomic reconstruction of the peptide backbone.[10,11]

Accordingly, a vast array of off-lattice minimal models have been parametrized by various means and for various purposes. For instance, the seminal studies of Thirumalai and coworkers[12,13] represented proteins with a "reduced alphabet" (i.e., amino acids are distinguished by their character as, e.g., hydrophobic or polar) and employed simple potentials to investigate universal features of protein folding. Conversely, native-based Gō models[14−16] and network models[17,18] represent proteins with an "extended alphabet" (i.e., amino acids are distinguished by their location in the protein sequence) and employ biased potentials that stabilize known structures in order to characterize the folding and fluctuations of specific proteins.[19,20] Additionally, minimal models have been employed to characterize generic aspects of cellular crowding, unfolded protein ensembles, and peptide aggregation.[21−24] These latter minimal models have often been parametrized via top-down approaches[25,26] that attempt to capture emergent, often thermodynamic, properties.

In contrast, several recent studies have employed bottom-up approaches to parametrize CG peptide models that accurately describe the ensembles sampled by all-atom (AA) models.[27−36] The many-body potential of mean force (PMF) is the appropriate potential for a CG model that reproduces all structural features of the mapped AA ensemble, i.e., the ensemble generated by mapping the AA ensemble to the CG representation. Bottom-up methods typically approximate the many-body PMF with simple molecular mechanics potentials that include separate terms for bond, angle, torsion, and pair "nonbonded" interactions. These various terms are often iteratively refined in order to reproduce target 1-D distribution functions for corresponding degrees of freedom in the CG model.[37−40] In general, these studies have described the

Received: November 5, 2014

DOI: 10.1021/ct5009922
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

mapped AA ensemble with reasonable accuracy. For instance, the CG model of Bezkorovaynaya et al.[31] reproduced the relevant 1-D distribution functions of the underlying ensemble quite accurately, but did not accurately describe the cross-correlations between the angle and torsion degrees of freedom along the CG peptide backbone.

The present work expands upon previous studies by further investigating the capabilities and limitations of bottom-up coarse-graining methods for determining minimal peptide models that accurately describe AA conformational ensembles for helical and disordered peptides. In particular, we employ the multiscale coarse-graining (MS-CG) method[41,42] to determine potentials that provide a variationally optimal approximation to the many-body PMF.[43,44] In contrast to many other bottom-up structure-based approaches, which iteratively refine the model potential to reproduce particular structural features, the MS-CG method employs a generalized Yvon−Born−Green (g-YBG) equation[45−47] to directly (i.e., noniteratively) determine the approximate CG potential from the correlations that are observed in the mapped AA ensemble.

In a certain sense, the MS-CG/g-YBG approach is quite elegant, since it provides a direct solution to the inverse problem of inferring potentials from the AA ensemble. Moreover, the MS-CG/g-YBG approach also holds considerable computational promise. While iterative methods require the solution to nonlinear optimization problems for a large number of parameters, the MS-CG/g-YBG method requires only the solution of a linear least-squares problem. However, the MS-CG/g-YBG procedure rests upon the fundamental assumption that the form of the CG potential is sufficiently flexible for reproducing the relevant cross-correlations of the mapped AA ensemble.[48] Clearly this assumption depends not only upon the AA model, but also upon the complexity of the CG potential and the CG representation of the AA model.[49,50]

This assumption is quite central to bottom-up CG methods. In cases that this assumption is valid, the MS-CG model will accurately reproduce the structure of the mapped AA ensemble. However, in cases that this assumption is not valid, the MS-CG model may not accurately describe this ensemble. Moreover, in this case, iterative bottom-up methods may reproduce the target 1-D distribution functions for individual degrees of freedom, but will do so at the expense of distorting the cross-correlations between these degrees of freedom. This distortion may prove especially detrimental for describing the complex, hierarchical structures of proteins and other biomolecules.

Accordingly, the objective of the present work is to assess this basic assumption in the context of parametrizing minimal CG peptide models with implicit solvent. We demonstrate that the MS-CG/g-YBG framework directly determines accurate minimal models for peptides that fluctuate about a single well-defined conformation. These results complement the results of previous studies[33−35] with the MS-CG method that represented peptides with slightly higher resolution and employed explicit solvent. However, given the minimal representation and molecular mechanics potential, the MS-CG/g-YBG framework provides a less accurate description for more complex ensembles that sample multiple conformations. In order to investigate this discrepancy, we employ an iterative g-YBG (iter-gYBG) method[51,52] to parametrize CG models for these more complex ensembles. These iter-gYBG models provide a reasonably accurate description of the mapped AA ensembles and reveal the structural features of the mapped ensembles that are not consistent with the assumptions of the MS-CG/g-YBG framework. Finally, by adapting the g-YBG framework to account for these problematic features of the mapped AA ensemble, we significantly improve the accuracy of the resulting MS-CG minimal peptide models.

THEORY

The MS-CG, g-YBG, and iterative g-YBG methods have been extensively discussed in previous papers.[41−43,47,50−54] We briefly summarize the details that are particularly relevant for the present work. We first consider an AA model with a configuration, $\mathbf{r}$, that is defined by the Cartesian coordinates for n atoms. We assume that a mapping function determines a configuration, $\mathbf{R}$, for the CG model as a linear function of $\mathbf{r}$. In the present work, the CG configuration $\mathbf{R}$ corresponds to the Cartesian coordinates of the α-carbons in the AA model. It is convenient to define the "mapped AA ensemble" as the ensemble of CG configurations that is generated by applying the mapping to each configuration that is sampled by the AA model. The many-body potential of mean force (PMF), $W(\mathbf{R})$, is the appropriate potential for a CG model that quantitatively reproduces all structural properties of this mapped AA ensemble.[5] The PMF may be defined, to within a configuration independent constant, by

$$W(\mathbf{R}) = -k_{\mathrm{B}}T \ln p_{\mathrm{R}}(\mathbf{R}) + \text{const} \tag{1}$$

where $p_{\mathrm{R}}(\mathbf{R})$ is the probability for the AA model to sample a configuration $\mathbf{r}$ that maps to the CG configuration $\mathbf{R}$.[5] In general, the PMF is a complex, many-body function. The MS-CG method determines the parameters, $\boldsymbol{\phi}^0$, for a molecular mechanics potential energy function, $U(\boldsymbol{\phi}^0)$, that provides a variationally optimal approximation to the PMF.[43,45] According to the MS-CG objective function,[41−44] this optimal approximation is determined by directly inverting the normal system[45] of linear equations:

$$\boldsymbol{b}^{\mathrm{AA}} = \mathbf{G}^{\mathrm{AA}}\boldsymbol{\phi}^0 \tag{2}$$

In eq 2, $\boldsymbol{b}^{\mathrm{AA}}$ is a vector of ensemble averages that can be expressed either in terms of AA forces[43,45] or in terms of a corresponding set of structural correlation functions.[46,47,54]

We focus on the common case that the nonbonded contribution to the CG potential, U, is represented by a sum of central pair potentials and each of these pair potentials is represented by a set of flexible basis functions (e.g., spline functions). In this case, a subset of the elements in $\boldsymbol{b}^{\mathrm{AA}}$ is in 1−1 relationship with the radial distribution functions (rdfs) for the CG sites in the mapped AA ensemble. $\mathbf{G}^{\mathrm{AA}}$ is a matrix of ensemble averages that quantify the cross-correlations between pairs of CG degrees of freedom in the mapped AA ensemble. This matrix allows the MS-CG method to decompose the force correlation vector, $\boldsymbol{b}^{\mathrm{AA}}$, into contributions from the various terms in the CG potential, $U(\boldsymbol{\phi}^0)$.

Equation 2 is related to a generalized Yvon−Born−Green equation[46,47] that exactly relates a given set of potential parameters, $\boldsymbol{\phi}$, to the vector, $\boldsymbol{b}(\boldsymbol{\phi})$, of force correlation functions and to the matrix, $\mathbf{G}(\boldsymbol{\phi})$, of cross-correlations that are generated by equilibrium sampling of a CG model with potential $U(\boldsymbol{\phi})$:

$$\mathbf{b}(\boldsymbol{\phi}) = \mathbf{G}(\boldsymbol{\phi})\boldsymbol{\phi} \tag{3}$$

According to eqs 2 and 3, the MS-CG method employs the g-YBG relation to determine potential parameters $\boldsymbol{\phi}^0$ that reproduce the AA force correlation vector, $\boldsymbol{b}^{\mathrm{AA}}$, but employs the

DOI: 10.1021/ct5009922
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

matrix, $\mathbf{G}^{AA}$, of cross-correlations that are observed in the mapped AA ensemble to approximate the cross-correlations, $\mathbf{G}(\boldsymbol{\phi}^0)$, that will be generated by the CG potential $U(\boldsymbol{\phi}^0)$. If $\mathbf{G}(\boldsymbol{\phi}^0) = \mathbf{G}^{AA}$, then $\boldsymbol{b}(\boldsymbol{\phi}^0) = \boldsymbol{b}^{AA}$ and the CG model will reproduce the corresponding AA rdfs. However, this will generally not be the case because it would require that the CG model reproduce the higher order cross-correlations of the AA model.[48] Thus, if an MS-CG model does reproduce the rdfs, this suggests that it also likely reproduces higher order correlations of the AA model.

Iterative bottom-up CG procedures seek to determine potential parameters $\boldsymbol{\phi}^*$ that will reproduce the 1-D equilibrium distributions of the mapped AA ensemble for the relevant degrees of freedom in the CG model. In the context of the g-YBG framework,[50−52] this corresponds to determining the force field coefficients $\boldsymbol{\phi}^*$ such that

$$\boldsymbol{b}^{AA} = \mathbf{G}(\boldsymbol{\phi}^*)\boldsymbol{\phi}^* \tag{4}$$

In contrast to eq 2 for the MS-CG potential, eq 4 corresponds to a self-consistent g-YBG equation that determines the force field coefficients, $\boldsymbol{\phi}^*$, that reproduce $\boldsymbol{b}^{AA}$, while using the cross-correlations $\mathbf{G}(\boldsymbol{\phi}^*)$ sampled by a CG model with the corresponding potential $U(\boldsymbol{\phi}^*)$.[51]

In brief, the iter-gYBG procedure first determines the MS-CG potential parameters $\boldsymbol{\phi}^0$ according to eq 2. Simulations with this CG model determine the resulting matrix, $\mathbf{G}(\boldsymbol{\phi}^0)$, of cross-correlations. The iter-gYBG method then determines a new set of potential parameters by solving eq 4 for $\boldsymbol{\phi}^*$, while approximating $\mathbf{G}(\boldsymbol{\phi}^*)$ with the correlations, $\mathbf{G}(\boldsymbol{\phi}^0)$, generated by the preceding CG model. This procedure is iterated until the CG potential adequately reproduces the AA force correlation vector $\boldsymbol{b}^{AA}$ and, thus, also the AA pair structure. Note that this implies $\mathbf{G}(\boldsymbol{\phi}^*) \neq \mathbf{G}^{AA}$, i.e., the final CG model reproduces rdfs of the AA model at the expense of distorting higher order cross-correlations. As discussed further below, we have heuristically modified the method to improve its robustness for systems with complex intramolecular structure.[50]

METHODS

This section summarizes the key details of our calculations. The Supporting Information provides a much more detailed description.

Simulation Details. All reported molecular dynamics (MD) simulations were performed in the constant NVT ensemble with the Gromacs 4.5.3 simulation suite[55] according to standard procedures.[56−61] All-atom (AA) peptide simulations were performed at a temperature T = 298 K, while employing the OPLS-AA force field[62] to model peptide interactions and the SPC/E model[63] to describe the solvent (when applicable). We employed LINCS[64] to rigidly constrain all bonds that involve H atoms. The AA peptide models were capped with an N-terminal acetyl group and a C-terminal N-methyl amide group.

High-Resolution Models. The objective of the present work is to assess the capability of the MS-CG/g-YBG approach to model complex free energy surfaces. Accordingly, we considered four distinct high-resolution peptide models that sample ensembles of varying helicity and complexity. Three of these high-resolution models represent short alanine peptides in conventional atomic detail. The fourth high-resolution model is a C-α native-based,[14−16] i.e., Gō, model. Although the Gō model represents each amino acid with a single site, it employs considerably higher resolution than the MS-CG model that we parametrize for the Gō model. While the Gō model treats each site as a distinct type and employs 17 distinct types of pair potentials, the corresponding MS-CG model employs only 3 distinct types of pair potentials to model the same set of interactions.

FCP1. We performed MD simulations of a C-α Gō model[14−16] for the C-terminal residues (944−961) of the FCP1 protein.[65] This region of FCP1 is intrinsically disordered in isolation.[66] However, when interacting with the C-terminal domain of the Rap74 subunit of Transcription Factor IIF, the FCP1 residues 944−957 fold to form an α helix, while the final 4 residues remain disordered.[67] We previously[68] employed the Structure-based Models in Gromacs Web server[69] (http://smog.ucsd.edu) to construct a Gō model for the Rap74-FCP1 system from the published crystal structure of the complex (PDBID: 1J2X).[67] The resulting intramolecular potential for the C-terminal FCP1 peptide defined a high resolution model for FCP1. We employed the Gō model to sample an ensemble with an average helical content of $\langle Q_{hel} \rangle = 0.65$, where Q_{hel} is defined below. Although prior studies suggest that the intrinsically disordered FCP1 peptide samples a similar degree of helicity *in vitro*,[68,70,71] we emphasize that the present calculations do not attempt to develop or investigate a realistic model for FCP1. Rather, our calculations only employ the FCP1 Gō model as a convenient means for generating an ensemble of peptide conformations that fluctuate about a simple helical structure. We sampled a total of 2.8 million configurations for this model and employed the last 2.5 million for subsequent analysis.

AA Model for Alanine 12-mer in Vacuum. We performed a 320 ns AA MD simulation of a capped alanine 12-mer in vacuum. After the first 20 ns, we sampled configurations every 1 ps to obtain a total of 300 000 configurations for subsequent analysis. We note that the OPLS-AA force field may not necessarily provide an accurate description for peptides in vacuum. However, as in the case of FCP1, we emphasize that we do not employ this model to accurately describe peptides in vacuum but rather as a convenient means for generating an ensemble of peptide conformations that fluctuate about a precise helical structure.

AA Models for Solvated Alanine Oligomers. We also performed AA MD simulations with explicit solvent for a capped alanine tetramer and a capped alanine 12-mer.

After equilibration, we performed five independent 200 ns simulations of the solvated alanine tetramer. We sampled configurations every 1 ps to obtain a total of 1 million configurations for subsequent analysis.

We performed a single 600 ns production simulation of the solvated alanine 12-mer. After the initial 50 ns, we sampled configurations every 1 ps, which yielded a total of 550 000 configurations for subsequent analysis. This simulation resulted in a heterogeneous ensemble that included helical, coil, and extended structures. Over the course of the AA trajectory, we observed 6 folding events, for which Q_{hel} increased from <0.50 to >0.98, and 15 partial folding events, for which Q_{hel} increased from <0.55 to >0.82. Although the resulting AA ensemble is unlikely to be completely converged, it provides a suitable ensemble for assessing the capability of the MS-CG method to accurately model a complex disordered AA ensemble.

CG Models. We constructed at least one CG model for each high-resolution peptide model. The CG peptide models represented amino acids with a single site that corresponded

DOI: 10.1021/ct5009922
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

to the residue α-carbon. These models did not explicitly describe the solvent environment. Instead, the solvent effects were implicitly incorporated into the molecular mechanics potentials that describe the interactions between the peptide sites, as is done in, e.g., the Honeycutt−Thirumalai model[12,13] or Gō models.[14−16] The CG model for FCP1 included 15 sites for residues 944−958. The CG models for AA models of alanine peptides associated a site with each α-carbon of the AA model. These CG models did not explicitly represent the capping groups that were present in the AA models.

For each CG model, we parametrized a distinct molecular mechanics potential with both bonded and nonbonded contributions. The bonded potential included bond, angle, and dihedral potentials for each pair, triple, and quadruple, respectively, of consecutive CG sites along the peptide chain. Each CG model employed a single bond potential to model all bonds. While the CG models of FCP1 and the alanine tetramer employed a single angle potential to model all angles, the CG models for the alanine 12-mers (both solvated and in vacuum) employed 5 distinct angle potentials, including distinct potentials for the first two and last two angles of the peptide chain, as well as a single angle potential for all interior angles. With the exception of the alanine tetramer, the CG models all employed distinct dihedral potentials for the first two and last two dihedral angles along the chain, i.e., they employed five distinct dihedral potentials.

The nonbonded potential included central pair potentials between each pair of sites that were separated by more than two bonds along the chain. In each CG model, a single "nonspecific" pair potential was employed to model all interactions between sites that were separated by more than 4 bonds. Interactions between sites separated by exactly 3 bonds (1−4 pairs) and exactly 4 bonds (1−5 pairs) were treated with distinct 1−4 and 1−5 potentials, respectively. In the CG model of FCP1, all 1−4 interactions were modeled with a single 1−4 potential and all 1−5 interactions were modeled with a single 1−5 potential. The CG models for alanine 12-mers, though, required a more complex treatment of the 1−4 and 1−5 interactions. These models employed 5 distinct 1−4 potentials and 5 distinct 1−5 potentials in order to distinguish the 1−4 and 1−5 interactions, respectively, of the two terminal sites on each side of the peptide. The Supporting Information presents results for CG models with simpler potentials.

We note that the increased complexity of the CG potential attempts to compensate for the dramatically reduced CG description of the peptide conformation. For instance, because they explicitly describe the atomic structure of each residue and the surrounding solvent, atomic models are capable of distinguishing the local environments of each residue while employing relatively few "atom types" and employing the same potential for each pair of atom types, irrespective of their context. In contrast, because they eliminate all such information about the residue structure and solvation environment, the CG models distinguish between different environments by employing different "types" of nonbonded interactions.

Force Field Calculations. Each calculated term in the CG potential was represented by a discrete set of basis functions of a single variable.[44] Different basis functions (e.g., linear spline, cubic B-spline, etc.) were employed for different types of interactions. The coefficients for these basis functions were parametrized via either the MS-CG or iter-gYBG method. We note that the present treatment of rigidly constrained bonds is not rigorously consistent with the MS-CG theory.[43] Nevertheless, we expect that this should not substantively impact the present results.

For each high-resolution ensemble, we parametrized a MS-CG model by solving eq 2 for the potential parameters. For the solvated alanine 12-mer, we also parametrized MS-CG models for specific regions of configuration space by solving eq 2, while determining $\boldsymbol{b}^{\mathrm{AA}}$ and $\mathbf{G}^{\mathrm{AA}}$ from the configurations that sampled the corresponding regions of the free energy surface (FES). Additionally, as described in the Results, we investigated the errors in the MS-CG model for the solvated alanine 12-mer by calculating the MS-CG force field according to eq 2, but using a modified correlation matrix, $\mathbf{G}^{\mathrm{AA\text{-}mod}}$, in the place of $\mathbf{G}^{\mathrm{AA}}$.

We employed a modified version of the iter-gYBG method[51] to parametrize CG models for the solvated alanine tetramer and 12-mer. As described in our prior work, this heuristic modification, which applies only for bond and angle interactions, employs an exact decomposition of $\mathbf{G}$ into direct and indirect contributions.[48] For bond and angle interactions, we determine the direct contribution to $\mathbf{G}$ from the mapped AA ensemble and then adapted the indirect contributions to $\mathbf{G}$ based upon the CG models that are generated during the iterative procedure. This modification increased the stability and robustness of the calculations, especially for models with complex intramolecular structure.

As described in the Supporting Information, the iter-gYBG calculations for the solvated alanine 12-mer employed reference potentials[72] for the bonded potential in order to improve the robustness of the method. We determined the reference potentials for the CG bonds and angles via direct Boltzmann inversion of the corresponding mapped AA distributions. We determined the reference potentials for the CG dihedrals from the MS-CG potential for the solvated alanine 12-mer. The iter-gYBG procedure determined corrections to the reference bond and dihedral potentials, which were represented by a linear spline and Fourier series, respectively, but not for the reference angle potential. As in our previous work,[50] we computed the reference contribution, $\boldsymbol{b}^{\mathrm{ref}}$, to the $\boldsymbol{b}^{\mathrm{AA}}$ vector from the mapped AA ensemble. We then iteratively solved for the CG potential parameters after replacing $\boldsymbol{b}^{\mathrm{AA}}$ with $\delta\boldsymbol{b} = \boldsymbol{b}^{\mathrm{AA}} - \boldsymbol{b}^{\mathrm{ref}}$ in eq 4.

Our previous study[50] demonstrated that the iter-gYBG method did not always converge. Instead, after initially converging upon an accurate model after a few iterations, the method sometimes diverged to less accurate models. In the present work we find that, over the course of many iterations, the iter-gYBG method repeatedly approaches to and diverges from an accurate force field. Accordingly, we have developed several metrics to identify the optimal iter-gYBG model. The Supporting Information reviews these criteria and briefly assesses the convergence of the iter-gYBG method for the present peptide models.

As in our prior calculations,[50] we modified the g-YBG system of linear algebraic equations in order to automate the iterative procedure, improve its robustness, and minimize user interference as previously described.[50] As described in the Supporting Information, these modifications included removing force field coefficients for interactions that were rarely sampled, introducing constraints to ensure periodicity of dihedral potentials, smoothly switching nonbonded forces to zero over the last sampled grid points, and regularizing central pair interactions to avoid overfitting the statistical noise. We extensively tested these modifications to ensure that they minimally impacted the MS-CG calculation for well-behaved

DOI: 10.1021/ct5009922
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

test cases. We then solved the modified linear equations[73] via singular value decomposition[74] after applying right−left preconditioning[40,75] to render the linear equations dimensionless. Finally, as described in the Supporting Information, we manually modified the calculated dihedral potentials for the vacuum alanine 12-mer, in order to reasonably extrapolate the potentials to dihedral angles that were not sampled in the AA simulation.

Structural Analysis. In addition to analyzing structural distribution functions along individual degrees of freedom, we also examined 2-D free energy surfaces (FES's) for pairs of order parameters. We considered several order parameters that are functions of the CG configuration, $\mathbf{R}$. The fractional helical content, or helicity, is defined: $Q_{\mathrm{hel}}(\mathbf{R}) = (1/N_{\mathrm{hel}})\sum_{i-j=3} \exp[-(1/2\sigma^2)(R_{ij} - R_0)^2]$, where R_{ij} is the distance between site i and j in configuration $\mathbf{R}$, N_{hel} is the number of 1−4 pairs, $R_0 = 0.5$ nm, and $\sigma^2 = 0.02$ nm^2. The radius of gyration, R_g, is defined as[55] $R_g(\mathbf{R}) = (\sum_{i=1}^{N}|\mathbf{R}_i - \mathbf{R}_{\mathrm{com}}|^2)^{1/2}$, where $\mathbf{R}_i$ is the position of site i, $\mathbf{R}^{\mathrm{com}}$ is the center of mass position, and N is the total number of CG sites. The deviation from a perfectly helical configuration, $\mathbf{R}^{\mathrm{hel}}$, is characterized by the RMSD and DRMS metrics.[55] The root mean squared deviation (RMSD) is defined as $\mathrm{RMSD}(\mathbf{R}|\mathbf{R}^{\mathrm{hel}}) = ((1/N)\sum_{i=1}^{N}|\mathbf{R}_i - \mathbf{R}_i^{\mathrm{hel}}|^2)^{1/2}$, where $\mathbf{R}_i$ and $\mathbf{R}_i^{\mathrm{hel}}$ are the coordinates of site i in configurations $\mathbf{R}$ and $\mathbf{R}^{\mathrm{hel}}$, respectively. The distance root-mean-square (DRMS) is defined as $\mathrm{DRMS}(\mathbf{R}|\mathbf{R}^{\mathrm{hel}}) = ((1/N^2)\sum_{i=1}^{N}\sum_{j=1}^{N}(R_{ij}-R_{ij}^{\mathrm{hel}})^2)^{1/2}$, where R_{ij} and R_{ij}^{hel} are the distances between sites i and j in structures $\mathbf{R}$ and $\mathbf{R}^{\mathrm{hel}}$, respectively. The root-mean-square fluctuation (RMSF) of a CG site i is defined as[55] $\mathrm{RMSF}(i) = (\langle|\mathbf{R}_i - \bar{\mathbf{R}}_i|^2\rangle)^{1/2}$, where $\bar{\mathbf{R}}_i$ is the average position of site i and the brackets denote an ensemble average. Before computing the RMSD or RMSF, a least-squares superposition was performed for each configuration, $\mathbf{R}$, with respect to the reference structure, $\mathbf{R}^{\mathrm{hel}}$.

RESULTS

In this study, we investigated the capability of bottom-up structure-based methods to parametrize minimal CG models that accurately describe the free energy surfaces of helical and disordered peptides. Accordingly, we considered four high-resolution peptide models that sampled ensembles with varying degrees of complexity and helicity. We considered two high-resolution models that fluctuated about well-defined helices: (1) a Gō model[14−16] for a 15-residue segment of FCP1[65] and (2) an all-atom (AA) model for a 12-residue alanine peptide in vacuum. We also considered explicitly solvated AA models of 4- and 12-residue alanine peptides, which generated more complex conformational ensembles that demonstrate helix−coil transitions. For each of these high-resolution models, we parametrized a CG model via the MS-CG method[41−44] and, in some cases, also via the iter-gYBG method.[50−52] In each case, the CG model represented the peptide with sites at α-carbons, while implicitly incorporating solvent effects into the potential for the peptide CG sites. We assessed the quality of each CG model by comparison with the corresponding mapped AA ensemble. We identified specific structural features of the mapped AA ensemble that are not accurately described by the minimal peptide resolution and simple molecular mechanics potential. Finally, we demonstrated that these features are the dominant source of error in the MS-CG models for disordered peptides.

Structured Peptides. *Flexible Helices.* We first employed a C-α Gō model for the FCP1 peptide in order to generate an ensemble of conformations with simple fluctuations about a helical structure. Although this Gō model represents each residue with a single site, for our purposes it is a "high-resolution" model since it employs 17 distinct pair potentials as well as distinct bond, angle, and dihedral potentials for each instance of these interactions. These potentials explicitly bias the ensemble to sample the folded peptide conformation from the corresponding PDB structure (PDBID: 1J2X).[67] In contrast, the MS-CG model treats the interactions between the 15 sites with only 3 types of pair potentials: (1) a 1−4 potential to model interactions between sites separated by exactly 3 bonds, (2) a 1−5 potential to model interactions between sites separated by exactly 4 bonds, and (3) a "nonspecific" pair potential to model interactions between all sites separated by more than 4 bonds. The bonded potential for the MS-CG model employs a single bond potential and a single angle potential to model all instances of these interactions. The MS-CG potential also employs 5 distinct dihedral potentials in order to distinguish the conformational tendencies of the two dihedrals at each terminal of the peptide. However, we obtain similar results for the FCP1 peptide when using only a single potential to model all dihedral angles.

The MS-CG model reproduces the ensemble sampled by the Gō model with reasonably high accuracy. The MS-CG model very accurately reproduces the distributions of the mapped ensemble for the bond, angle, and dihedral degrees of freedom. The MS-CG model also accurately reproduces the mapped distributions for 1−4, 1−5, and nonspecific pairs, i.e., pairs separated by more than 4 bonds (see panel a of Figure S1 in the Supporting Information). Figure 1 further characterizes the ensembles that are sampled by the two models for the FCP1 peptide. Panels a and b present the free energy surface (FES) as a function of helicity, Q_{hel}, and the radius of gyration, R_g, for the Gō and MS-CG models, respectively. Panels c and d of Figure 1 compare the distributions that are sampled by the Gō and MS-CG models for the DRMS metric, which describes deviations from the FCP1 crystal structure, and for the RMSF metric, which quantifies the fluctuations of each residue. In comparison to the Gō model, the MS-CG model slightly undersamples highly helical conformations. The MS-CG model also slightly overestimates the average helicity of the Gō model ($\langle Q_{\mathrm{hel}}\rangle_{\mathrm{MSCG}} = 0.68$, $\langle Q_{\mathrm{hel}}\rangle_{G\bar{o}} = 0.65$). Nevertheless, Figure 1 demonstrates that the MS-CG model reproduces the free energy surface of the Gō model quite accurately. Additionally, the MS-CG model reproduces the DRMS and RMSF distributions quite accurately

Precise Helices. In order to generate a second high-resolution ensemble with a well-defined helical structure, we performed simulations with the OPLS-AA model[62] for a capped 12-residue alanine peptide in vacuum. In comparison to the Gō-model, which sampled a rather "floppy" helix, this AA model sampled a very precise helix that only exhibited slight unfolding from the peptide termini. As in the preceding case, the MS-CG model included distinct potentials to model 1−4 and 1−5 interactions, as well as a nonspecific pair potential between sites separated by more than 4 bonds. However, because the interior residues sampled very precise helical conformations, while the terminal residues sampled less rigid helical conformations, we found it necessary to employ distinct potentials to model the interactions of the two terminal sites on each end of the peptide. Thus, the MS-CG model for the alanine 12-mer in vacuum employed 5 distinct angle and 5 distinct dihedral potentials. In addition, the model employed 5 distinct 1−4 potentials and 5 distinct 1−5 potentials, along with

DOI: 10.1021/ct5009922
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

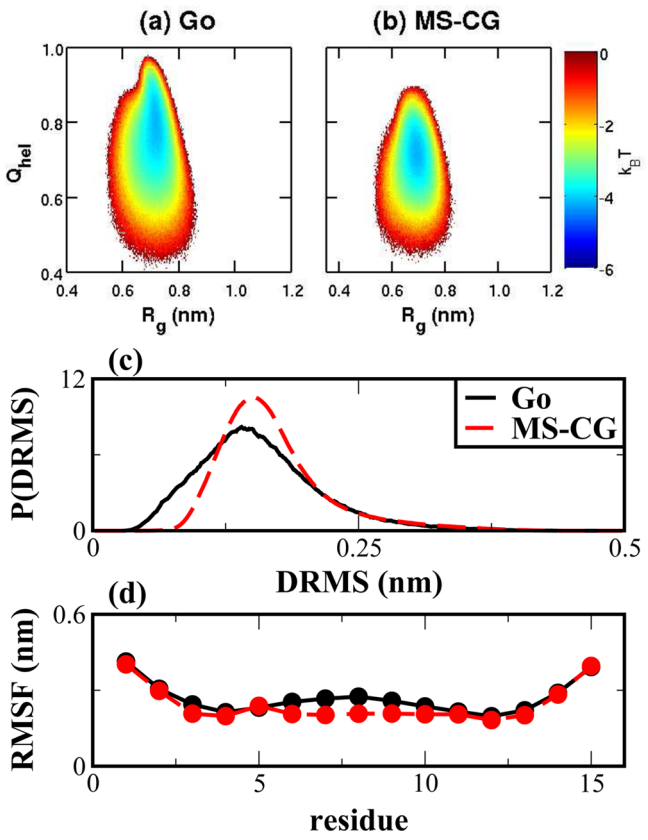


Figure 1. Comparison of Gō and MS-CG models for the FCP1 peptide. Panels (a) and (b) present the FES as a function of helicity, Q_{hel}, and the radius of gyration, R_g, for the Gō and MS-CG models, respectively. Panels (c) and (d) present the DRMS distribution, which describes the deviations of the peptide from a perfect helix, and the RMSF, which describes the fluctuations of each residue, respectively, for the Gō (solid, black curves) and MS-CG models (dashed, red curves).

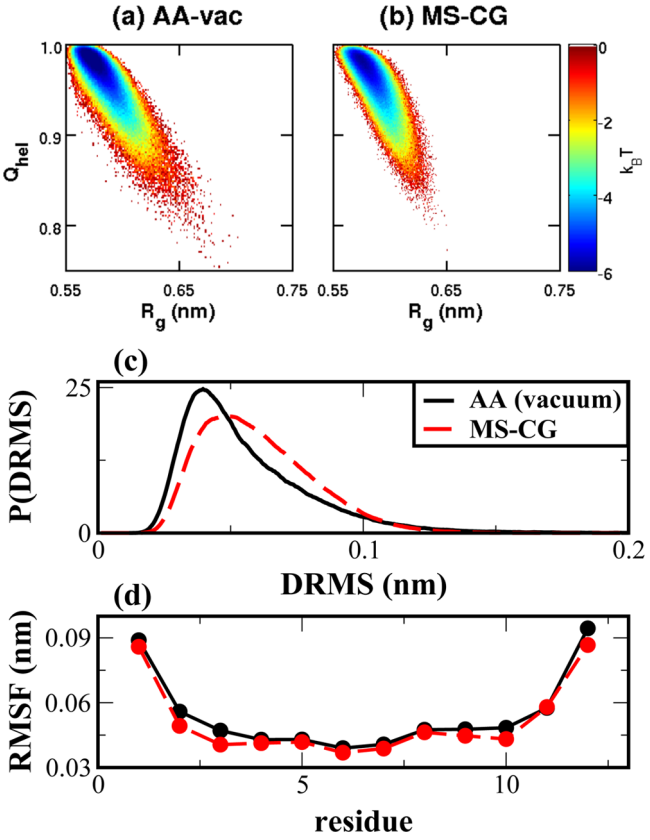


Figure 2. Comparison of AA and MS-CG models for the alanine 12-mer in vacuum. Panels a and b present the FES as a function of helicity, Q_{hel}, and the radius of gyration, R_g, for the AA and MS-CG models, respectively. Panels c and d present the DRMS distribution and the RMSF of each residue, respectively, for the AA (solid, black curves) and MS-CG models (dashed, red curves).

a single nonspecific pair potential. When we did not distinguish between the interior and terminal sites, the resulting MS-CG model underestimated the helicity of the interior residues and overestimated the helicity of the termini.

The MS-CG model quantitatively reproduces the bond, angle, dihedral, 1−4, and 1−5 distributions of the mapped AA ensemble and also qualitatively reproduces the corresponding nonspecific pair distribution (see panel b of Figure S1 in the Supporting Information.) In correspondence with Figure 1, Figure 2 employs the same metrics to compare the ensembles generated by the AA and MS-CG models for the alanine 12-mer in vacuum. Clearly, the MS-CG model provides a very accurate description of the mapped AA ensemble.

Helix−Coil Transition for a Single Peptide Unit. As demonstrated by the preceding calculations, the MS-CG method determines accurate minimal models for peptides that fluctuate about a well-defined helical structure. We next considered whether bottom-up coarse-graining methods can determine minimal peptide models that accurately describe the helix−coil transition of a high-resolution AA model. We constructed a high-resolution ensemble for the helix−coil transition by simulating the OPLS-AA model for a capped 4-residue alanine peptide in explicit SPC/E solvent.[63] For this ensemble, we parametrized CG models with 4 sites that correspond to the α-carbons of the AA model and modeled the CG interactions with 3 equivalent bond potentials, 2 equivalent angle potentials, 1 dihedral potential, and a single 1−4 pair potential. Because the MS-CG model provided limited accuracy for this ensemble, we employed the iter-gYBG method to investigate the source of this discrepancy.

Panels a, b, and c of Figure 3 present the simulated distributions for the two angles, θ, the one dihedral angle, ψ, and the 1−4 distance, r_{1-4}, between the first and last CG sites, respectively. These degrees of freedom all sample multimodal distributions in the mapped AA ensemble (solid black curves). In particular, helical conformations correspond to the peaks at $\theta \approx 91°$, $\psi \approx 57°$, and $r_{1-4} \approx 0.53$ nm, while extended configurations correspond to $\theta \approx 122°$, $\psi \approx -120°$, and $r_{1-4} \approx 0.93$ nm. The MS-CG model (dashed red curves) qualitatively reproduces these distributions, while the iter-gYBG model (dashed-dotted green curves) reproduces each distribution with nearly quantitative accuracy. The slight errors in the r_{1-4} and ψ distributions of the iter-gYBG model appear to result from a competition between accurately modeling the two interactions simultaneously. This may result either from fundamental limitations of the minimal peptide representation or, alternatively, from the insensitivity of the iter-gYBG procedure (see Figure S2 of the Supporting Information).

Figure 4 presents the FES's that are sampled by the three models as a function of r_{1-4} and ψ. The AA model (panel a) samples helical (H), extended (E1 and E2), and intermediate (I) conformations. The MS-CG model (panel b) samples the I,

DOI: 10.1021/ct5009922
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

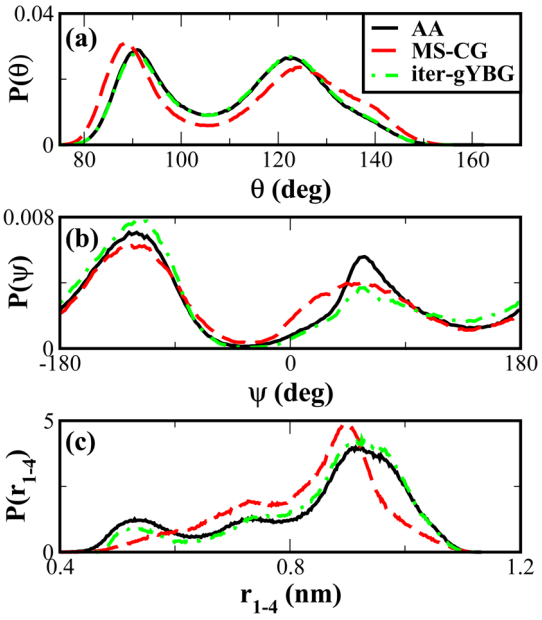


Figure 3. Comparison of AA, MS-CG, and iter-gYBG models for the solvated alanine tetramer. Panels a, b, and c present the 1-D distributions sampled in each model by the CG angles, θ, the CG dihedral angle, ψ, and the distance, r_{1-4}, between the first and last sites, respectively. The solid black, dashed red, and dashed-dotted green curves correspond to distributions sampled by the AA, MS-CG, and iter-gYBG models, respectively.

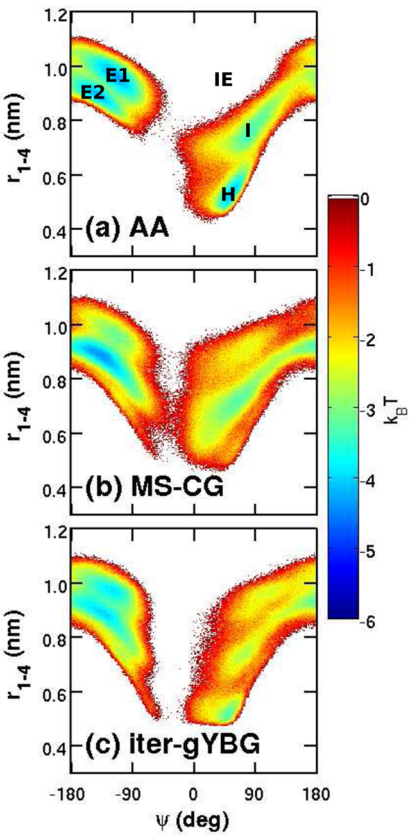


Figure 4. FES's for the solvated alanine tetramer as a function of the 1−4 distance, r_{1-4}, and the dihedral angle, ψ. Panels a, b, and c present results for the AA, MS-CG, and iter-gYBG models, respectively. In panel a, the labels identify helix (H), intermediate (I), extended (E1/E2), and intermediate-extended (IE) regions of the configuration space.

E1, and E2 regions, albeit with incorrect propensities, but fails to significantly sample the H region. The iter-gYBG model (panel c) samples the FES with considerably greater accuracy, which is expected since it is explicitly parametrized to reproduce force correlation functions that are related to the AA distributions along r_{1-4} and ψ.

However, both the MS-CG and iter-gYBG models sample regions of the FES that were not sampled by the AA model, including extensions of the E2 region and also an intermediate/extended (IE) region. As discussed above, the MS-CG method determines potentials based upon the assumption that the chosen form of the CG potential is capable of reproducing the cross-correlations of the AA model. The IE region that is "forbidden" from the AA model demonstrates an instance that this assumption fails. This region corresponds to values for r_{1-4} that are sampled in extended conformations and values for ψ that are sampled in helical conformations. Because the atomic geometry of the peptide backbone precludes this combination of r_{1-4} and ψ, the AA model transitions from extended to helical conformations via the I region and not through the IE region. In contrast, the simple molecular mechanics potential and minimal peptide resolution of the CG model do not adequately couple these two degrees of freedom. Thus, in order to sample both the H and E2 regions of the FES, the CG model also samples the IE region. Consequently, the MS-CG model provides a relatively poor description of the 1-D distributions of the mapped AA ensemble because $\mathbf{G}^{AA}$ provides a relatively poor approximation to the cross-correlations generated by the CG model. Moreover, iterative methods, such as the iter-gYBG method, which quite accurately reproduce the 1-D distributions of the mapped ensemble, do so by distorting the cross-correlations of the mapped AA ensemble and, in particular, sampling the forbidden IE region of the FES. Figure S3 of the Supporting Information explicitly compares the cross-correlations generated by the AA and iter-gYBG models and demonstrates their effect on the resulting CG potentials.

Disordered Peptide Ensemble. We next considered whether a minimal resolution, bottom-up CG model can accurately describe the ensemble sampled by the OPLS-AA model for a capped 12-residue alanine peptide in explicit SPC/E solvent. This high-resolution model sampled a very heterogeneous ensemble that included helical, coil, and extended structures. For this high-resolution ensemble, we parametrized 12-site CG models via both the MS-CG and iter-gYBG methods. As for the 12-residue alanine peptide in vacuum, we distinguished between the interior sites and the two sites at each terminus of the peptide, while employing a total of 11 distinct pair potentials to model the interactions between sites that are separated by exactly 3, exactly 4, and more than 4 bonds.

Quite recently, Carmichael and Shell[30] also employed a bottom-up method to parametrize CG models from a simulation of an atomically detailed model for a 15-residue alanine peptide with implicit solvent. In particular, they parametrized CG models with 1-, 2-, and 3-sites per residue by minimizing the relative entropy[39,76] with respect to the atomically detailed ensemble. In order to make direct comparison with their study, Figure 5 presents simulated FES's in terms of the order parameters that they considered, i.e., the root-mean-square displacement (RMSD) from a perfect helical conformation and the radius of gyration, R_g.

DOI: 10.1021/ct5009922
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

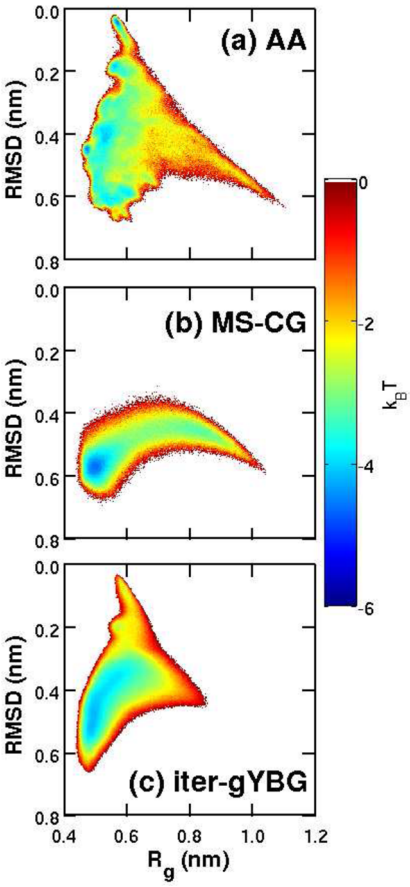


Figure 5. FES's for the solvated alanine 12-mer as a function of the deviation, RMSD, from a perfectly helical conformation and the radius of gyration, R_g. Panels a, b, and c present results for the AA, MS-CG, and iter-gYBG models, respectively.

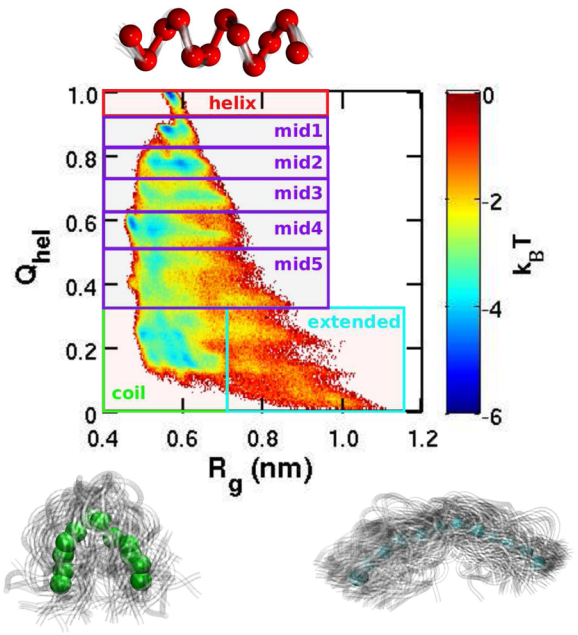


Figure 6. FES as a function of helicity, Q_{hel}, and radius of gyration, R_g, sampled by the AA model for the solvated alanine 12-mer. The rectangles indicate 8 partitions of this FES. The opaque configurations present the average structures sampled in the helix, coil, and extended regions of the FES. The superimposed transparent images present traces of the peptide backbone for representative conformations sampled in these regions.

Figure 5a demonstrates that our high-resolution peptide model, i.e., the OPLS-AA model in explicit SPC/E solvent, samples a rather diffuse FES with several shallow minima, including a metastable minima for a nearly perfect helical structure. Figure 5a suggests that this explicit solvent model samples a more complex ensemble and demonstrates greater helical tendency than the implicit solvent model considered by Carmichael and Shell. Figure 5b demonstrates that the MS-CG minimal model samples a range of collapsed and extended structures with little tendency for helical structures. Interestingly, this MS-CG model appears to sample a similar ensemble to the minimal model parametrized by Carmichael and Shell, although it is important to recognize that the MS-CG model employed a more complex potential function and was parametrized for a mapped AA ensemble with greater complexity. Finally, Figure 5c demonstrates that the iter-gYBG model samples an ensemble that is much more similar to the mapped AA ensemble. In particular, the iter-gYBG model samples helical conformations with much greater tendency than the MS-CG model, although with slightly less tendency than the AA model. Moreover, the iter-gYBG model clearly smooths over the fine structure of the AA FES and underestimates the stability of extended conformations.

In order to more closely compare the three models and, in particular, the helix−coil transitions that they sample, we next analyzed the ensembles as a function of Q_{hel} and R_g. Figure 6 presents the corresponding FES for the explicitly solvated AA model. The AA FES demonstrates basins that correspond to helical and coil conformations as well as a "tail" of extended conformations. Figure 6 presents the average structure sampled by the AA model in each of these regions of the FES, as well as representative structures that demonstrate the fluctuations sampled about these average structures. In addition, Figure 6 also reveals horizontal bands in the AA FES that correspond to metastable intermediates that form as the AA model transitions from coil conformations to fully helical conformations. We partitioned this FES into eight regions for subsequent analysis.

Panel a of Figure 7 presents the FES sampled by the MS-CG model for the solvated alanine 12-mer as a function of Q_{hel} and R_g. Figure 7a demonstrates that the MS-CG model samples coil and extended conformations but does not sample conformations with $Q_{hel} > 0.5$. We considered several possible causes for the discrepancies between the MS-CG and AA ensembles, including (1) the MS-CG method cannot accurately describe the helical structures that are sampled by the AA model for the solvated alanine 12-mer; (2) given the minimal peptide representation and simple molecular mechanics form, there does not exist a single potential that can sample the entire range of structures in the mapped AA ensemble; or (3) the minimal peptide representation and simple molecular mechanics potential cannot reproduce the cross-correlations of the mapped AA ensemble that characterize the transitions between different regions of the FES.

Accordingly, we employed the MS-CG method to determine CG models that were specific to the helical, coil, and extended regions of the FES. Panels b, c, and d of Figure 7 present the FESs sampled by these MS-CG models and demonstrate that they accurately sample the corresponding regions of configuration space. These results are consistent with the results in Figures 1 and 2 for MS-CG models of well-defined helices.

DOI: 10.1021/ct5009922
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

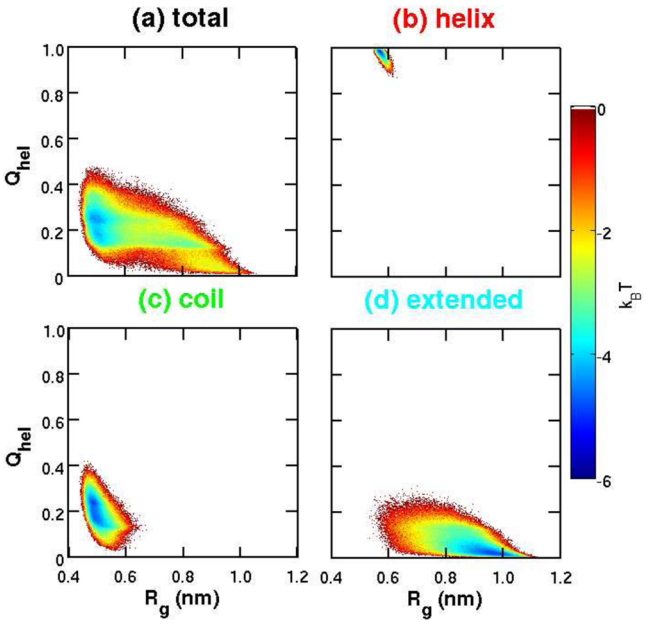


Figure 7. FES's as a function of Q_{hel} and R_g sampled by various CG models for the solvated alanine 12-mer. Panel a corresponds to the MS-CG model. Panels b, c, and d correspond to models parametrized via the MS-CG method for the helix, coil, and extended regions, respectively, of the FES.

Thus, minimal MS-CG models appear capable of reasonably reproducing the structure and cross-correlations within each region of configuration space that is characterized by a well-defined average structure.

Figure 8 presents the FES sampled by the iter-gYBG model for the solvated alanine 12-mer as a function of Q_{hel} and R_g. The iter-gYBG model reasonably reproduces the AA FES in Figure 6. In particular, the FES's for the iter-gYBG and AA models demonstrate similar bands in the helix−coil transition region of the FES, although the iter-gYBG FES does not demonstrate the same minima in this region. Moreover, the iter-gYBG model samples helical conformations with similar weight to the AA model. However, the iter-gYBG model overestimates the stability of the mid3, mid4, and mid5 regions of the FES, which correspond to the helix−coil transition, while underestimating the stability of the extended and coil regions of the FES. Nevertheless, Figure 8 suggests that a minimal model with a molecular mechanics potential is capable of sampling the stable conformations in the heterogeneous AA ensemble. Thus, the discrepancies between the AA and MS-CG ensembles appear to stem primarily from the inability of the MS-CG method to accurately describe the transitions between the different regions of conformational space.

As discussed above, the MS-CG method assumes that the CG model is capable of reproducing the cross-correlations $\mathbf{G}^{AA}$ in the mapped AA ensemble. In contrast, iterative methods, such as the iter-gYBG method, more accurately reproduce the mapped AA distributions along individual degrees of freedom by accounting for the limited ability of the CG model to reproduce the corresponding cross-correlations. Thus, by comparing the cross-correlations sampled by the AA and iter-gYBG models, we can identify specific features of the AA ensemble that cause discrepancies between the AA and MS-CG models. Our analysis of these cross-correlations indicated, perhaps unsurprisingly, that the minimal peptide representation

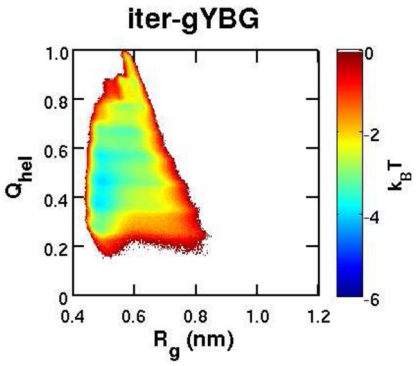


Figure 8. FES as a function of Q_{hel} and R_g sampled by the iter-gYBG model for the solvated alanine 12-mer.

and simple molecular mechanics potential are incapable of reproducing the cooperativity of the helix−coil transition in the AA model. In particular, we observed four interrelated manifestations of this limitation: (1) The CG model does not reproduce the complex correlations between the CG angle and the other intramolecular degrees of freedom that arise during this transition. (2) The AA and iter-gYBG models sample considerably different structural correlations in the transition regions of the FES (i.e., the mid3, mid4, and mid5 regions) in Figure 6. (3) The iter-gYBG model samples the helix region of the FES with reasonable weight but significantly oversamples the transition regions and undersamples the coil regions of the FES. (4) The iter-gYBG model samples helices that are slightly less precise than the helices of the AA model.

We hypothesized that the errors in the MS-CG model largely stem from these cross-correlations in the mapped AA ensemble that cannot be reproduced by the minimal CG model. We numerically tested this hypothesis by modifying the matrix, $\mathbf{G}^{AA}$, of cross-correlations in order to account for the limitations of the minimal peptide representation and simple molecular mechanics potential. To perform these modifications, we decomposed $\mathbf{G}^{AA}$ into distinct contributions from each region ν of the FES: $\mathbf{G}^{AA} = \sum_\nu w_\nu \mathbf{G}^{AA}_\nu$, where ν identifies a particular region of the FES in Figure 6, $\mathbf{G}^{AA}_\nu$ indicates the matrix of cross-correlations that is calculated from the configurations that map to region ν, and w_ν is the probability for the AA model to sample region ν. We then manually modified the matrix that describes the cross correlations that are present in the mapped ensemble, $\mathbf{G}^{AA}$, in order to determine a cross-correlation matrix, $\mathbf{G}^{AA\text{-}mod}$, that more accurately reflects the limitations of the minimal representation and molecular mechanics potential:

(1) Because the CG model appears incapable of reproducing the cross-correlations of the CG angle with the other intramolecular degrees of freedom, we treated the angle interactions independently of the remaining interactions. In particular, we eliminated from $\mathbf{G}^{AA}$ all indirect contributions that reflect cross-correlations involving the CG angle.

(2) Because the CG model appears incapable of reproducing the cross-correlations sampled by the AA model in the mid3, mid4, and mid5 regions of the FES that correspond to the transition between coil and helical conformations, we replaced the contributions to the $\mathbf{G}$ matrix from these regions by linearly interpolating between the mid2 and coil regions, i.e., $\mathbf{G}^{AA\text{-}mod}_\nu = \gamma_\nu \mathbf{G}^{AA}_{mid2} + (1 - \gamma_\nu)\mathbf{G}^{AA}_{coil}$ for ν = mid3, mid4, mid5, where γ_ν is determined by linearly interpolating the helicity Q_{hel} at the center of region ν.

DOI: 10.1021/ct5009922
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

(3) Because the CG model tends to overestimate the stability of the mid5 region and underestimate the stability of the coil region of the AA FES, we replaced the atomistic weights for these regions with weights more closely resembling those sampled by the iter-gYBG model (see Figure S11 in the Supporting Information).

(4) Because the CG model appears incapable of reproducing the precise structural features of the AA helix, we slightly smoothed the indirect contributions from the helical regions of the AA FES (i.e., the helix, mid 1, and mid 2 regions) to the blocks of the **G** matrix that describe the 1−4 interactions[77] (see Figure S12 in the Supporting Information). The Supporting Information (Figures S9−S12) describes these modifications in much greater detail.

We then determined a modified MS-CG model by solving the system of MS-CG normal equations for the potential parameters after replacing $\mathbf{G}^{AA}$ with the $\mathbf{G}^{AA\text{-mod}}$. In comparison to the matrix of cross-correlations, $\mathbf{G}^{AA}$, that is determined from the mapped AA ensemble, $\mathbf{G}^{AA\text{-mod}}$ provides a much more accurate description of the cross-correlations that can be generated by the CG model. Thus, if these structural features of the mapped AA ensemble are the major sources of error in the MS-CG model, the modified MS-CG model should provide a significantly improved description of the AA FES. Figure 9

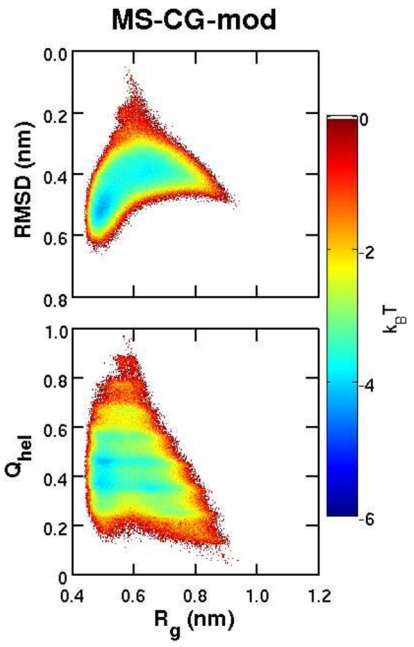


Figure 9. FESs sampled by the "modified" MS-CG model for the solvated alanine 12-mer. The top panel presents the FES as a function of RMSD and R_g; the bottom panel presents the FES as a function of Q_{hel} and R_g.

presents the FES's sampled by the modified MS-CG model as a function of both RMSD and R_g (top) and also Q_{hel} and R_g (bottom). Indeed, after making these modifications to **G**, the MS-CG equations determine a model that describes the AA ensemble with much greater accuracy than the original MS-CG model. This strongly supports our hypothesis regarding the errors in the MS-CG model.

DISCUSSION

The present work investigates the potential and limitations of bottom-up coarse-graining methods for accurately modeling the ensembles of structures that are sampled by high-resolution peptide models. In particular, we considered several high resolution peptide models that sampled ensembles with varying degrees of complexity and helicity. For each high resolution ensemble, we employed the MS-CG method to parametrize an implicit solvent, minimal resolution CG model that represented each residue with a single site and incorporated solvent effects into the interactions between the peptide sites. In contrast to other bottom-up methods, the MS-CG method does not require simulations with multiple CG models in order to parametrize the CG potential. Rather, it employs the g-YBG relation to directly determine the CG potential from the high-resolution ensemble. However, this calculation effectively assumes that the CG model will reproduce the cross-correlations that are present in the mapped AA ensemble. Since these cross-correlations may be essential for describing hierarchical protein structures, this study assessed the validity of the MS-CG assumption in the particular case that the CG model employs a minimal peptide representation and a simple molecular mechanics potential.

The MS-CG method determined accurate minimal models for two different high-resolution peptide ensembles that fluctuated about well-defined helices. The MS-CG model very accurately reproduced the ensemble of floppy helices sampled by a simple Gō model. The MS-CG model also very accurately reproduced the ensemble of precise helices sampled by an AA model for an alanine 12-mer in vacuum. In this latter case, the MS-CG model required a more complex potential in order to distinguish the tendencies of the terminal and interior residues. In particular, this model employed distinct angle, dihedral, and pair potentials to describe interactions involving the two terminal residues on either side of the peptide. One generally expects that, given a fixed CG representation, increasingly complex potentials may be required to reproduce increasingly complex ensembles.

As a third high-resolution peptide model, we considered an explicitly solvated AA model for the alanine 12-mer. The solvated high-resolution model sampled a much more complex ensemble that included helical, coil, and extended structures, as well as various intermediate structures. Interestingly, when we extracted subensembles with well-defined average structures, the MS-CG method proved capable of determining models that sampled helical, coil, or extended regions of conformational space. However, given the minimal representation and the simple molecular mechanics potential, the MS-CG method did not succeed in determining a single potential for sampling the entire conformational space with appropriate weight. The pioneering studies of Voth and co-workers[33−35] demonstrated similar challenges for determining transferable MS-CG models that accurately described both helical and turn conformations with a slightly higher resolution peptide model and an explicit CG solvent model.

It is interesting to compare these results with the recent studies of Carmichael and Shell,[30] who employed a clever iterative nonlinear optimization method to parametrize implicit solvent CG peptide models that minimized the relative entropy at various levels of resolution. Figure 5 suggests that the present study considered a high-resolution ensemble with somewhat greater complexity than the high-resolution model considered

DOI: 10.1021/ct5009922
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

by Carmichael and Shell. Interestingly, though, Figure 5 also suggests that the MS-CG model, which we directly determined from the high-resolution ensemble, appears to sample the same regions of conformational space as the corresponding minimal model that was obtained by iteratively minimizing the relative entropy. This minimal MS-CG model does not adequately sample helical conformations and, instead, too frequently samples collapsed coil structures.

We employed an iterative g-YBG (iter-gYBG) method[50−52] to parametrize an implicit solvent, minimal CG model that more accurately reproduced the AA ensemble for the explicitly solvated alanine 12-mer. The iter-gYBG method is quite similar to other iterative bottom-up methods, since it employs multiple CG simulations in order to determine potentials that reproduce target 1-D correlation functions of the mapped AA ensemble. The resulting iter-gYBG model reasonably reproduces the 1-D distributions of the mapped AA ensemble, but does so at the expense of distorting the cross-correlations between these degrees of freedom.

Despite these distortions, the iter-gYBG model samples the various regions of the FES with much greater accuracy than the MS-CG model. The similarity of the AA and iter-gYBG ensembles is perhaps somewhat surprising, since the iter-gYBG model is parametrized to reproduce a set of 1-D correlation functions that reflect weighted averages over the various conformations in the heterogeneous AA ensemble. Since different ensembles might possibly give rise to very similar 1-D correlation functions, one might expect that iterative structure-based CG methods may be quite insensitive to the underlying distribution of conformations. In fact, the distributions of the mapped AA ensemble appear quite similar to the distributions sampled by intermediate states in the helix−coil transition and, indeed, the iter-gYBG model demonstrates too great a tendency for sampling this region of conformational space. Nevertheless, the iter-gYBG model provides a reasonably accurate description of the entire AA FES. Thus, while it may sometimes prove useful to employ different potentials for sampling different types of helices,[78] this study provides evidence that bottom-up structure-based methods can determine a single potential that reasonably samples the complex ensemble sampled by an AA peptide model.

By comparing the ensembles sampled by the AA and iter-gYBG peptide models, we identified the cross-correlations that cannot be reproduced by the CG model and, thus, invalidate the fundamental MS-CG assumption. This analysis suggested that the errors in the MS-CG model for the solvated alanine 12-mer stem primarily from four particular manifestations of the cooperative interactions in the AA model: (1) The minimal representation and simple molecular mechanics potential appear incapable of reproducing the cross-correlations that involve the CG angle. In fact, the MS-CG method determines a more accurate peptide model if these cross-correlations are systematically neglected, similar to the "hybrid force-matching" method of Rühle and Junghans.[79] (2) The CG model appears incapable of reproducing the precise cross-correlations that the AA model samples during the helix−coil transition, as indicated by our calculations for the solvated alanine tetramer. (3) The CG potentials that are determined via the g-YBG equation appear to systematically overestimate the stability of transition structures that arise during the helix−coil transition. (4) Minimal models may be incapable of reproducing the precise cross-correlations that result from the cooperative interactions that stabilize helices in AA models. Significantly, we explicitly demonstrated that these four considerations significantly limit the accuracy of the MS-CG model. By implementing corresponding modifications to the MS-CG cross correlation matrix, $\mathbf{G}^{\mathrm{AA}}$, the modified MS-CG model described the mapped AA ensemble much more accurately.

The cooperative interactions that stabilize helices in the AA model generate sharp features in the $\mathbf{G}^{\mathrm{AA}}$ cross-correlation matrix. Since the simple molecular mechanics potential and minimal CG representation do not provide sufficient coupling to reproduce these precise cross-correlations, the sharp features in $\mathbf{G}^{\mathrm{AA}}$ determine relatively weak MS-CG forces, $\boldsymbol{\phi}^0 = (\mathbf{G}^{\mathrm{AA}})^{-1}\boldsymbol{b}^{\mathrm{AA}}$, that do not adequately stabilize helical structures. This suggests an amusing "reciprocal" relation for the MS-CG method: since $\boldsymbol{b}^{\mathrm{AA}}$ describes the structural features of the AA helix, it follows that weakening the corresponding features of $\mathbf{G}^{\mathrm{AA}}$ leads to stronger MS-CG forces $\boldsymbol{\phi}^0$ that better stabilize helices. Thus, given the limitations of the minimal resolution and the molecular mechanics potential, the iter-gYBG model appears to stabilize helices by replacing the relatively weak but cooperative interactions of the AA model with stronger, but less cooperative interactions.

In summary, the present work demonstrates both the promise and limitations of bottom-up structure-based methods for developing minimal CG models of helical and disordered peptide ensembles. Notably, the present work does not resolve the outstanding challenge of developing a single transferable potential that accurately models the structures of multiple proteins. Indeed, while previous studies have developed transferable CG protein models via top-down or knowledge-based approaches,[29,80−83] relatively few studies have attempted to develop transferable protein models from high-resolution simulations.[36,84−87] We have previously proposed a rigorous method for combining statistics from an "extended ensemble" in order to determine a single transferable CG potential that accurately describes the free energy surface for multiple systems.[88,89] However, the success of such an approach will likely depend upon identifying an appropriate choice of CG representation and potential functions. This remains beyond the scope of the present study, but certainly motivates future work.

More generally, our calculations suggest that minimal peptide models with molecular mechanics potentials may not accurately describe the helix−coil transition that is sampled by atomically detailed models. Although the iter-gYBG model for the alanine tetramer accurately reproduces the AA distributions along each relevant degree of freedom and samples both helical and extended conformations with reasonable weight, Figure 4 demonstrates that this model can sample the helix−coil transition via a mechanism that is excluded by the atomic structure of the peptide backbone. Similarly, although the iter-gYBG and AA models for the solvated alanine 12-mer sample conformations of similar size (i.e., R_g) when transitioning between helical and coil structures, the actual transition structures demonstrate considerably different cross-correlations in the two models. It is possible that alternative parametrization methods may determine minimal models that more accurately describe this transition. However, it seems more likely that the minimal representation and the simple molecular mechanics potential may be fundamentally incapable of reproducing the cooperative interactions that result from an atomically detailed peptide backbone and explicit hydrogen bonding interactions.[31,90] As noted above, it appears that such CG models will

DOI: 10.1021/ct5009922
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

tend to stabilize protein structures by replacing weak, cooperative atomic interactions with strong, noncooperative coarse interactions. These considerations echo previous conclusions that cooperative interactions are essential for protein folding and thermodynamics.[91−93] Moreover, they strongly motivate the further development of CG potentials that either mimic the cooperativity of hydrophobic collapse[94−96] or that couple multiple interactions in order to preserve a more accurate description of the peptide backbone geometry[7] and hydrogen bonding.[97,98]

Finally, this work further demonstrated the utility of the iter-gYBG method for modeling complex molecular systems. As in previous work,[50] we employed a heuristic approach for robustly treating intramolecular interactions. As documented in the Supporting Information, although the iter-gYBG method determines reasonably accurate potentials after relatively few iterations, it demonstrates suboptimal convergence properties and often diverges away from, or oscillates about, these accurate potentials. This clearly motivates subsequent work to address limitations of the iter-gYBG method. Nevertheless, the iter-gYBG method provides a convenient mechanism for (1) efficiently sampling the space of CG models, (2) determining reasonably accurate models for the structure of complex molecular systems, (3) identifying structural features that limit the accuracy of the MS-CG method, and (4) revealing necessary improvements to the CG representation and interaction set.

CONCLUSION

This work investigated the potential and limitations of bottom-up methods for developing minimal, implicit solvent peptide models from AA simulations. The MS-CG method determines models that quite accurately reproduce the ensembles generated by high-resolution models for flexible helices, rigid helices, coils, and extended structures. However, when the AA models sampled multiple distinct conformations, the MS-CG models did not correctly weight the various regions of configuration space. We demonstrated that these discrepancies in the MS-CG models for disordered peptides primarily result from the inability of the minimal peptide resolution and simple molecular mechanics potential to reproduce the complex cross-correlations that arise in the cooperative helix−coil transition of the AA model. Consequently, the MS-CG model can be improved by smoothing over the cross-correlations in the mapped AA ensemble, e.g., by neglecting cross-correlations involving the angle degree of freedom or by blurring the cross-correlations that arise in precise helical conformations.

We also demonstrated that the iter-gYBG method reasonably reproduces the 1-D distributions of the mapped AA ensemble for the disordered peptide. Moreover, although it clearly averages over the fine details of the AA FES and overestimates the stability of intermediate transition structures, the iter-gYBG method determines a CG potential that reasonably reproduces the FES of the AA model. This is somewhat surprising since the 1-D distributions may provide relatively little information regarding either the distribution of conformations in the ensemble or the global structure of the peptide.

More generally, these calculations demonstrate that CG models with minimal resolution and simple molecular mechanics potentials are unlikely to provide an accurate description of the cooperative helix−coil transition that occurs in AA models. In particular, minimal CG models appear to stabilize helical conformations by replacing the weak, but cooperative interactions of the AA model with stronger, but less cooperative, interactions. These results provide insight into the potential of bottom-up methods for accurately modeling biomolecular structure and motivate future investigations of the relationship between the CG representation, the complexity of the CG potential, and the accuracy of the CG model.

ASSOCIATED CONTENT

Supporting Information

Detailed description of all methods, simulations, and calculations employed in this work, as well as additional analysis of these calculations. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: wnoid@chem.psu.edu.

Funding

The present work has been financially supported by an NSF CAREER award (Grant No. MCB 1053970) from the National Science Foundation. J.F.R. was also funded by the Penn State Academic Computing Fellowship. Numerical calculations were performed using support and resources from Research Computing and Cyberinfrastructure, a unit of Information Technology Services at Penn State. Figure 6 employed VMD. VMD is developed with NIH support by the Theoretical and Computational Biophysics group at the Beckman Institute, University of Illinois at Urbana−Champaign.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The intensity plots in Figures S3a, S4b, S9a, S10a, and S12 in the Supporting Information were generated using gnuplot 4.4.0.[99]

REFERENCES

(1) Schlick, T.; Collepardo-Guevara, R.; Halvorsen, L. A.; Jung, S.; Xiao, X. Biomolecular modeling and simulation: a field coming of age. *Q. Rev. Biophys.* **2011**, *44*, 191−228.

(2) Morriss-Andrews, A.; Shea, J.-E. Simulations of Protein Aggregation: Insights from Atomistic and Coarse-grained Models. *J. Phys. Chem. Lett.* **2014**, *5*, 1899−1908.

(3) Hyeon, C.; Thirumalai, D. Capturing the essence of folding and functions of biomolecules using coarse-grained models. *Nat. Commun.* **2011**, *2*, 487.

(4) Riniker, S.; Allison, J. R.; van Gunsteren, W. F. On developing coarse-grained models for biomolecular simulation: a review. *Phys. Chem. Chem. Phys.* **2012**, *14*, 12423−12430.

(5) Noid, W. G. Systematic methods for structurally consistent coarse-grained models. *Methods Mol. Biol.* **2013**, *924*, 487−531.

(6) Potoyan, D. A.; Savelyev, A.; Papoian, G. A. Recent successes in coarse-grained modeling of DNA. *WIREs Comput. Mol. Sci.* **2013**, *3*, 69−83.

(7) Tozzini, V. Coarse-grained models for proteins. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144−150.

(8) Tozzini, V.; Rocchia, W.; McCammon, J. Mapping all-atom models onto one-bead coarse-grained models: General properties and applications to a minimal polypeptide model. *J. Chem. Theory Comput.* **2006**, *2*, 667−673.

(9) Tozzini, V. Minimalist models for proteins: a comparative analysis. *Q. Rev. Biophys.* **2010**, *43*, 333−371.

(10) Purisima, E.; Scheraga, H. Conversion from a Virtual-bond Chain to a Complete Polypeptide Backbone Chain. *Biopolymers* **1984**, *23*, 1207−1224.

DOI: 10.1021/ct5009922
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

(11) Rey, A.; Skolnick, J. Efficient Algorithm for the Reconstruction of a Protein Backbone from the Alpha-carbon Coordinates. *J. Comput. Chem.* **1992**, *13*, 443−456.
(12) Honeycutt, J. D.; Thirumalai, D. Metastability of the folded states of globular proteins. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 3526−3529.
(13) Honeycutt, J. D.; Thirumalai, D. The Nature of Folded States of Globular Proteins. *Biopolymers* **1992**, *32*, 695−709.
(14) Taketomi, H.; Ueda, Y.; Go, N. Studies on Protein Folding, Unfolding and Fluctuations by Computer-Simulation. 1. Effect of Specific Amino-Acid Sequence Represented by Specific Inter-Unit Interactions. *Int. J. Pept. Protein Res.* **1975**, *7*, 445−459.
(15) Go, N. Theoretical-Studies of Protein Folding. *Annu. Rev. Biophys. Biol.* **1983**, *12*, 183−210.
(16) Nymeyer, H.; Garcia, A.; Onuchic, J. Folding funnels and frustration in off-lattice minimalist protein landscapes. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 5921−5928.
(17) Tirion, M. M. Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Phys. Rev. Lett.* **1996**, *77*, 1905−1908.
(18) Bahar, I.; Atilgan, A. R.; Erman, B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.* **1997**, *2*, 173−181.
(19) Hills, R. D.; Brooks, C. L. Insights from Coarse-Grained Go Models for Protein Folding and Dynamics. *Int. J. Mol. Sci.* **2009**, *10*, 889−905.
(20) Whitford, P. C.; Sanbonmatsu, K. Y.; Onuchic, J. N. Biomolecular dynamics: order-disorder transitions and energy landscapes. *Rep. Prog. Phys.* **2012**, *75*, 076601.
(21) Friedel, M.; Baumketner, A.; Shea, J.-E. Effects of surface tethering on protein folding mechanisms. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 8396−8401.
(22) Enciso, M.; Rey, A. Simple model for the simulation of peptide folding and aggregation with different sequences. *J. Chem. Phys.* **2012**, *136*, 215103.
(23) Enciso, M.; Schutte, C.; Delle Site, L. A pH-dependent coarse-grained model for peptides. *Soft Matter* **2013**, *9*, 6118−6127.
(24) Ghavami, A.; van der Giessen, E.; Onck, P. R. Coarse-Grained Potentials for Local Interactions in Unfolded Proteins. *J. Chem. Theory Comput.* **2013**, *9*, 432−440.
(25) Nielsen, S. O.; Lopez, C. F.; Srinivas, G.; Klein, M. L. Coarse grain models and the computer simulation of soft materials. *J. Phys.: Condens. Matter* **2004**, *16*, R481.
(26) Muller, M.; Katsov, K.; Schick, M. Biological and synthetic membranes: What can be learned from a coarse-grained description? *Phys. Rep.* **2006**, *434*, 113−176.
(27) Villa, A.; van der Vegt, N. F. A.; Peter, C. Self-assembling dipeptides: including solvent degrees of freedom in a coarse-grained model. *Phys. Chem. Chem. Phys.* **2009**, *11*, 2068−2076.
(28) Villa, A.; Peter, C.; van der Vegt, N. F. A. Self-assembling dipeptides: conformational sampling in solvent-free coarse-grained simulation. *Phys. Chem. Chem. Phys.* **2009**, *11*, 2077−2086.
(29) Bereau, T.; Deserno, M. Generic coarse-grained model for protein folding and aggregation. *J. Chem. Phys.* **2009**, *130*, 235106.
(30) Carmichael, S. P.; Shell, M. S. A new multiscale algorithm and its application to coarse-grained peptide models for self-assembly. *J. Phys. Chem. B* **2012**, *116*, 8383−8393.
(31) Bezkorovaynaya, O.; Lukyanov, A.; Kremer, K.; Peter, C. Multiscale simulation of small peptides: Consistent conformational sampling in atomistic and coarse-grained models. *J. Comput. Chem.* **2012**, *33*, 937−949.
(32) Ni, B.; Baumketner, A. Reduced atomic pair-interaction design (RAPID) model for simulations of proteins. *J. Chem. Phys.* **2013**, *138*, 064102.
(33) Zhou, J.; Thorpe, I. F.; Izvekov, S.; Voth, G. A. Coarse-grained peptide modeling using a systematic multiscale approach. *Biophys. J.* **2007**, *92*, 4289−4303.
(34) Thorpe, I. F.; Zhou, J.; Voth, G. A. Peptide folding using multiscale coarse-grained models. *J. Phys. Chem. B* **2008**, *112*, 13079−13090.
(35) Thorpe, I. F.; Goldenberg, D. P.; Voth, G. A. Exploration of transferability in multiscale coarse-grained peptide models. *J. Phys. Chem. B* **2011**, *115*, 11911−26.
(36) Hills, R. D.; Lu, L. Y.; Voth, G. A. Multiscale Coarse-Graining of the Protein Energy Landscape. *PLoS Comput. Biol.* **2010**, *6*, e1000827.
(37) Müller-Plathe, F. Coarse-graining in polymer simulation: From the atomistic to the mesoscopic scale and back. *ChemPhysChem* **2002**, *3*, 754−769.
(38) Lyubartsev, A. P.; Laaksonen, A. Calculation of effective interaction potentials from radial distribution functions: A reverse Monte Carlo approach. *Phys. Rev. E* **1995**, *52*, 3730−3737.
(39) Shell, M. S. The relative entropy is fundamental to multiscale and inverse thermodynamic problems. *J. Chem. Phys.* **2008**, *129*, 144108.
(40) Savelyev, A.; Papoian, G. A. Molecular renormalization group coarse-graining of electrolyte solutions: Applications to aqueous NaCl and KCl. *J. Phys. Chem. B* **2009**, *113*, 7785−7793.
(41) Izvekov, S.; Voth, G. A. A multiscale coarse-graining method for biomolecular systems. *J. Phys. Chem. B* **2005**, *109*, 2469−2473.
(42) Izvekov, S.; Voth, G. A. Multiscale coarse graining of liquid-state systems. *J. Chem. Phys.* **2005**, *123*, 134105.
(43) Noid, W. G.; Chu, J.-W.; Ayton, G. S.; Krishna, V.; Izvekov, S.; Voth, G. A.; Das, A.; Andersen, H. C. The Multiscale Coarse-graining Method. I. A Rigorous Bridge between Atomistic and Coarse-grained Models. *J. Chem. Phys.* **2008**, *128*, 244114.
(44) Noid, W. G.; Liu, P.; Wang, Y. T.; Chu, J.-W.; Ayton, G. S.; Izvekov, S.; Andersen, H. C.; Voth, G. A. The Multiscale Coarse-graining Method. II. Numerical implementation for molecular coarse-grained models. *J. Chem. Phys.* **2008**, *128*, 244115.
(45) Noid, W. G.; Chu, J.-W.; Ayton, G. S.; Voth, G. A. Multiscale Coarse-graining and Structural Correlations: Connections to Liquid State Theory. *J. Phys. Chem. B* **2007**, *111*, 4116−4127.
(46) Mullinax, J. W.; Noid, W. G. A generalized Yvon-Born-Green theory for molecular systems. *Phys. Rev. Lett.* **2009**, *103*, 198104.
(47) Mullinax, J. W.; Noid, W. G. A Generalized Yvon-Born-Green Theory for Determining Coarse-grained Interaction Potentials. *J. Phys. Chem. C* **2010**, *114*, 5661−5674.
(48) Rudzinski, J. F.; Noid, W. G. The role of many-body correlations in determining potentials for coarse-grained models of equilibrium structure. *J. Phys. Chem. B* **2012**, *116*, 8621−35.
(49) Das, A.; Lu, L.; Andersen, H. C.; Voth, G. A. The multiscale coarse-graining method. X. Improved algorithms for constructing coarse-grained potentials for molecular systems. *J. Chem. Phys.* **2012**, *136*, 194115.
(50) Rudzinski, J. F.; Noid, W. G. Investigation of Coarse-grained Mappings via an Iterative Generalized Yvon-Born-Green Method. *J. Phys. Chem. B* **2014**, *118*, 8295−8312.
(51) Cho, H. M.; Chu, J. W. Inversion of radial distribution functions to pair forces by solving the Yvon-Born-Green equation iteratively. *J. Chem. Phys.* **2009**, *131*, 134107.
(52) Lu, L.; Dama, J. F.; Voth, G. A. Fitting coarse-grained distribution functions through an iterative force-matching method. *J. Chem. Phys.* **2013**, *139*, 121906.
(53) Rudzinski, J. F.; Noid, W. G. Coarse-graining entropy, forces, and structures. *J. Chem. Phys.* **2011**, *135*, 214101.
(54) Ellis, C. R.; Rudzinski, J. F.; Noid, W. G. generalized-Yvon-Born-Green model for toluene. *Macromol. Theory Sim.* **2011**, *20*, 478−495.
(55) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435−447.
(56) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684−3690.
(57) Nose, S. A molecular dynamics method for simulations in the canonical ensemble. *Mol. Phys.* **1984**, *52*, 255−268.

DOI: 10.1021/ct5009922
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

(58) Hoover, W. G. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A* **1985**, *31*, 1695−1697.
(59) Parrinello, M.; Rahman, A. Strain fluctuations and elastic constants. *J. Chem. Phys.* **1982**, *76*, 2662−2666.
(60) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: AnN log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *99*, 8345−8348.
(61) Allen, M. P.; Tildesley, D. P. *Computer Simulation of Liquids*; Oxford Press: New York, 1987.
(62) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and testing of the OPLS All-Atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.
(63) Berendsen, H.; Grigera, J.; Straatsma, T. The Missing Term in Effective Pair Potentials. *J. Phys. Chem.* **1987**, *91*, 6269−6271.
(64) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraajie, J. G. E. M. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **1997**, *18*, 1463−1472.
(65) Archambault, J.; Pan, G.; Dahmus, G.; Cartier, M.; Marshall, N.; Zhang, S.; Dahmus, M.; Greenblatt, J. FCP1, the RAP74-interacting subunit of a human protein phosphatase that dephosphorylates the carboxyl-terminal domain of RNA polymerase IIO. *J. Biol. Chem.* **1998**, *273*, 27593−27601.
(66) Nguyen, B.; Abbott, K.; Potempa, K.; Kobort, M.; Archambault, J.; Greenblatt, J.; Legault, P.; Omichinski, J. NMR structure of a complex containing the TFIIF subunit RAP74 and the RNA polymerase II carboxyl-terminal domain phosphatase FCP1. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *18*, 5688−5693.
(67) Kamada, K.; Roeder, R.; Burley, S. Molecular mechanism of recruitment of TFIIF-associating RNA polymerase C-terminal domain phosphatase (FCP1) by transcription factor IIF. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 2296−2299.
(68) Kumar, S.; Showalter, S. A.; Noid, W. G. Native-Based Simulations of the Binding Interaction Between RAP74 and the Disordered FCP1 Peptide. *J. Phys. Chem. B* **2013**, *117*, 3074−3085.
(69) Noel, J. K.; Whitford, P. C.; Sanbonmatsu, K. Y.; Onuchic, J. N. SMOG@ctbp: simplified deployment of structure-based models in GROMACS. *Nucleic Acids Res.* **2010**, *38*, W657−W661.
(70) Lawrence, C. W.; Bonny, A.; Showalter, S. A. The disordered C-terminus of the RNA Polymerase II phosphatase FCP1 is partially helical in the unbound state. *Biochem. Biophys. Res. Commun.* **2011**, *410*, 461−465.
(71) Camilloni, C.; de Simone, A.; Vranken, W. F.; Vendruscolo, M. Determination of Secondary Structure Populations in Disordered States of Proteins Using Nuclear Magnetic Resonance Chemical Shifts. *Biochemistry* **2012**, *51*, 2224−2231.
(72) Mullinax, J. W.; Noid, W. G. Reference state for the generalized Yvon-Born-Green theory: Application for a coarse-grained model of hydrophobic hydration. *J. Chem. Phys.* **2010**, *133*, 124107.
(73) Anderson, E.; Bai, Z.; Bischof, C.; Blackford, S.; Demmel, J.; Dongarra, J.; Croz, J. D.; Greenbaum, A.; Hammarling, S.; McKenney, A. et al. *LAPACK Users'Guide*; SIAM: Philadelphia, PA, 1999.
(74) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in FORTRAN: The Art of Scientific Computing*; Cambridge University Press: New York, 1992.
(75) Savelyev, A.; Papoian, G. A. Molecular renormalization group coarse-graining of polymer chains: Applications to double-stranded DNA. *Biophys. J.* **2009**, *96*, 4044−4052.
(76) Chaimovich, A.; Shell, M. S. Coarse-graining errors and numerical optimization using a relative entropy framework. *J. Chem. Phys.* **2011**, *134*, 094112.
(77) Shapiro, L.; Stockman, G. *Computer Vision*; Prentice Hall: Upper Saddle River, NJ, 2001.
(78) Lia, G.; Spampinato, B.; Maccari, G.; Tozzini, V. Minimalist Model for the Dynamics of Helical Polypeptides: A Statistic-Based Parametrization. *J. Chem. Theory Comput.* **2014**, *10*, 3885−3895.
(79) Ruhle, V.; Junghans, C. Hybrid Approaches to Coarse-Graining using the VOTCA Package: Liquid Hexane. *Macromol. Theory Sim.* **2011**, *20*, 472−477.
(80) Cheon, M.; Chang, I.; Hall, C. K. Extending the PRIME model for protein aggregation to all 20 amino acids. *Proteins* **2010**, *78*, 2950−2960.
(81) Chebaro, Y.; Pasquali, S.; Derreumaux, P. The coarse-grained OPEP force field for non-amyloid and amyloid proteins. *J. Phys. Chem. B* **2012**, *116*, 8741−8752.
(82) de Jong, D. H.; Singh, G.; Bennett, W. F. D.; Arnarez, C.; Wassenaar, T. A.; Schäfer, L. V.; Periole, X.; Tieleman, D. P.; Marrink, S. J. Improved Parameters for the Martini Coarse-Grained Protein Force Field. *J. Chem. Theory Comput.* **2013**, *9*, 687−697.
(83) Kapoor, A.; Travesset, A. Folding and stability of helical bundle proteins from coarse-grained models. *Proteins: Struct., Funct., Bioinf.* **2013**, *81*, 1200−1211.
(84) Betancourt, M. R.; Omovie, S. J. Pairwise energies for polypeptide coarse-grained models derived from atomic force fields. *J. Chem. Phys.* **2009**, *130*, 195103.
(85) Gopal, S. M.; Mukherjee, S.; Cheng, Y.-M.; Feig, M. PRIMO/PRIMONA: A coarse-grained model for proteins and nucleic acids that preserves near-atomistic accuracy. *Proteins* **2010**, *78*, 1266−1281.
(86) Engin, O.; Villa, A.; Peter, C.; Sayar, M. A Challenge for Peptide Coarse Graining: Transferability of Fragment-Based Models. *Macromol. Theory Sim.* **2011**, *20*, 451−465.
(87) Andrews, C. T.; Elcock, A. H. COFFDROP: A Coarse-Grained Nonbonded Force Field for Proteins Derived from All-Atom Explicit-Solvent Molecular Dynamics Simulations of Amino Acids. *J. Chem. Theory Comput.* **2014**, *10*, 5178−5194.
(88) Mullinax, J. W.; Noid, W. G. Extended Ensemble approach for deriving transferable Coarse-grained potentials. *J. Chem. Phys.* **2009**, *131*, 104110.
(89) Mullinax, J. W.; Noid, W. G. Recovering physical potentials from a model protein databank. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 19867−19872.
(90) Larini, L.; Shea, J.-E. Coarse-Grained Modeling of Simple Molecules at Different Resolutions in the Absence of Good Sampling. *J. Phys. Chem. B* **2012**, *116*, 8337−8349.
(91) Kolinski, A.; Galazka, W.; Skolnick, J. On the origin of the cooperativity of protein folding: Implications from model simulations. *Proteins: Struct., Funct., Bioinf.* **1996**, *26*, 271−287.
(92) Klimov, D. K.; Thirumalai, D. Cooperativity in protein folding: From lattice models with sidechains to real proteins. *Fold. Des.* **1998**, *9*, 127−139.
(93) Kaya, H.; Chan, H. S. Polymer principles of protein calorimetric two-state cooperativity. *Proteins: Struct., Funct., Bioinf.* **2000**, *40*, 637−661.
(94) Takada, S.; Luthey-Schulten, Z.; Wolynes, P. G. Folding dynamics with nonadditive forces: A simulation study of a designed helical protein and a random heteropolymer. *J. Chem. Phys.* **1999**, *110*, 11616−11629.
(95) Makowski, M.; Liwo, A.; Scheraga, H. A. Simple physics-based analytical formulas for the potentials of mean force for the interaction of amino acid side chains in water. 1. Approximate expression for the free energy of hydrophobic association based on a Gaussian-overlap model. *J. Phys. Chem. B* **2007**, *111*, 2910−2916.
(96) Badasyan, A.; Liu, Z.; Chan, H. S. Probing Possible Downhill Folding: Native Contact Topology Likely Places a Significant Constraint on the Folding Cooperativity of Proteins with similar to 40 Residues. *J. Mol. Biol.* **2008**, *384*, 512−530.
(97) Ding, F.; Buldyrev, S. V.; Dokholyan, N. V. Folding Trp-cage to NMR resolution native structure using a coarse-grained protein model. *Biophys. J.* **2005**, *88*, 147−155.
(98) Enciso, M.; Rey, A. A refined hydrogen bond potential for flexible protein models. *J. Chem. Phys.* **2010**, *132*, 235102.
(99) Williams, T.; Kelley, C. et al. *Gnuplot 4.4: An Interactive Plotting Program*; http://gnuplot.sourceforge.net/, 2010.

DOI: 10.1021/ct5009922
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX