The Journal of Chemical Physics

Perspective: Coarse-grained models for biomolecular systems

Cite as: J. Chem. Phys. **139**, 090901 (2013); https://doi.org/10.1063/1.4818908
Submitted: 04 May 2013 . Accepted: 22 July 2013 . Published Online: 03 September 2013

W. G. Noid

Perspective: Coarse-grained models for biomolecular systems

W. G. Noid[a)]
Department of Chemistry, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

(Received 4 May 2013; accepted 22 July 2013; published online 3 September 2013)

By focusing on essential features, while averaging over less important details, coarse-grained (CG) models provide significant computational and conceptual advantages with respect to more detailed models. Consequently, despite dramatic advances in computational methodologies and resources, CG models enjoy surging popularity and are becoming increasingly equal partners to atomically detailed models. This perspective surveys the rapidly developing landscape of CG models for biomolecular systems. In particular, this review seeks to provide a balanced, coherent, and unified presentation of several distinct approaches for developing CG models, including top-down, network-based, native-centric, knowledge-based, and bottom-up modeling strategies. The review summarizes their basic philosophies, theoretical foundations, typical applications, and recent developments. Additionally, the review identifies fundamental inter-relationships among the diverse approaches and discusses outstanding challenges in the field. When carefully applied and assessed, current CG models provide highly efficient means for investigating the biological consequences of basic physicochemical principles. Moreover, rigorous bottom-up approaches hold great promise for further improving the accuracy and scope of CG models for biomolecular systems. © *2013 AIP Publishing LLC*. [http://dx.doi.org/10.1063/1.4818908]

I. INTRODUCTION

The most powerful theoretical models provide not only quantitatively accurate predictions, but also fundamental and transparent insight.[1] In this sense, models that focus on essential features, while "coarse-graining" over extraneous details, provide the foundation for most, if not all, scientific efforts.[2] In recent years, though, coarse-grained (CG) models have become especially associated with a class of particle-based computational models for soft materials.[3] These models represent molecular systems by interaction "sites" (or "superatoms") that correspond to groups of atoms. By eliminating (or, more formally, integrating out) atomistic details that are considered "unnecessary," these models may provide three or more orders of magnitude greater efficiency than atomically detailed models.[4] Because human imagination and computational demands both significantly outpace Moore's law, the popularity of CG models continues to grow rapidly, despite tremendous recent advances in simulation methods, atomically detailed models, and computational resources.[5]

The objective of this perspective is to survey the current landscape of CG models for biomolecular systems, to highlight recent progress, and to identify outstanding challenges in the field. Because CG models have been applied for investigating a staggering range of physical phenomena and a vast array of biomolecular systems, this review will necessarily be of limited scope. This review will focus on particle-based, off-lattice CG models of systems that are at or near equilibrium. However, it will neglect a very large body of important work on many aspects of multiscale modeling, including hybrid resolution models, advanced sampling techniques, and CG dynamics. Moreover, this review will reflect the author's personal perspective in developing and applying statistical mechanical formalism for CG modeling. Nevertheless, despite these limitations, it is hoped that this perspective will prove useful to researchers, both as an introduction and survey of the field. In particular, we hope that this work will be distinguished from many excellent recent reviews[3,6–19] by its scope, (relative) brevity, and framework for rigorously unifying several rather diverse approaches, as well as by our perspective in identifying outstanding challenges in the field.

After the introduction of notation and discussion of particularly common CG representations of biomolecules in Sec. II, the remainder of this review is divided into two types of sections. Sections III, VIII, IX, and X briefly consider basic philosophies for coarse-graining, provide a very broad overview of CG models for biomolecules, identify certain outstanding challenges in the field, and provide concluding comments, respectively. These sections may prove most useful as a first survey and introduction to the field. The remaining sections provide concise, but somewhat more technical, reviews of certain leading approaches. In particular, these sections address "top-down" approaches for modeling emergent phenomena (Sec. IV), the many-body potential of mean force (Sec. V), "knowledge-based" approaches that harness information from experimentally determined structures (Sec. VI), and "bottom-up" approaches that integrate information from more detailed models (Sec. VII). These sections seek to outline the fundamental assumptions and philosophies of various approaches, elucidate their inter-relationships, and also provide a nonexhaustive, but hopefully useful, introduction to the relevant literature.

a)Electronic mail: wnoid@chem.psu.edu.

II. REPRESENTATION

When constructing a particle-based model for a specific system, one must first define the particles that are used to represent the system. In the case of an atomically detailed model, these particles simply correspond to the atoms in the system. In contrast, the particles in a CG model, which we shall refer to as sites, correspond to groups of one or more atoms. We shall refer to the procedure for transforming an atomically detailed structure into a CG representation of the same structure as the "CG mapping." Clearly, one may tailor the mapping to precisely capture the key features of a specific system or phenomenon of interest, while simultaneously eliminating atomic details that are considered "unimportant." This flexibility in designing the mapping provides one of the greatest advantages of CG models.

It is useful to distinguish two components of this mapping.[20] The "system mapping" specifies the number, types, and connectivity of the sites that are used to describe the system. In many cases, this mapping associates each site with a particular atomic group, defines the physicochemical character (or type) of the site on the basis of this group, and then connects these sites with "CG bonds" on the basis of chemical bonds between the associated atomic groups. Given this system mapping, a "coordinate mapping," $\mathbf{M}$, determines the configuration, $\mathbf{R}$, of the CG model as a function of the configuration, $\mathbf{r}$, of an underlying atomistic model. As illustrated in Fig. 1, the Cartesian coordinates, $\mathbf{R}_I$, of site I are typically determined as a linear combination of atomic Cartesian coordinates, $\mathbf{r}_i$, with constant, positive coefficients that often correspond to, e.g., the center of mass or geometry for the associated atomic group

$$\mathbf{R}_I = \mathbf{M}_I(\mathbf{r}) = \sum_i c_{Ii}\mathbf{r}_i. \tag{1}$$

This mapping applies for a wide range of CG models, including models with explicit or implicit solvent and also models in which different molecules are represented with different resolutions.[21] However, adaptive resolution models, in which the resolution of a molecule may change during the simulation,[22–27] require a more complex mapping. Although some models explicitly treat the internal structure[28–30] or anisotropy of sites,[31,32] we will primarily focus on the common case that the CG sites are structureless point particles.

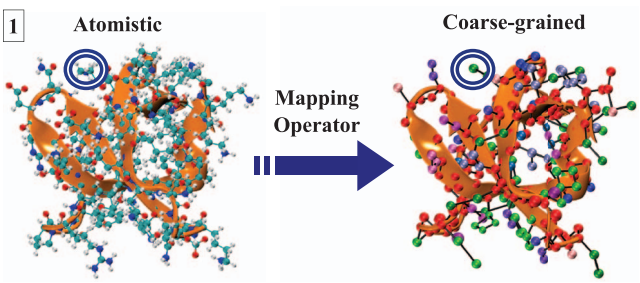

FIG. 1. Schematic for the mapping of an atomistic structure, $\mathbf{r}$, to a CG structure, $\mathbf{R}$. For clarity, the figure omits the solvent that would be present in the atomistic model, but that would typically be eliminated in the CG model. The ribbons indicate the underlying protein fold.

The construction of the mapping is not only the first step in defining a CG model, but is also intimately related to the accuracy, efficiency, and transferability of the model. In particular, the accuracy and transferability of a CG model do not necessarily increase with increasing resolution of the system mapping.[20,33] As very clearly articulated by van Gunsteren and co-workers,[18] one intuitively expects that the ideal CG mapping will: (1) preserve the features that are necessary to describe both the phenomena of interest and also the relevant slow, large amplitude motions of the system; (2) eliminate sufficient detail to provide significant gains in computational efficiency and, moreover, to filter out the high frequency, low amplitude fluctuations that are only weakly coupled to the slower, global motions; and (3) facilitate a transparent and computationally efficient treatment of the physical forces that govern the phenomena of interest. Unfortunately, though, relatively few approaches have been developed for systematically optimizing this mapping. One approach is to design high-resolution CG mappings that allow accurate reconstructions of the atomistic structure[34–37] or potential energy[38,39] (as opposed to the free energy). In the case of relatively low resolution CG models for large biomolecules or biomolecular complexes,[40,41] mappings have been systematically optimized by considering correlated fluctuations[42–48] and experimentally determined density maps.[49–51] In practice, though, CG mappings are most often based upon the "chemical intuition" of the researcher. Consequently, in the remainder of this section, we briefly and non-exhaustively survey some of the most popular mappings for proteins, nucleic acids, and water. For lack of space, we will not specifically address CG mappings for lipid membrane systems, which have enjoyed a rich history,[52–58] or for carbohydrate systems, which are an emerging area for CG modeling.[59–66]

A. Proteins

Since the pioneering studies of Levitt and Warshel,[67,68] many CG protein models have represented each amino acid with one or a few sites having properties that are completely specified by the amino acid type. One site is often associated with the α-carbon of the amino acid, since this allows for detailed reconstructions of the protein backbone.[36,37] Moreover, a generalized Ramachandran map can distinguish secondary structures on the basis of the α-carbon positions.[69] In some cases, such "minimal" models do not distinguish between all types of amino acids, but instead classify them into a small number of site types based upon, e.g., hydrophobicity.[70,71] Despite their remarkable simplicity, such simple top-down models have proven tremendously influential for elucidating generic principles that govern, e.g., protein folding and interactions.[17,72–75] In contrast, network-based and Gō-type models also often represent each amino acid with a single site, but distinguish each site on the basis of its position in the three-dimensional folded structure. By explicitly biasing the CG potential to stabilize this structure, these models have proven remarkably useful for characterizing protein fluctuations and folding.[76,77]

More detailed models typically include sites for both the peptide backbone and sidechain.[67,78,79] For instance, the celebrated UNRES model describes the amide peptide group with a point site and the sidechain with an ellipsoidal site.[80–82] Models with more detailed sidechains have proven useful for modeling the interactions of proteins with other proteins or with nucleic acids.[83,84] Conversely, because models with 3 or more backbone sites per residue provide a better description of secondary structure and hydrogen bonding, they have enjoyed considerable success in simulating peptide aggregation and in folding small proteins without relying upon explicit bias towards the native state.[85–95] On the other end of the spectrum, protein models with even lower resolution have provided insight into protein flexibility, functional fluctuations, biomolecular assemblies, cytosolic crowding, and protein-protein interactions.[29,40,41,44,47,49,51,96–101] We direct the reader to several excellent reviews for further discussion of common CG representations of proteins.[102–105]

B. Nucleic acids

Early computational studies of double stranded DNA employed very low resolution CG models to address length scales similar or larger to the DNA persistence length, which is approximately 50 nm under physiological conditions.[106] These models often employed rigid base pairing or polymer elasticity theory to investigate, e.g., DNA supercoiling.[107–109] While elasticity-based models continue to be actively developed,[110,111] recent CG models have introduced greater detail[112–119] for describing, e.g., DNA melting,[118] renaturation,[120] or nanotweezers.[121] Several models have represented each DNA nucleotide with three sites[122] that usually correspond to the phosphate, sugar, and base moieties.[123–125] Some models have adopted relatively complex potentials to describe the planarity of base pairs and the anisotropy of their interactions.[122,125–127] Higher resolution models have employed additional sites to describe the planarity of the nucleobases[39,128] and, in some cases, also the sugar backbone.[129–132]

The remarkable flexibility and conformational diversity of RNA enable vital biological functions, but also present considerable challenges for computational modeling. Nevertheless, in recent years, several groups have introduced CG RNA models. Often these models represent each nucleotide with three sites, corresponding to the phosphate, sugar, and base moieties.[133–135] More detailed models have employed additional sites to model the base[136] and sugar groups,[137,138] while lower resolution models have employed a single site per nucleotide.[139,140] We direct the reader to several recent reviews that describe advances in the coarse-graining of nucleic acids.[141–144]

C. Water

Water presents significant challenges for coarse-graining. Because water can easily comprise more than 80% of the particles and computational effort in an atomistic MD simulation, it is highly desirable to optimize the representation of CG solvent. However, although water molecules are small and (relatively) simple in isolation, their collective effect generates powerful and highly cooperative hydrophobic forces that drive protein folding[145] and stabilize base-pair stacking in nucleic acids.[146] Moreover, because it determines the dielectric environment, the description of water is also intimately coupled to the treatment of long-ranged electrostatic interactions.

Arguably the most straightforward strategy is to represent each water molecule with a single site, in which case the CG model necessarily requires nearly one-third as many particles as the atomistic model. These models typically employ spherically symmetric solvent sites that can accurately reproduce the pair correlations of water and provide significant efficiency gains by replacing long-ranged Coulombic interactions with simple short-ranged pair potentials.[147–152] However, these models provide a limited description of the tetrahedral structure and electrostatic properties of water. Two exceptional one-site water models are the soft-sticky dipole model,[153,154] which includes explicit dipole and orientation-dependent hydrogen-bonding interactions, and the monatomic water model,[155,156] which adopts the three-body Stillinger-Weber potential[157] to enforce the tetrahedral ordering of water. These two models have been adopted for CG models of lipid bilayers[158] and DNA[159] with explicit electrostatic interactions.

Several CG models associate multiple (typically 3-5) water molecules with a single site[160,161] and, in some cases, replace explicit long-range electrostatic interactions in favor of soft short-ranged potentials.[160,162–164] While this approach provides tremendous computational efficiency,[165,166] it presents conceptual challenges. Because water molecules diffuse apart in time, in this case, no simple mapping relates atomistic and CG configurations and, consequently, it is no longer straightforward to rigorously relate the structure or interactions between the two models.[167–169] In addition, the CG solvent interaction potentials must be carefully parameterized, since the reduced density and enhanced cohesion of the remaining solvent sites can lead to overstabilization of the solid phase.[166,170] Quite recently, several low resolution water models have introduced explicit "polarizability" to model solvent electrostatics.[171–174] In particular, the "big multipole water" model[175] employs three bonded sites to describe the local dipole and quadrupole of 4-water clusters. Intriguingly, CG models that preserve the anisotropy of water interactions may provide a more realistic description of hydrophobic effects.[176,177]

Finally, perhaps the most appealing and commonly adopted approach for treating water in CG models is to formally integrate out the entire solvent. The electrostatic interactions between the remaining CG sites have been modeled with Debye-Huckel potentials,[178] with explicit Coulomb potentials in a constant dielectric,[179] or, as is most commonly done, simply subsumed into short-ranged nonbonded potentials.[180] Many-body hydrophobic forces are then approximately addressed either by implicitly incorporating their effects into the remaining nonbonded potentials or by explicitly modeling their effects.[85,120,181–183] Somewhat surprisingly, although implicit solvent models based upon the generalized Born and Poisson-Boltzmann frameworks have been

extensively developed for atomically detailed models,[184,185] these approaches have not yet been significantly developed in biomolecular CG models.

We direct the interested reader to recent reviews that discuss CG water models in greater detail.[186,187] In addition, an excellent review by Dill *et al.*[188] provides a complementary perspective that clearly demonstrates the power of very simple CG models, such as the two-dimensional Mercedes-Benz model, for understanding the thermodynamic properties of pure water and the physical basis of hydrophobic effects.[189]

III. PERSPECTIVE: PHILOSOPHIES OF COARSE-GRAINING

Upon determining a low-resolution representation for a particular system, the remaining challenge is to determine the appropriate interactions between the CG sites. These interactions should capture the effects of the atomic details that have been eliminated from the CG model. For instance, in an implicit solvent CG model for a folded protein, the interactions between the CG sites must capture the hydrophobic stabilization that is provided in an atomistic model by interactions with the explicit solvent molecules. Thus, the interactions in the CG model should reflect the "correct physics" so that the model is not only highly efficient, but also provides, ideally, both fundamental insight and quantitatively accurate predictions.

A vast array of approaches and philosophies have been adopted for achieving this goal. It seems quite challenging to catalog the resulting "landscape" of CG models in a completely coherent fashion. However, two important paradigms in the field are the "top-down" and "bottom-up" philosophies of coarse-graining. Although the distinction is quite intuitive, it should be emphasized from the outset that the distinction between top-down and bottom-up approaches is becoming increasingly blurred as more models integrate aspects of both philosophies. Nevertheless, it may prove useful to briefly make some philosophical comments about these paradigms, which are schematically illustrated in Fig. 2.

By definition, a "coarse-grained" model implies the existence of a more detailed "fine-grained" model. We assume that the fundamental principles of physics provide such a fine-grained model for interpreting and predicting experimental observables based upon the "real" description of the material. In particular, we commonly assume that quantum mechanics, if it were exactly applied, would not only predict the experimental observables with quantitative accuracy, but would also correctly describe the material properties and dynamics on any and every relevant length scale. Of course, this "fundamental" model is itself empirical and may represent a low resolution representation of a yet more detailed model. Moreover, this model cannot be exactly applied in practice because it is prohibitively costly.

A "bottom-up" CG model is a model of a particular system that is constructed on the basis of a more detailed model for the same system. In principle, this high resolution model may be based upon fundamental first principles. More commonly, though, bottom-up models are constructed on the basis of a classical atomistic model that provides an empirical, but

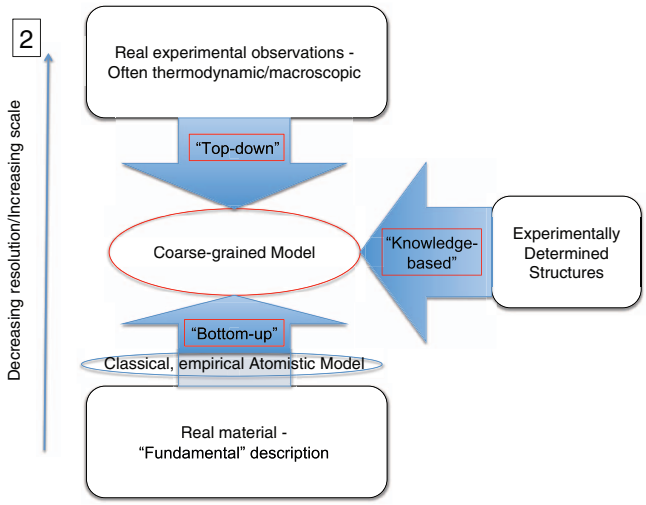

FIG. 2. Schematic illustrating the distinctions between top-down, bottom-up, and knowledge-based strategies for coarse-graining.

highly optimized, approximation to the fundamental model. If no such high resolution model exists, or if it is not sufficiently accurate, then bottom-up strategies cannot succeed. However, given the existence of a high-resolution model, statistical mechanics provides a rigorous framework for determining a CG model that will exactly reproduce, on the scales that can be addressed by the CG model, all thermodynamic and structural properties of the assumed high-resolution model. The central quantity within this framework is the many-body potential of mean force, which is discussed further below. At present, this statistical mechanics framework cannot be exactly applied in practice and many bottom-up models do not explicitly consider this procedure at all. Nevertheless, bottom-up models are unified by their dependence upon information from a more detailed model. Significantly, because they do not rely upon experimental observations of meso- or macro-scopic phenomena, bottom-up models can be applied for making predictions in the case that no such observations exist.

In contrast, "top-down" models do not rely upon or directly relate to a more detailed model for the particular system. Instead, they are related to the "real" system by addressing phenomena that are experimentally observed on length scales that are accessible to the CG model. In this sense, they may be considered a "fine-grained" model constructed on the basis of experimental observables that are even more coarse. "Generic" top-down models are constructed for the sole purpose of assessing the consequences of universal physical principles, sometimes without consideration of any particular system. "Chemically specific" top-down models can be constructed by addressing experimental observations for a particular system. In either case, top-down models provide an elegant framework for determining the highly non-trivial consequences and emergent phenomena that result from the physical principles that were included in the model.[190] Their success in reproducing experimental observables and in predictive modeling may arise from general symmetry principles that determine the structures and properties that are preserved on increasing length scales, as exemplified by thermodynamic theory.[191] However, because a top-down model

only relates to the real system by addressing large-scale emergent phenomena, it is not obvious that a top-down model can be related to any more detailed or more realistic model of the system. In particular, because many different models may lead to the same emergent phenomena, it may prove challenging to assess the validity of the assumptions that underly a top-down model. Similarly, it may prove challenging to systematically improve upon these assumptions. Thus, while bottom-up models are typically over-constrained and necessarily involve loss of information from more detailed models, top-down models are typically under-constrained by the available experimental data.

A second common distinction among CG models for biomolecular systems distinguishes between "physics-based" and "knowledge-based" approaches. Physics-based approaches employ physical theories to infer the interactions in CG models, typically either via bottom-up strategies that address more detailed models or via top-down strategies that address less detailed observables. In contrast, knowledge-based models are constructed on the basis of information extracted from experimentally determined three-dimensional structures. Knowledge-based approaches typically employ empirical relations to determine the interactions in CG models (or scoring functions for bioinformatic calculations), especially in the context of structure prediction and computational protein engineering. Of course, this distinction is also becoming increasingly blurred.

In the present work, we expand the knowledge-based paradigm to include network models and native-centric "Gō-models," which also employ knowledge of known biomolecular structures. We refer to this broader class of models as "native structure-based." Moreover, rather than treating these models as empirical, we will consider them as well-defined approximations within a more general physics-based, bottom-up approach. In particular, if it is assumed that the known structures correspond to results (that would be) generated by a "fundamental" high resolution model, then it may be reasonable to apply physics-based, bottom-up strategies for developing CG models from these structures.

IV. TOP-DOWN MODELS

In top-down models, the interactions are typically parameterized without explicit consideration of a more detailed model. In particular, these interactions are not intended as a well-defined approximation to the many-body potential of mean force for any specific system. Instead, the interactions are often determined either on the basis of physicochemical intuition, or to reflect generic physical principles, or to reproduce certain emergent structural or thermodynamic properties that are observed on even larger scales.

A. Generic top-down models

Historically, CG models have enjoyed great success in assessing the generic consequences of basic physical principles.[8,57,190] These phenomenological top-down models often adopt a relatively low resolution representation and, in many cases, lack the chemical detail necessary to describe any particular system.[70,180,192] In addition, these models often employ relatively simple potentials with relatively few parameters. These parameters can be determined from phenomenological or continuum models.[193,194] Alternatively, they may be systematically varied to investigate the significance of a particular aspect of the model, for instance, the significance of conformational preferences for peptide aggregation,[105,195,196] the significance of chain connectivity and hydrophobic packing for protein folding,[70,197] or the significance of polymer stiffness and charge valency for DNA condensation.[198] In other cases, the parameters in top-down models are tuned to provide a reasonable description of either microscopic properties, such as peptide secondary structure preferences,[93] or macroscopic properties, such as bilayer rigidity[53] or DNA persistence length and melting thermodynamics.[122]

B. Chemically specific top-down models

The top-down strategy can also be adapted for constructing CG models with sufficient chemical specificity to describe a particular system of interest. Chemically specific top-down models often employ interaction potentials with simple functional forms that are parameterized to reproduce thermodynamic properties. Often the sites correspond to 3–4 heavy atoms (aromatic groups are sometimes modeled with slightly higher resolution) or, in the case of solvent sites, 3–4 water molecules. For instance, the multiproperty fitting approach parameterizes potentials for a particular system to reproduce the experimentally observed density and interfacial tension.[162,199] The popular Martini model[163,200–202] extends this paradigm by providing transferable potentials that describe the effects of hydrophobic, van der Waals, and electrostatic interactions between sites as a function of their polarity and charge. These potentials have been optimized to reproduce the partitioning of model compounds between aqueous and hydrophobic environments.[61,163,200,203] These approaches have enjoyed remarkable success for modeling liquid and membrane environments. By incorporating elements of bottom-up strategies, these models can be fine-tuned to reproduce the molecular structure of bilayers,[162] hydrocarbons,[204] and short peptides.[205] At present, *ad hoc* elastic interactions[206] appear necessary to stabilize the structure of more complex biomolecules.[203,207]

V. THE MANY-BODY POTENTIAL OF MEAN FORCE

Bottom-up approaches determine the interactions between CG sites on the basis of a more detailed model. Statistical mechanics provides a formal framework for determining the effective interactions that emerge as a "coarse-grained" manifestation of the interactions in a more detailed model.[208,209] The many-body potential of mean force (PMF) is the central quantity in this approach to coarse-graining.[210–212] The many-body PMF, W, is completely specified by the underlying atomistic model and the CG mapping

$$W(\mathbf{R}) = -k_B T \ln z(\mathbf{R}), \tag{2}$$

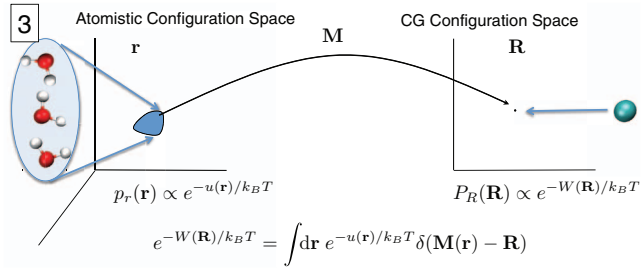

FIG. 3. The many-body PMF, $W(\mathbf{R})$, is defined by the total Boltzmann weight for the volume element of atomistic configurations $\mathbf{r}$ that are mapped to a specified CG configuration, $\mathbf{R}$.

$$z(\mathbf{R}) = \int d\mathbf{r} \exp[-u(\mathbf{r})/k_BT]\delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}), \qquad (3)$$

where u is the potential of the assumed atomistic model, which may be arbitrarily complex, $\mathbf{M}$ is the mapping defined in Eq. (1), and $\delta(\mathbf{M}(\mathbf{r}) - \mathbf{R}) = \prod_{I=1}^{N} \delta(\mathbf{M}_I(\mathbf{r}) - \mathbf{R}_I)$. As schematically illustrated in Fig. 3, W weights each CG configuration, $\mathbf{R}$, according to the total Boltzmann weight for the atomistic configurations, $\mathbf{r}$, that map to $\mathbf{R}$. Consequently, if (for a given temperature T and volume V) $W(\mathbf{R})$ is known to within an additive constant, then this function can be used to sample a canonical distribution of CG configurations (at the given T and V) that is equivalent to the one implied by the atomistic model and the CG mapping. (We note in passing that similar analysis of the entire phase space integral determines site masses such that the CG model reproduces the equilibrium momenta distribution implied by the atomistic model.[213] However, it should be stressed that this certainly does not imply that the CG model will necessarily reproduce the dynamical properties, such as time correlation functions, of any atomistic model.[214])

Importantly, the many-body PMF is **not** a conventional potential energy function, but should rather be considered a configuration-dependent free energy function that reflects both energetic and entropic contributions. The CG forces determined from gradients of the PMF equal conditioned averages of atomistic forces, i.e., W is truly a potential that generates mean forces.[213] Consequently, the PMF quantifies the reversible work associated with changing the CG configuration of (an atomistic model for) a molecular system.[81,215] Although not explicitly indicated above, $W(\mathbf{R})$ is a function of the thermodynamic state. In addition to depending upon the molecular composition of the system, W depends upon V through the limits of integration and upon T through the Boltzmann weight given to each $\mathbf{r}$. Given W as a function of $\mathbf{R}$, V, and T, one can, in principle, reconstruct both structural and thermodynamic properties of the atomistic model. However, because u is typically assumed to be state-point independent, the representation of thermodynamic quantities is fundamentally different in the CG model,[216,217] as discussed further below.

Unfortunately, $W(\mathbf{R})$ is impossible to calculate for most systems, since evaluating Eq. (3) requires performing a free energy calculation as a function of the Cartesian coordinates for all of the CG sites.[218,219] Moreover, as a consequence of integrating out (i.e., renormalizing) the atomistic coordinates, W is generally a many-body function with new couplings between the remaining CG coordinates that cannot be represented by a simple molecular mechanics form.[220] Thus, even if it could be determined, W is generally too complex for subsequent computations.

In essence, the outstanding challenge for "bottom-up" coarse-graining strategies is to determine approximations to W that are tractable to calculate, efficient to simulate, sufficiently accurate for describing a given phenomena, and, ideally, useful (i.e., transferable) for modeling molecular systems and thermodynamic state points other than the one for which they were parameterized. At present, no single method completely addresses this challenge. Rather, a vast array of diverse approaches have been developed for determining the interactions in CG models. In the following, we consider "native structure-based" and "bottom-up" approaches to approximating W.

VI. NATIVE STRUCTURE-BASED MODELS

Since its inception, the protein databank[221] (PDB) has provided the foundation for the development and parameterization of many CG models.[68,222] Network models, native-centric models (also referred to as Gō, native-based, or structure-based models), and knowledge-based models are three remarkably popular implicit solvent CG models that are explicitly constructed on the basis of known PDB structures. We shall treat these "native structure-based" approaches as providing various approximations to the many-body PMF. As the quantity and quality of high resolution structures continues to rapidly increase, one may expect that native structure-based approaches should only become increasingly important in the future.

"Native contacts" feature particularly prominently in the construction of native structure-based models. Given the CG representation of a folded structure, $\mathbf{R}_0$, a pair of amino acid residues (or, more generally, a pair of CG sites) form a "native contact" if the distance between the residues is less than a predefined threshold distance. On this basis, the interaction between the pair is classified as either "native" or "non-native," irrespective of the physicochemical character of the residues. By explicitly ensuring that the pair potentials governing native interactions are all minimized in the folded structure and by ensuring that nonnative interactions are significantly less favorable, network and native-centric models stabilize the experimentally determined three-dimensional structure. As a consequence of this feature, as well as their remarkable simplicity and tremendous computational efficiency, these models are perhaps the most popular and successful models for studying the folding, fluctuations, and interactions of proteins, protein complexes, and, to a lesser extent, also nucleic acids with known equilibrium structures. In contrast, knowledge-based approaches employ information obtained from the experimentally determined structures for multiple distinct proteins to determine potentials that are intended to be transferable for modeling other proteins.

A. Network models

Network models provide perhaps the simplest approximation to the many-body PMF. Network models[223–225] approximate the PMF for a protein about its folded structure, $\mathbf{R}_0$, as a quadratic function of either the distances between sites (Elastic Network Model, ENM) or the Cartesian displacements of sites (Gaussian Network Model, GNM). In principle, free energy calculations can determine the coefficients of this expansion, which corresponds to the curvature of the PMF about $\mathbf{R}_0$. Instead, network models determine these coefficients on the basis of the native contact map. The ENM potential[223] may be expressed

$$U_{\mathrm{ENM}}(\mathbf{R}|\mathbf{R}_0) = \frac{1}{2}\sum_{I<J} k_{IJ}\Delta_{IJ}(\mathbf{R}_0)|R_{IJ} - R_{IJ;0}|^2, \qquad (4)$$

where R_{IJ} and $R_{IJ;0}$ are the distances between CG sites I and J in configuration $\mathbf{R}$ and in the folded configuration, $\mathbf{R}_0$, respectively. In Eq. (4), k_{IJ} is a spring constant and $\Delta_{IJ}(\mathbf{R}_0)$ is the matrix of native contacts that equals 1 if $R_{IJ;0}$ is less than a specified cutoff, R_c, (i.e., if sites I and J form a native contact) and is zero otherwise. The GNM potential[226,227] may be expressed as

$$U_{\mathrm{GNM}}(\mathbf{R}|\mathbf{R}_0) = \frac{1}{2}\sum_{I<J} k_{IJ}\Delta_{IJ}(\mathbf{R}_0)|\mathbf{R}_{IJ} - \mathbf{R}_{IJ;0}|^2, \qquad (5)$$

where $\mathbf{R}_{IJ}$ and $\mathbf{R}_{IJ;0}$ are the vector displacements from site I to site J in configurations $\mathbf{R}$ and $\mathbf{R}_0$, respectively.

Clearly, the folded structure corresponds to the minimum of both the ENM and GNM potentials. Because the GNM potential is a quadratic function of Cartesian coordinates, the free energy, the average fluctuations of each site, and the covariance of these fluctuations can be analytically determined from simple Gaussian integrals.[227] However, because the GNM potential can be separated into independent terms governing motion along each Cartesian direction, the GNM predicts spatially isotropic fluctuations and its normal modes do not correspond to meaningful displacements in three-dimensional space. In contrast, the ENM, which is sometimes referred to as an anisotropic network model, cannot be so easily analyzed, but provides more realistic normal modes.[225]

Network models often represent each amino acid with a single site corresponding to the α carbon and employ a single spring constant for modeling all interactions (i.e., $k_{IJ} = k$). Despite employing only two parameters (R_c and k), these network models have reasonably reproduced the fluctuations that are reported by experimental B-factors[228,229] and essential dynamics analysis[230] of MD simulations.[231] These successes have suggested that protein fluctuations may be largely dominated by shape and packing considerations.[232–234] Network models have been extensively developed to incorporate distance-dependent[224,235–237] and optimized[238–240] spring constants, to address models of various or spatially heterogeneous resolution,[96,223,241–245] to describe conformational transitions between multiple reference configurations,[246–251] and also to treat many other aspects of protein motion.[229,252–257] They have been applied for a similarly wide range of applications, including modeling and refining experimentally determined structures,[258–260] investigating conformational transitions,[261–268] and modeling protein docking.[269–271] In addition, several web servers have been designed for constructing and analyzing network models.[272–275] We direct the reader to several excellent reviews for more information on elastic network models.[76,276–278]

B. Native-centric models

Native-centric models are variously referred to as Gō,[279] native-based, or structure-based models. Native-centric models share with network models the dramatic "teleological" assumption that the strength of nonbonded interactions between protein residues is determined primarily by the folded three-dimensional structure of a protein, rather than by the physicochemical character of the residues.[280] Similarly to network models, native-centric models explicitly stabilize native contacts so that the known folded reference structure, $\mathbf{R}_0$, minimizes the approximate potential. However, in contrast to the harmonic potential employed by network models, native-centric models approximate the many-body PMF with an anharmonic molecular mechanics potential that can be used to describe the interactions and folding of unfolded conformations.[281] In the simplest native-centric models,[281] each residue is represented by a single site that corresponds to the α carbon; the nonbonded interactions that correspond to native contacts are modeled via attractive Lennard-Jones-type potentials with equal attractive strength that are all minimized in $\mathbf{R}_0$; the nonbonded interactions between all other pairs, i.e., non-native interactions, are modeled with simple repulsive potentials to describe excluded volume; and the connectivity of the molecule is enforced by a simple bonded potential that is often biased to enforce the local geometry of the reference structure.

These models are motivated by Gō's observation that the folded structure of many natural proteins achieves mutual "consistency" among all relevant interactions.[282] These proteins are "minimally frustrated" in the sense that the native state not only optimizes the configurational free energy for the entire protein, but also simultaneously optimizes the individual interactions in the protein.[283–285] By eliminating the energetic frustration that arises from non-native interactions, native-centric models provide an idealized "funnel" for folding: the CG potential (which approximates the many-body PMF and thus corresponds to a configurational free energy) decreases if and only if the protein configuration becomes increasingly similar to the known folded native structure, i.e., by forming a native contact.[286–288] However, because it explicitly treats both excluded volume and the protein backbone connectivity, the model reasonably describes the "topological frustration" that results from the loss of configurational entropy as the protein folds.[289–292] In this sense, it has been argued that "Gō models may still be closer to reality than our current realistic models,"[293,294] which may not provide adequate bias towards folded structures.[295,296] In support of the native-centric hypothesis that protein structure determines the folding mechanism, experiments have suggested that in many,[297] though not all,[298] cases the folding rates of

small proteins correlate with the complexity of the resulting structure.[299,300]

As a consequence of their simplicity and capability to guide proteins from an unfolded conformation to the correct folded conformation, native-centric models have enjoyed remarkable popularity. Atomically detailed native-centric models[301,302] and "flavored" C-α models, for which the strength of native interactions varies with residue identity,[303] have proven capable of distinguishing between the folding mechanisms of proteins with remarkably similar native structures. In order to provide an improved description of the interactions in unfolded proteins and the cooperativity of folding, native-centric models have introduced phenomenological potentials for modeling desolvation effects,[304] non-native interactions,[305–307] cooperative many-body interactions,[308,309] and electrostatic interactions.[179] Native-centric models have been extended to describe conformational transitions between multiple structures,[310–312] to address the effects of denaturants and osmolytes,[313–315] and to address experimentally determined information regarding free energies of mutation.[316] Native-centric models have been extensively applied not only for modeling protein folding, but also for modeling conformational transitions in biomolecular complexes,[317,318] for investigating the effects of confinement and crowding upon folding,[319–321] for studying nucleic acids,[123,322] for describing AFM experiments,[133,323] and even for studying the remarkable mechanism by which intrinsically disordered proteins couple their folding to ligand binding.[324–326] Several studies have suggested that the conclusions obtained with native-centric models are robust to many details of the model.[293,327] Interested readers are referred to several webservers[328,329] and recent reviews[77,330] for more information regarding native-centric models.

C. Knowledge-based models

Network models and native-centric models employ an experimentally determined structure (or, in some cases, a few structures) for a single protein to construct a potential for a CG model that explicitly stabilizes this structure (or these structures). Clearly, the resulting potentials cannot be applied to other proteins and, more generally, these models cannot be developed for proteins with unknown structure. In contrast, knowledge-based approaches employ known PDB structures for multiple nonhomologous proteins (i.e., proteins with sufficiently distinct sequences) to determine a set of transferable, i.e., protein-independent, potentials that are intended for modeling multiple proteins, including proteins that were not in the training set and proteins that may have unknown structure. Although a tremendous diversity of knowledge-based models have been developed, Subsections VI C 1–VI C 3 briefly consider three classes: (1) models with protein-independent potentials that have been determined from PDB statistics, (2) models with protein-independent potentials that have been optimized on the basis of known PDB structures, and (3) models that are explicitly tailored for a particular protein on the basis of PDB structures.

1. Statistical potentials

As originally suggested by Tanaka and Scheraga,[222] "statistical potentials" estimate the effective interaction between a pair of amino acid residues based upon the statistical frequency of finding contacts between the pair in PDB structures. Miyazawa and Jernigan extended this approach by adopting a quasichemical approximation to treat solvent effects and correlations between amino acids.[331,332] The resulting Miyazawa-Jernigan potential, which was initially developed for lattice protein models and later extended for off-lattice models,[333] emphasizes hydrophobic interactions[334] and remains commonly adopted as a first estimate of the interaction between amino acids in CG models.[335,336] Subsequently, Sippl[337,338] proposed a "Boltzmann device" for determining statistical potentials in analogy to the pair potentials of mean force that are obtained from direct Boltzmann inversion, as discussed further below. According to the Boltzmann device, the distance-dependent potential, $U_\zeta(r)$, that governs the interaction between a particular pair of residue (or CG site) types, ζ, may be determined

$$U_\zeta(r) = -k_B\tau \ln[N_\zeta(r)/N_{\mathrm{ref};\zeta}(r)]. \quad (6)$$

In Eq. (6), $k_B\tau$ defines an energy scale (which is often, though not always,[339] associated with room temperature[340]) and $N_\zeta(r)$ is a histogram that counts the instances in which the residue pair, ζ, is separated by a distance r in a set of PDB structures. As discussed in greater detail later, the distribution of distances between a pair of amino acids reflects not only the direct interaction between the pair, but also the effects of the protein environment, including the packing of surrounding residues and backbone connectivity. The normalizing factor, $N_{\mathrm{ref};\zeta}(r)$, attempts to account for these effects and is defined as the histogram that would be observed in a "reference state" of imagined protein-like structures that lack specific direct interactions between amino acids.[337,338]

The Boltzmann device has been widely adopted for determining very many statistical pair potentials that primarily differ in their representation of proteins and in their treatment of the reference state.[341–348] Statistical potentials have also been developed for modeling backbone torsions,[349,350] as well as for orientation-dependent[351–355] and many-body interactions.[356–359] The uncertainty in the reference state has motivated iterative self-consistent approaches[360,361] for determining statistical potentials. These self-consistent approaches bear striking similarity to the iterative Boltzmann inversion approach,[362] which is widely adopted for developing bottom-up polymer models and discussed further below. We note in passing that Shakhnovich and co-workers have developed an alternative knowledge-based strategy for determining atomically detailed statistical potentials that generalizes native-centric approaches and that has successfully folded several small proteins.[363–365]

The basis for determining potentials from PDB statistics in analogy to the Boltzmann distribution for a canonical ensemble has sparked extensive discussion.[339,340,366–375] In particular, similar relations have been justified on the basis of information theoretic arguments, in which case the statistical potentials are regarded as statistical preferences.[371,376,377]

Aside from controversies regarding the physiological relevance of PDB structures,[378–380] it is clear that Eq. (6) will lead to significant errors in determining interaction potentials, unless the unknown reference state is carefully designed to address the correlations between different interactions.[370,381] Furthermore, because conventional ensembles address individual physical systems, conventional statistical thermodynamics cannot be used to relate structural information collected for different proteins to physical potentials.[382,383] However, if PDB structures correspond to configurations sampled from physiologically relevant canonical ensembles for distinct proteins,[376,384–387] then it is possible to employ a physics-based theory to rigorously address both of these considerations.[388] Regardless of these theoretical considerations, statistical potentials have continued to enjoy considerable popularity in computational protein science and are often incorporated into complex models with contributions from both knowledge- and physics-based terms.[348,389]

2. Potential optimization

In order for an approximate CG potential to stabilize a protein sequence in its folded native structure, this structure should correspond to the global minimum of the approximate potential (or, in the case of an explicit solvent CG model at finite temperature, of the corresponding free energy surface). Network models and native-centric models accomplish this for a single protein (with known structure) by construction. In contrast, statistical potentials do not ensure this stabilization. Consequently, a second common knowledge-based strategy employs PDB structures for multiple proteins (with distinct sequences) to optimize a set of transferable (i.e., protein-independent) parameters for CG potentials that stabilize these structures.

The simplest optimization strategy was pioneered by Crippen and co-workers.[390,391] In this approach, a collection of proteins with known PDB structures was identified as a training set. The potential parameters were then optimized via linear programming methods such that, for each protein in the training set, the corresponding PDB structure minimized the potential (for that protein sequence) relative to any other structure considered. Similar strategies have determined CG potentials to optimize the entropy, foldability, or thermodynamic stability of known protein structures.[392–397] These optimization procedures have been extended to provide a "funneled" free energy surface that correlates the approximate CG potential with similarity to the folded native structure.[398–402] Potentials have also been optimized on the basis of intermolecular interactions between proteins and nucleic acids.[132,138]

In addition to folded structures, such optimization procedures also require knowledge of alternative, competing "decoy" conformations as a basis for comparison when determining the structure that minimizes a potential.[403,404] Clearly, the quality of the optimized potential should depend upon the quality of these decoys. Decoy conformations have historically been treated in various ways, e.g., by explicitly generating structures via energy minimization or molecular simulations,[405,406] by applying random energy models to estimate the statistical features of decoy structures,[392] or by "threading," i.e., evaluating the potential for a particular protein sequence using PDB structures for other proteins.[391]

Importantly, these approaches have provided strong evidence that, if each amino acid is represented by a single site and if the properties of that site are completely specified by the amino acid type, then pair additive interactions are likely inadequate for determining accurate and transferable potentials for CG models of folded proteins.[407,408] (Of course, native-centric models surmount this difficulty by determining the interactions of each site on the basis of structure, rather than amino acid type.) Although these optimization procedures were initially developed for simple lattice-based protein models, they have since been adopted for optimizing the weights associated with individual terms (including statistical pair potentials or many-body terms) in much more complex CG potentials.[81,406,409,410]

3. Fragment and homology based approaches

Finally, a third class of knowledge-based models explicitly employs structural information from known PDB structures during simulations of the CG model. These approaches often employ transferable protein-independent potentials that may be determined from statistical analysis or optimization, as described above. However, the total potential is then supplemented with additional terms that explicitly tailor the potential for a particular protein of interest, often with *unknown structure*, on the basis of structures that have been experimentally determined for other proteins with similar sequences. For instance, models have been tailored for specific proteins of unknown structure by supplementing transferable statistical potentials with Gō-like potentials that stabilize *predicted* native contacts.[411,412] This strategy may prove increasingly powerful due to very recent improvements in the prediction of native contacts from multiple sequence alignments.[413–415] A second strategy, which has been adopted by leading methods for protein structure prediction,[416–418] employs transferable statistical potentials with optimized weights, but then employs Monte Carlo moves that explicitly sample structural fragments from known PDB structures when simulating the folding process.[377,419–422] Finally, the associative memory Hamiltonian model[423,424] synthesizes components of both knowledge- and physics-based approaches by adopting a physically-motivated potential with additional knowledge-based terms that correspond to "memories" of protein structures with similar sequences. This approach has also been extended to incorporate fragments from a library of known structures.[425,426]

VII. "BOTTOM-UP" APPROACHES

"Bottom-up" coarse-graining approaches employ information from a more detailed model when constructing the potential for a CG model of the same system. A wide range of bottom-up strategies have been proposed on the basis of, e.g., reproducing energetics,[81,427,428] matching structural

properties,[429] and integral equation theories.[430,431] This section surveys several leading structure-motivated, bottom-up coarse-graining strategies that determine the CG potential as a systematic approximation to the many-body PMF. We direct interested readers to a recent book chapter[432] for a detailed treatment of the basic statistical thermodynamics principles underlying bottom-up coarse-graining and to recent software packages that implement these approaches.[433,434]

In many such approaches, the approximate CG potential, $U(\mathbf{R})$, adopts a generalized molecular mechanics form, in which each bond, angle, torsion, and pair nonbonded interaction is modeled by a single potential that is a function of the corresponding mechanical degree of freedom. This may be compactly represented

$$U(\mathbf{R}) = \sum_{\zeta} \sum_{\lambda} U_{\zeta}(\psi_{\zeta}(\mathbf{R}_{\lambda})). \tag{7}$$

In Eq. (7), ζ identifies a particular type of interaction that is modeled by a potential, U_{ζ}, which is a function of a single scalar variable, ψ_{ζ}, which is itself a function of the Cartesian coordinates, $\mathbf{R}_{\lambda}$, for a particular set, λ, of particles. For instance, if ζ corresponds to a type of nonbonded pair interaction, then U_{ζ} is the corresponding pair potential, ψ_{ζ} is a pair distance, and $\mathbf{R}_{\lambda}$, are the Cartesian coordinates for a particular pair, λ, of sites.

Given Eq. (7), the approximate CG potential, U, is determined by an appropriate set of simpler potentials, U_{ζ}, each of which depends upon a single variable. In contrast to top-down, network-based, and native-centric models, which typically represent each potential U_{ζ} with a simple analytic function, bottom-up models typically use more complex functions in order to more accurately approximate the many-body PMF. In practice, bottom-up models often represent these potentials on a grid or with flexible basis functions, such as splines. The coefficients of these basis functions then become the parameters determining the CG potential. Potentials of more complicated form can also be incorporated into the framework.[435]

A. Correlation function approaches

Because it is generally impossible to calculate the many-body PMF, W, for complex systems, the approximate potential, U, is often parameterized with a much more modest goal. Each type of potential function, U_{ζ}, included in Eq. (7) models a particular type of interaction as a function of a scalar variable, ψ_{ζ}, that describes a corresponding mechanical degree of freedom in the CG model. One can then define $P_{\zeta}(x|U)$ as the probability distribution sampled by that degree of freedom when the CG model has a potential U. The potentials, U_{ζ}, are often determined so that, for each degree of freedom treated in Eq. (7), the CG model reproduces the corresponding target distribution, $p_{\zeta}(x)$, that is determined by the atomistic model and CG mapping, i.e., $P_{\zeta}(x|U) = p_{\zeta}(x)$. For instance, each CG pair potential may be parameterized to reproduce the corresponding atomistic radial distribution function. Of course, this condition does not imply that the CG and atomistic models will necessarily generate the same cross-correlations between different degrees of freedom.[436] Also, we note in passing that the system may include multiple instances, λ, of an interaction, ζ, that are modeled with the same potential, U_{ζ}. In this case the probability distributions $P_{\zeta}(x|U)$ and $p_{\zeta}(x)$ reflect an additional averaging over these instances.[435]

Although it can be proven for simple systems,[437,438] it is not obvious *a priori* whether there exists a molecular mechanics potential that will reproduce a given set of target structural correlation functions generated by an atomistic model. However, if these potentials should exist, then, at least in principle, they should be unique.[150,435,439–441] (In practice, though, many structural features are quite insensitive to the CG potentials.[442]) Unfortunately, no direct (i.e., noniterative) mechanism exists for determining potentials that are guaranteed to reproduce target correlation functions. Consequently, structure-motivated bottom-up methods often employ iterative nonlinear optimization techniques to systematically refine the CG potential until it accurately reproduces the target correlation functions of an atomistic model.[443–447]

1. Direct Boltzmann inversion

Direct Boltzmann inversion (DBI) provides the simplest and most straightforward approach for determining CG potentials from atomistic distribution functions.[448] Each term in the CG potential is determined directly from the corresponding atomistic distribution function,

$$U_{\zeta}(x) = -k_B T \ln[p_{\zeta}(x)/J_{\zeta}(x)], \tag{8}$$

where $J_{\zeta}(x)$ is a corresponding Jacobian factor. In the following, we primarily focus on the case that $U_{\zeta}(x)$ corresponds to a nonbonded pair potential, although analogous considerations apply for other types of potentials. In this case, the potential determined by DBI via Eq. (8) is equivalent to the atomistic pair potential of mean force (ppmf):

$$w_{\zeta}(r) = -k_B T \ln g_{\zeta}(r), \tag{9}$$

where $g_{\zeta}(r)$ is the corresponding radial distribution function (RDF) determined by the atomistic model and CG mapping. As noted earlier, this definition motivates the Boltzmann device for determining statistical pair potentials.[337,338,374] It should be emphasized that the ppmf, w_{ζ}, which corresponds to a pair distance r, is analogous to, but distinct from, the many-body PMF, W, which corresponds to the entire CG configuration $\mathbf{R}$.

DBI has enjoyed considerable success in polymeric systems.[449] DBI has also been employed in parameterizing models for short peptides[450,451] and RNA,[136] as well as for improving the bonded potentials in Martini models.[205] In particular, DBI determines potentials that provide an accurate description for interactions that are effectively isolated in the CG model, since in this case the corresponding CG distribution only reflects the single potential. Consequently, DBI is often useful for nonbonded pair potentials if the CG sites are "dilute" and for bonded potentials if the CG bonds are "stiff."[448,452] Importantly, this criterion relies upon both the system and also the CG mapping. More generally, though, the interactions in CG models may be strongly coupled and the potentials obtained via DBI will not reproduce atomistic distribution functions.

A recent generalization[453,454] of the Yvon-Born-Green (g-YBG) equation[455–459] provides a quantitative framework for understanding this discrepancy. If $W_\zeta(r|U) = -k_BT \times \ln[P_\zeta(r|U)/J_\zeta(r)]$ is the ppmf for the ζ pair interaction in a CG model with potential U, then the pair mean force $-W'_\zeta(r|U)$ equals the average magnitude of the net force on one site of the pair when the second site is a distance r away.[455,456] Importantly, if the CG model reproduces the pair mean force of the atomistic model, $-w'_\zeta(r)$, at every distance r, then the CG model will also reproduce the atomistic ppmf, $w_\zeta(r)$, and, thus, the corresponding atomistic RDF.

As illustrated in Fig. 4, the g-YBG theory exactly decomposes the CG pair mean force, $-W'_\zeta(r|U)$, into a "direct" contribution due to the force, $F_\zeta(r) = -dU_\zeta(r)/dr$, between the pair of sites and a correlated "indirect" contribution from the surrounding environment. This indirect contribution can itself be decomposed into forces $F_{\zeta'}(x) = -dU_{\zeta'}(x)/dx$ due to every other term in the CG potential,

$$-W'_\zeta(r|U) = F_\zeta(r) + \sum_{\zeta'} \int dx' \tilde{G}_{\zeta\zeta'}(r, x'|U) F_{\zeta'}(x'), \quad (10)$$

where $\tilde{G}_{\zeta\zeta'}(r, x'|U)$ is a relatively simple structural correlation function describing the coupling between the ζ and ζ' interactions in the CG model.[453,454] If the CG potential is determined via DBI according to Eq. (8) so that $F_\zeta(r) = -w'_\zeta(r)$ and if, in the CG model, the degree of freedom ζ is statistically uncoupled from all other interactions so that $\tilde{G}_{\zeta\zeta'} = 0$, then $-W'_\zeta(r|U) = -w'_\zeta(r)$ at all distances and the CG model will also reproduce the atomistic distribution, $p_\zeta(r)$.[460,461] More generally, though, the interactions in the CG model are coupled and, consequently, the potentials determined via DBI do not reproduce the atomistic pair mean forces or, equivalently, the RDFs.

In principle, if one could determine the indirect contribution to the CG pair mean force (i.e., the second term in Eq. (10)), then one could determine CG force functions to reproduce atomistic distributions by direct subtraction. This indirect contribution corresponds to the "cavity distribution function" in liquid state theories[462] and, given an extended ensemble framework[20] for combining statistics from distinct proteins, to the reference state in knowledge-based statistical potentials.[337,463] This contribution to the ppmf can be estimated by free energy calculations, as is done in the conditional reversible work method,[464,465] or can be minimized by independently determining CG potentials for small molecular fragments.[466] More commonly, though, bottom-up approaches employ multiple CG simulations with trial potentials to estimate this indirect contribution to the pair mean force and, thus, determine the appropriate direct interaction for reproducing target distributions.

2. Iterative Boltzmann inversion

As indicated by Soper,[467] the integrated form of Eq. (10),

$$W_\zeta(r|U) = U_\zeta(r) + \sum_{\zeta'} W_{\zeta;\zeta'}(r|U), \quad (11)$$

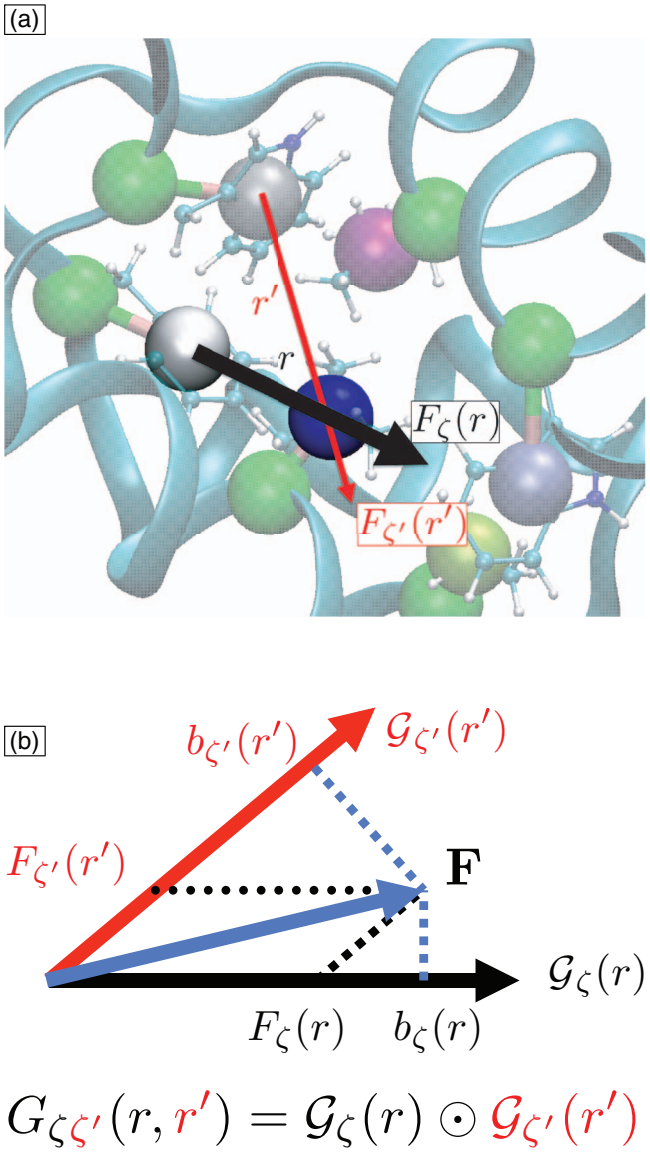

FIG. 4. (a) The pair mean force, $-W'_\zeta(\mathbf{R}|U)$, describing the effective interaction between the Phe34 sidechain (represented by the white site) and the Leu14 sidechain (represented by the blue site) includes a direct contribution from the force between the pair, $F_\zeta(r)$, and also an indirect contribution due to the forces from the surrounding CG sites, including the force, $F_{\zeta'}(r')$, from the Trp6 sidechain (represented by the opaque gray site). (b) The g-YBG/MS-CG theory relates this pair mean force to $b_\zeta(r)$, which is the projection of the CG force field, $\mathbf{F}$, onto the corresponding basis vector, $\mathcal{G}_\zeta(r)$. The metric tensor, $G_{\zeta\zeta'}$, is the "angle" between these basis vectors and quantifies the indirect contributions from, e.g., Trp6 to the pair mean force. For clarity, the figure indicates only the nearest CG sites.

suggests a systematic framework for improving the agreement between the atomistic, $p_\zeta(r)$, and CG distributions, $P_\zeta(r|U)$. If, for a given CG potential U, the CG model samples a particular distance r too frequently (i.e., $P_\zeta(r|U) > p_\zeta(r)$), then the corresponding CG ppmf is too attractive at this distance (i.e, $W_\zeta(r|U) < w_\zeta(r)$). According to Eq. (11), the most straightforward way to increase the ppmf in the CG model (and thus improve the agreement in the corresponding distribution) is to simply increase the corresponding potential, U_ζ. This suggests the iterative Boltzmann Inversion (IBI) algorithm[362,468] for determining a correction, $\delta U_\zeta(r)$, to the CG

potential,

$$\delta U_\zeta(r) \propto w_\zeta(r) - W_\zeta(r|U) \tag{12}$$

$$\propto k_B T \ln[P_\zeta(r|U)/p_\zeta(r)]. \tag{13}$$

The modified potentials, $U_\zeta(r) + \delta U_\zeta(r)$, are then employed in CG simulations, the distributions obtained from these simulations are compared with the atomistic distributions, and the potentials are modified again. This process is repeated until convergence. The IBI approach has proven remarkably popular and successful for modeling complex liquids and polymers.[469] It has also been applied to optimize UNRES potentials for proteins,[470] to model membrane systems,[182,471,472] disordered peptides,[473] and cellulose fibers,[65] as well as to approximate solvation effects in atomically detailed, implicit solvent peptide models.[474] The power of the IBI approach is its tremendous simplicity in treating each interaction U_ζ and each distance r independently. Correlations between different interactions and between different distances are treated implicitly through iteration. In some cases, though, this can lead to convergence difficulties that may be exacerbated by complex couplings between interactions or by the relative insensitivity of structural correlation functions to many aspects of the CG potential.[475–477]

3. Inverse Monte Carlo

Originally motivated by renormalization group calculations,[478] Lyubartsev and Laaksonen introduced an Inverse Monte Carlo (IMC) algorithm for addressing these correlations when refining CG potentials.[479,480] The IMC algorithm considers a susceptibility matrix,

$$K_{\zeta\zeta'}(x, x'|U) = \frac{\delta P_\zeta(x|U)}{\delta U_{\zeta'}(x')}, \tag{14}$$

that describes the sensitivity of each simulated CG distribution to small variations in the CG potential functions. The IMC approach employs Newton's method to determine potentials that minimize the difference between the simulated CG and target atomistic distributions.[481] The corrections to the CG potentials, $\delta U_\zeta(x)$, are determined by solving the system of coupled linear equations:

$$p_\zeta(x) - P_\zeta(x|U) = \sum_{\zeta'} \int dx' K_{\zeta\zeta'}(x, x'|U)\delta U_{\zeta'}(x'). \tag{15}$$

In practice, this and subsequent integral equations are typically solved either by discretization or by employing a linear basis for the potentials. The corrected CG potentials are simulated, the resulting distributions are again compared with the target atomistic distributions, the susceptibility matrix is recalculated, Eq. (15) is solved again to correct the potentials, and the process is repeated to convergence. Importantly, $K_{\zeta\zeta'}(x, x'|U)$ is a positive definite matrix describing the covariance in the degrees of freedom that are governed by U_ζ and $U_{\zeta'}$.[432,479] Consequently, in principle, the IMC algorithm should not become trapped in local minima when optimizing U.[28] The IMC algorithm has been applied to model water as well as implicit solvent models of amino acid solutions[481] and lipid bilayer membranes.[482,483] Savelyev and Papoian have extended the IMC method within a molecular renormalization group framework[484] and demonstrated remarkable results for modeling ionic systems[115] and DNA.[116]

B. Variational approaches

Variational approaches determine a unique CG potential that provides an optimal approximation to the many-body PMF by minimizing an appropriate functional. Because these functionals are independent of the form used to represent the potential, the corresponding variational principles provide great flexibility for numerically optimizing the CG potential. In particular, several of the approaches discussed above can be considered specific numerical methods for minimizing an underlying variational functional. Importantly, by providing a quantitative measure of error, variational approaches provide a rigorous framework for systematically improving the CG potential, in principle, to within desired accuracy by introducing additional interactions into the model.

1. Relative entropy framework

Shell and co-workers[485,486] have developed an elegant variational framework for optimizing CG models based upon the information theoretic relative entropy. Following Kullback and Leibler,[487] one can define

$$\Phi(\mathbf{R}|U) = \ln\left[\frac{p_R(\mathbf{R})}{P_R(\mathbf{R}|U)}\right] \tag{16}$$

to quantify the information available in a single configuration, $\mathbf{R}$, for discriminating between the distribution of CG configurations generated by the atomistic model (and associated mapping), $p_R(\mathbf{R})$, and the distribution generated by a CG model with potential U, $P_R(\mathbf{R}|U)$. If both models sample a canonical ensemble, then Φ may be interpreted as a configuration-dependent entropy: $\Phi(\mathbf{R}|U) = -(W(\mathbf{R}) - U(\mathbf{R}))/k_B T + \text{const}[U]$, where the constant is independent of configuration, but is a functional of U.[432,485] The relative entropy may be defined

$$S_{rel}[U] = k_B \int d\mathbf{R}\, p_R(\mathbf{R})\Phi(\mathbf{R}|U), \tag{17}$$

where the average is evaluated over the CG configuration space, but weighted according to the atomistic probability distribution, p_R. The relative entropy can also be derived as the (negative) log-likelihood for the atomistic and CG models to sample the same distribution of CG configurations.[432,485]

According to the Gibbs inequality,[456] $S_{rel}[U] \geq 0$ and only vanishes when the CG model perfectly reproduces the distribution of configurations sampled by the atomistic model, which determines the many-body PMF W to within an additive constant.[485] Moreover, if the CG potential is of the form given by Eq. (7), then S_{rel} is minimized with respect to a particular term, $U_\zeta(x)$, when the CG model reproduces the corresponding atomistic distribution, i.e., when $P_\zeta(x|U) = p_\zeta(x)$.[435,485] Thus, minimizing $S_{rel}[U]$ provides a variational framework for approximating W and, in particular, for determining approximate potentials that reproduce

target atomistic distributions.[432] In practice, the relative entropy can be numerically minimized by, e.g., the Newton-Raphson method or stochastic optimization techniques.[488] In the former case, the method then corresponds[28] to the IMC method discussed above.[479,480] Shell has recently proposed an algorithm for more efficiently using the statistics sampled from CG trajectories when minimizing S_{rel} and employed the resulting model to study peptide aggregation.[489] More generally, the relative entropy formalism can be applied to optimize parameters that enter in a nonlinear fashion into a CG model[486] and has been applied for optimizing network models.[264] The relative entropy may also prove a powerful framework for optimizing both CG mappings and dynamics.[245,490,491]

2. Multiscale coarse-graining

Izvekov and Voth[492,493] pioneered the multiscale coarse-graining (MS-CG) method as an early variational method for determining CG potentials. The MS-CG method determines an optimal CG potential by minimizing a force-matching[494–496] functional of the trial CG force field, $\mathbf{F}$:

$$\chi^2[\mathbf{F}] = \frac{1}{3N}\left\langle \sum_{I=1}^{N} |\mathbf{F}_I(\mathbf{M}(\mathbf{r})) - \mathbf{f}_I(\mathbf{r})|^2 \right\rangle, \tag{18}$$

where the angular brackets denote a canonical ensemble average evaluated according to the atomistic model.[213,497] In Eq. (18), $\mathbf{f}_I(\mathbf{r})$ is the net force on site I in atomistic configuration $\mathbf{r}$ and $\mathbf{F}_I(\mathbf{M}(\mathbf{r}))$ is the force on the same site that is determined by the trial force field when evaluated in the mapped configuration, $\mathbf{M}(\mathbf{r})$. The many-body mean force, $\mathbf{F}^0$, i.e., the force field determined from the many-body PMF, W, provides the unique global minimum of χ^2.[498] Consequently, minimization of χ^2 provides an independent variational principle for determining, to within an additive constant, the many-body PMF or an optimal approximation to it.[213,497]

The form of the approximate CG potential, Eq. (7), determines a basis set for force fields:

$$\mathbf{F} = \sum_{\zeta} \int \mathrm{d}x\, \boldsymbol{\mathcal{G}}_\zeta(x) F_\zeta(x), \tag{19}$$

where $\boldsymbol{\mathcal{G}}_\zeta(x)$ is a basis vector that defines the direction of forces generated by the potential function $U_\zeta(x)$ and the associated coefficient, which determines the magnitude of these forces, is simply the force function $F_\zeta(x) = -dU_\zeta(x)/dx$. The MS-CG approximation to the PMF corresponds to the geometric projection of the many-body mean force field (and, equivalently, the atomistic force field) into the subspace of force fields that is determined by the terms included in the approximate CG potential, Eq. (7).[454,461,496,497,499] The force functions that determine the MS-CG force field can be calculated from the coupled system of linear equations:[453,497,498]

$$b_\zeta(x) = \sum_{\zeta'} \int \mathrm{d}x' G_{\zeta\zeta'}(x, x') F_{\zeta'}(x'). \tag{20}$$

In Eq. (20), $b_\zeta(x) = \mathbf{F}^0 \odot \boldsymbol{\mathcal{G}}_\zeta(x)$ is the projection of the many-body MF onto the basis vector $\boldsymbol{\mathcal{G}}_\zeta(x)$ and may be calculated as a correlation function of forces sampled from atomistic simulations; while $G_{\zeta\zeta'}(x, x') = \boldsymbol{\mathcal{G}}_\zeta(x) \odot \boldsymbol{\mathcal{G}}_{\zeta'}(x')$ is a structural correlation function describing the correlation between the forces generated by the U_ζ and $U_{\zeta'}$ terms in $U(\mathbf{R})$.[498] In both cases, the inner product, $\odot$, is evaluated as a canonical ensemble average for the atomistic model.[453,497] As illustrated in Fig. 4, the metric tensor, $G_{\zeta\zeta'}$, plays a central role in the MS-CG method by addressing the correlations between different interactions and by quantifying their contributions to each projection, b_ζ, of the many-body MF.[454,461]

There are several important differences between the MS-CG method and other methods that address atomistic correlation functions. While IBI, IMC, and relative entropy minimization involve a nonlinear optimization problem that requires iterative simulations, the MS-CG method involves a linear optimization problem that can be solved directly. Consequently, while these iterative methods should reproduce atomistic distribution functions by construction, the MS-CG method is not guaranteed to perfectly reproduce any particular atomistic distribution functions.

In fact, Eq. (20) for the MS-CG force functions is a simple force balance condition: the atomistic and CG models should generate the same average force along each degree of freedom, ψ_ζ, treated by the CG potential.[453,497] While b_ζ describes the average force generated by the atomistic model, the metric tensor, $G_{\zeta\zeta'}$, provides the structural information that is necessary to estimate the corresponding average force that will be generated by a CG model with force functions, F_ζ. Importantly, though, in Eq. (20), $G_{\zeta\zeta'}$ contains structural information regarding cross-correlations that are observed in the *atomistic* model. If simulations with the calculated MS-CG forces, F_ζ, accurately reproduce these atomistic cross-correlations, then the resulting MS-CG model will quantitatively reproduce the average forces along each degree of freedom and, thus, also the corresponding atomistic distributions. However, the cross-correlations between different degrees of freedom, in general, will not be perfectly reproduced by the CG model.[436] Consequently, the calculated MS-CG force field may not quantitatively reproduce the atomistic distributions for the individual degrees of freedom.[461,497] This argument also motivates an iterative YBG framework[500,501] for improving the MS-CG potential by self-consistently solving for the force functions that satisfy the MS-CG force balance relation (i.e., Eq. (20)) when using the cross-correlations generated by the CG model.[461]

At the same time, there are also remarkable similarities between the MS-CG method and other correlation function-based approaches. For instance, by treating the correlations between different interactions, the metric tensor, $G_{\zeta\zeta'}$, performs a similar function to the susceptibility matrix, $K_{\zeta\zeta'}$, in the IMC and relative entropy approaches. Intriguingly, while the relative entropy variational principle minimizes the average of $\Phi(\mathbf{R}|U)$, the MS-CG variational principle minimizes the average of $|\nabla\Phi(\mathbf{R}|U)|^2$.[435] Consequently, the uniqueness properties of the potentials obtained by minimizing the relative entropy and by minimizing the MS-CG functional are closely related. The MS-CG and relative entropy variational principles both determine the many-body PMF, to within an additive constant, in the limit of a complete basis set and also

obtain the same approximate potential in the case that the CG potential is quadratic in Cartesian coordinates.[435] In addition, the MS-CG "force-matching" method may be applied without any explicit force information by recognizing that the normal MS-CG equations, Eq. (20), are equivalent to the g-YBG equations, Eq. (10).[453,454] For instance, if U_ζ corresponds to a pair potential, then the corresponding projection of the many-body mean force, b_ζ, is related by the g-YBG theory to the pair mean force, $-w'_\zeta(r)$, which can be determined from the RDF. Similarly, $G_{\zeta\zeta'}(x, x')$ can be related to $\tilde{G}_{\zeta\zeta'}(x, x'|U)$.[461]

The theory and methodology of the MS-CG approach have been extensively developed.[502,503] In particular, the MS-CG method has been applied to study membrane systems,[21,181,492,504,505] carbohydrates,[60,62,66] polyglutamine aggregation,[506,507] peptide secondary structures,[508–510] and also larger proteins.[511] Furthermore, the MS-CG method has also been extended for modeling dynamics,[495,512] for coarse-graining fluid field theories,[513] and for treating the internal structure of CG sites.[30]

VIII. PERSPECTIVE: SUMMARY AND OVERVIEW

In the 50 years since their initial application,[514] atomically detailed molecular dynamics (MD) simulations[515,516] have profoundly advanced our current understanding of soft materials, including biological, polymeric, and liquid systems. In these simulations, atoms are typically described by point particles with fixed charges. Atomic interactions are then governed by potentials with simple functional forms that provide an empirical, but physically motivated, approximation to the Born-Oppenheimer effective energy surface for nuclear dynamics.[517] The parameters for these potentials have been heavily optimized to provide reasonable accuracy in a wide range of environments.[518–524] These simulations then propagate atomic dynamics by numerically integrating classical equations of motion that are cleverly augmented to enforce appropriate thermodynamic conditions or to enhance efficiency.[525–527] Given the simplicity of these models and the severity of the associated approximations, the predictive power and basic molecular insight provided by atomistic MD simulations is truly remarkable.[5]

Nevertheless, atomically detailed simulations remain limited in accuracy, transferability, and scope. Recent advances in computational infrastructure, methodology, and resources[5,527–538] have made it possible to systematically quantify significant discrepancies between different force fields and between simulated and measured observables.[539–543] As a consequence of "coarse-graining" over electronic degrees of freedom, atomically detailed potentials are themselves effective potentials[544] that possess somewhat limited transferability, as demonstrated by the need for different parameter sets when modeling the same atom in different environments,[518–523] by the myriad water models,[545,546] by the increasing popularity of polarizable models,[547–549] and by the challenge in extending current force fields to model, e.g., carbohydrates.[550–552] Perhaps, though, the greatest limitation of atomically detailed simulations stems from their computational expense. Despite the rapid advances noted above, their computational expense limits atomically detailed simulations to relatively small systems, e.g., typically one or two protein domains, on relatively short timescales, e.g., typically on the order of 100–1000 ns.[5] In particular, this computational expense often precludes the sampling necessary for accurate computations of thermodynamic and dynamical quantities.[553] These limitations may become increasingly severe for future simulations of more complex systems, since the relevant time scales can grow very rapidly with system size and complexity.[57]

The computational expense of atomically detailed simulations strongly motivates the development of low-resolution CG models that provide much greater efficiency. By reducing the number of particles used in modeling a particular system, CG models significantly reduce the number of force calculations required per simulated time step, the computational memory associated with these calculations, and also the hard drive space required for each saved configuration. Because they result from averaging over atomic interactions, the interactions between CG sites can be of shorter range and, in some cases, can eliminate the need for explicit electrostatic interactions,[76,281] which further reduces the computational cost associated with each force calculation in CG simulations. The averaging that is implicit in CG models can lead to "softer" interactions, which permit a larger time step in numerically propagating motion.[166,554] This averaging can also reduce the activation barriers that are present in the potential landscape of the CG model, which leads to more diffusive motion through configuration space.[555–557] Furthermore, by eliminating explicit atomic interactions that generate "friction" to motion (i.e., the ruggedness in the potential surface), the environment in CG models is much less viscous, which further accelerates dynamical processes and sampling in the CG model.[208,209,558] (However, these effects also present significant challenges for interpreting CG dynamics because not all processes in the CG model are necessarily accelerated to the same extent.[557]) The net result of these considerations is that CG models can be three or more orders of magnitude more efficient than atomistic models.[4] These gains enable simulations of systems and phenomena that could not otherwise be studied, allow for quantitative comparisons with experiments, and also provide the efficiency necessary for achieving adequate convergence of free energies and dynamical quantities.[559]

Finally, in addition to these tremendous gains in computational efficiency, CG models also provide a powerful conceptual advantage by allowing the modeler to precisely tailor the model to address emergent phenomena on the relevant mesoscopic length- and time-scales, without invoking unnecessary atomic details.[57,67,190] In particular, top-down approaches harness this opportunity to great effect by designing elegant and highly efficient models to clearly demonstrate the remarkably nontrivial phenomena that can emerge from simple physical principles.[190] Consequently, despite great strides in computational power, CG models may well become an increasingly equal partner to atomically detailed models for molecular simulations.[560]

While CG models provide powerful computational and conceptual advantages, they also introduce new challenges for the modeler. When developing CG models to investigate

a particular system or phenomena, the single most significant challenge is to ensure that the model reflects the correct physics, i.e., the CG model should not only provide the correct answer, but also provide this answer for the correct reason.[6,18,561] Clearly, CG models may prove misleading unless their construction and results are carefully assessed. These potential dangers were emphatically demonstrated shortly after the first CG protein simulations[562] and significantly stymied subsequent development.[563]

In stark contrast to atomically detailed models, the sites in CG models correspond to relatively abstract chunks of matter that, in some cases, have significant internal flexibility. Consequently, the interactions that emerge in CG models are not simply energies of atomic interactions. Rather they are effective interactions that may reflect strong entropic contributions, which can be sensitive to the molecular environment and thermodynamic state point. These entropic effects may prove challenging to predict or accurately model, may limit the transferability of the effective interactions, and may also complicate the interpretation of thermodynamic properties in the CG model,[216,217] for instance, by altering the apparent balance between energy and entropy.[176]

Various coarse-graining strategies address this challenge in different ways and, consequently, demonstrate different strengths and limitations. Top-down models typically address the universal consequences of fundamental physical principles, such as excluded volume or hydrophobic association. Generic top-down models can sidestep the challenge of capturing the correct physics for a particular system by severing the connection to any atomic description of a particular system.[190] More detailed top-down models can be indirectly related to a particular system by comparison with experimental observations of emergent phenomena.[199] In particular, the Martini model extends the top-down paradigm by designing transferable potentials that attempt to capture the consequences of generic principles, such as hydrophobic mixing, but that can be related to specific systems.[200] Nevertheless, because they are not directly connected to any more detailed (and presumably more realistic) model, it may prove challenging to demonstrate that top-down models provide the right answer for the right reason. Moreover, while top-down models have contributed important insight into the universal significance of basic physical principles and often provide a reasonable description of thermodynamic properties, they have enjoyed only modest success in modeling the structure and fluctuations of complex biomolecules.[206]

By directly incorporating knowledge of experimentally determined structures, native structure-based methods have enjoyed considerable success in modeling the structure and fluctuations of complex biomolecular systems. Network models assume that shape governs these motions and that, in particular, the free energy cost of fluctuations is a quadratic function of the displacement from the folded native state.[223] These models have proven remarkably useful for modeling biologically relevant fluctuations and conformational transitions.[76] Similarly, native-centric Gō models[279] assume that the configuration-dependent free energy can be described by a fairly smooth funnel-shaped surface[283,286,288] that systematically decreases with similarity to the folded structure in such a way that all interactions are optimally favorable (i.e., mutually consistent[282]) in this structure.[281,290] These models are heavily employed for modeling protein folding and also the interactions in larger biomolecular complexes.[77,330] However, the physical significance of the resulting interactions remain quite unclear, since they are determined not by the physicochemical properties of the CG sites, but rather by the target native structure.[280] Moreover, and, from a practical perspective, much more importantly, these models cannot be used to predict unknown structures for new systems. Knowledge-based approaches address this limitation by either explicitly or implicitly employing information from a databank of known structures.[402] The resulting potentials are often quite physically motivated and have proven successful in folding studies, structure prediction, and many other areas of computational protein (and nucleic acid) science.[402,564] Nevertheless, their transferability remains unclear. Furthermore, because they often adopt uncontrolled heuristic approximations, it may prove challenging to systematically improve their accuracy.[389]

From this perspective, physics-based bottom-up approaches provide several advantages. Given a particular atomically detailed model (and subject to the limitations of this model), the many-body PMF quantitatively describes all equilibrium thermodynamic and structural properties that emerge on the length scales of the CG model.[210,211,213] Importantly, although it may be quite complex, this PMF always exists for any system and any mapping. Various correlation-based approaches have been developed to systematically approximate the PMF.[432] Quite recently, the MS-CG[502,503] and relative entropy[486] approaches have provided two robust, variational frameworks for computationally determining the exact PMF (to within an additive constant and subject to adequate information about the atomistic model[565]). When applied in practice, the MS-CG and relative entropy variational principles determine two distinct, well-defined approximations to the PMF that possess physical significance and that can be systematically improved.[435] Thus, bottom-up strategies hold tremendous promise for correctly describing the true physical forces that drive a particular process and for providing not only the computational efficiency, but also the quantitative accuracy that is necessary for predictive simulations of complex biomolecular systems. Moreover, rigorous bottom-up approaches will likely be essential for adaptive- and multi-resolution computational models that require quantitative compatibility between atomistic and CG models.[24,25,566,567]

IX. PERSPECTIVE: CHALLENGES FOR BOTTOM-UP MODELS

Despite their tremendous promise and remarkable success for liquids, polymers, and even biological membranes, bottom-up strategies have not yet achieved their full potential for modeling biomolecular systems with more complex structures. This section briefly suggests some remaining interrelated challenges, that, if addressed, would significantly improve the accuracy, scope, and, thus, the advantages of CG modeling for biomolecular systems.

A. Basis set

The interactions that are included in the CG potential and the functional forms that are adopted for describing these interactions determine a "basis set" for approximating the many-body PMF. Variational methods determine the CG potential that, within the subspace of potentials spanned by this basis, provides an "optimal" approximation to this PMF.[435] Of course, the accuracy of this approximation and, thus, also the accuracy of the CG model depend upon the adequacy of the basis set. Bottom-up CG methods typically approximate the PMF with molecular mechanics potentials that have proven suitable for liquids and polymers, but that may be inadequate for more complex biomolecular systems. It should be emphasized that, while molecular mechanics potentials provide great efficiency for simulations and are familiar from atomistic models,[517] they may not always be suitable for CG models.[69,568] Bottom-up approaches may greatly benefit from considering more complex potential terms to model, e.g., hydrogen-bonding, electrostatic interactions, solvation forces, and anisotropic interactions.[91,93,125,409,569–572] More generally, it would be highly desirable to develop an algorithm or framework for systematically identifying "missing" basis vectors that are computationally efficient to parameterize and simulate and that would also provide significantly improved descriptions of biomolecular structure, dynamics, and thermodynamics.

B. Mapping

The mapping significantly impacts structural, dynamic, and thermodynamic properties of a CG model. Unfortunately, as discussed in Sec. II, relatively little theoretical work has addressed its importance. However, it seems intuitively reasonable that the ability of a CG model to describe the correct physics governing a particular system fundamentally relies upon the CG model capturing the key physical features underlying this physics.[18] It also seems intuitively reasonable that "better" mappings will allow for a "simpler" description of this physics. Bottom-up methods would tremendously benefit from practical, rigorous algorithms that apply this intuition when developing CG mappings. More generally, fundamental insight into the following considerations would greatly advance coarse-graining methodologies:

- Dynamics: The mapping impacts the dynamics of a CG model by altering the molecular shape[49] and also the surrounding free volume in the system.[573] Moreover, the mapping determines a subensemble of atomistic initial conditions that map to a single initial condition in the CG model.[208] Since each of the initial conditions within this subensemble leads to a distinct atomistic trajectory, it is no longer trivial to relate the dynamics of atomistic and CG models. Similarly, the mapping also influences the "friction" and non-Markovian "memory" that arise when an atomistic trajectory is observed from the CG perspective.[574] Can one optimize the mapping to improve agreement with dynamical correlation functions and to simplify the relationship between atomistic and CG dynamics? As noted earlier, the extensive literature addressing the dynamics of CG models lies beyond the scope of this review.[209,575]
- Structure: Given an atomistic model, the mapping determines the PMF, which is the exactly correct potential for modeling structure and thermodynamics in the CG model. Variational methods provide "optimal approximations" to the PMF that can, in principle, be systematically improved by introducing more complex terms in the CG potential. In practice, though, the PMF is typically approximated with relatively simple functional forms (e.g., pair additive non-bonded potentials). Given this practical restriction upon the complexity of the approximate CG potential, the accuracy of the resulting "optimal approximation" depends upon the complexity of the PMF. Can one systematically design the mapping so that the model still describes the phenomena of interest, but so that the PMF can be well approximated by relatively simple functions?[48] Importantly, the accuracy and transferability of CG models do not always improve with increasing model resolution.[20,33]
- Thermodynamics: By mapping multiple atomistic configurations into a single configuration and by averaging over interactions, the process of coarse-graining incorporates entropic effects into the PMF.[216] This not only incorporates a system- and state-point dependence into the PMF, but also alters the apparent entropy and enthalpy present in the CG model.[176,435] Can one systematically optimize the mapping to enhance the fidelity of thermodynamic properties and minimize the sensitivity of the PMF to thermodynamic state point and system?

C. Assessment/optimization

As discussed in Sec. VII A, the many-body PMF is often approximated with simple potentials that are parameterized to reproduce the distributions observed in an atomistic model along the corresponding degrees of freedom. For instance, non-bonded pair potentials are often optimized to reproduce the corresponding pair distribution functions. (Of course, these pair distributions should not be used to validate the resulting model.) While these pair distribution functions may provide a useful description of structure in liquids, bilayers, and amorphous polymers, they may be relatively insensitive to the many-body global structural features of more complex biomolecules. Depending upon the CG mapping, an accurate description of local or low order structural properties may not guarantee an accurate description of global, higher-order structure, such as protein tertiary structure.[470] In particular, the CG model that employs pair potentials to reproduce a set of pair correlations will generate the ensemble of structures that has maximum entropy relative to any other ensemble that reproduces the given set of pair correlation functions.[480] However, the set of pair correlations that are employed in optimizing CG potentials may be insufficient

to determine the biologically relevant ensemble of structures, i.e., the biologically relevant ensemble may not be of maximum entropy relative to any other ensemble with the same pair correlations.[576]

Moreover, as noted in Sec. VII A, a given potential function, U_ζ, is often applied to model multiple instances, λ, of the same interaction. These instances may be chemically equivalent (e.g., the interactions between indistinguishable solvent molecules) or they may be chemically inequivalent (e.g., the pair nonbonded interactions between two pairs of alanine residues that are located at different positions in a protein). In this latter case, bottom-up, structure-based methods determine potentials that reproduce the distribution functions obtained from averaging over these inequivalent environments.[432,435] This averaging over environments may further deteriorate the agreement between the higher order structure observed in the atomistic and CG ensembles without impacting the agreement between, e.g., distance distribution functions describing pairs of alanine residues.

Consequently, structure-based bottom-up methods will significantly benefit from a fundamental understanding of the following considerations.

- How can global structural properties, such as tertiary protein structure, be incorporated into the parameterization of interactions in bottom-up approaches? While additional terms can be introduced into the CG potential to directly govern these global order parameters,[116,486] the resulting potential may demonstrate less transferability. One intriguing strategy would be to develop rigorous theories for incorporating low-resolution experimental data (i.e., top-down information) into the development of bottom-up CG models.
- How can important many-body effects and correlations be more efficiently identified, incorporated, approximated, and assessed in CG models? Quantitative analyses of energy landscapes[577] via, e.g., the inherent structure formalism[578] or energy disconnectivity graphs,[579,580] may provide a powerful framework for analyzing and improving the thermodynamic and many-body structural properties of CG models.
- Can one estimate *a priori* the accuracy of a CG model on the basis of the atomistic model and the CG mapping via, e.g., information theoretic approaches?[490]

D. Representability and thermodynamic consistency

As a consequence of averaging over atomic structures, the many-body PMF incorporates significant entropic effects from the "hidden" atomistic degrees of freedom. Consequently, thermodynamic properties cannot be represented in their conventional manner.[216,217,581] In essence, the expressions for thermodynamic observables are renormalized by the coarse-graining procedure. For instance, the configurational entropy is altered as a direct consequence of viewing the atomistic model from a lower dimensional configuration space and the associated averaging of the configuration distribution.[435] Similarly, the elimination of degrees of freedom requires a new treatment and analysis of the pressure. As a simple concrete example, if a CG model eliminates a fraction of molecules, this also eliminates a related fraction of the force applied to the system boundaries and thus reduces the pressure. Bottom-up strategies often directly modify the intermolecular potentials to account for this missing contribution to the pressure.[362,493,582] Das and Andersen have introduced a rigorous framework for modeling the pressure in CG models, but this strategy has not yet been demonstrated for complex biomolecular systems.[583] This suggests the following considerations:

- Can one formulate a consistent treatment of thermodynamic properties for effective CG potentials that is both accurate and practical?
- Can one accurately reproduce phase transitions and other thermodynamic properties in complex systems with bottom-up models[584] that are typically optimized to reproduce structural properties?

E. Transferability

CG models adopt a reduced representation that eliminates certain details from the description of molecular systems. The potential governing the degrees of freedom that are retained in the CG model must then capture the physical effects of the details that were eliminated. Because these details may vary from system to system and may have varying significance with changes in thermodynamic state or local environment, an effective potential that provides a reasonable description for modeling one particular system or thermodynamic state may not be transferable, i.e., it may not provide sufficient accuracy for modeling other systems or thermodynamic state points. The transferability of CG potentials may be limited with respect to changes in the molecular system (e.g., which molecules are present), thermodynamic state (e.g., changes in temperature, pressure, or even pH[585]), or even local environment (e.g., changes between bulk and interfacial regions of a system). Quite generally, transferability remains an outstanding issue in the field that is most commonly assessed and improved on the basis of relatively heuristic metrics.

In particular, as discussed earlier, the many-body PMF necessarily depends upon the system and thermodynamic state point for which it is defined. Similarly, a potential that is optimized to approximate the PMF will likely also depend upon system and thermodynamic state point. However, this state point dependence remains poorly understood. Although many studies have investigated the temperature dependence of potentials obtained from bottom-up methods, these approaches have not yet proved predictive or generally applicable.[150,586–594] Interestingly, CG pair potentials that are determined by considering isolated molecular fragments may provide improved transferability for liquid mixtures and amorphous polymers.[465,595,596] Alternatively, an extended ensemble framework[20] provides a rigorous framework for systematically optimizing transferable potentials to approximate the PMFs for multiple systems. Nevertheless, these

approaches have not yet been successfully demonstrated for more complex biomolecular systems.

Moreover, issues of transferability present challenges even for modeling different regions of space in a single inhomogeneous system. In principle, the many-body PMF should provide sufficient information for modeling such a system. However, it remains challenging to devise simple approximations to this PMF that provide adequate transferability for describing how intermolecular interactions vary between bulk and interfacial environments or how the local electrostatic environment changes between aqueous and membrane environments.[597,598] Two distinct approaches are to either consider both environments when constructing a single CG potential[481,599] or, alternatively, construct CG potentials that vary as a function of the local environment.[600,601]

This suggests the following considerations:

- Can one predict the state point dependence of the PMF?
- Can one estimate and incorporate the system- and state point-dependence of approximate potentials?
- Can one optimize approximate CG potentials for transferability to a range of thermodynamic state points or different systems? This challenge seems particularly daunting in the case of systems that are sensitive to the solvent conditions or for systems, such as proteins, with many distinct types of interactions.
- Can one, for complex biomolecular systems, develop practical, systematic means to parameterize accurate and efficient CG potentials that accurately treat changes in the local environment?

X. PERSPECTIVE: CONCLUDING THOUGHTS

The success of any modeling effort critically depends upon precisely defining the objective of the study and upon applying appropriate modeling tools. Significantly, CG models already provide simplified, yet predictive and powerful tools for many areas of quantitative science. Intuitively, the simplicity and accuracy of these models often relies upon an adiabatic separation between the scales that are addressed by the model and those that have been eliminated, e.g., the separation between atomic and macroscopic scales for thermodynamics or between atomic and mesoscopic scales for colloidal theories. From this perspective, the development of CG models for biomolecular systems presents new challenges. In many cases, CG models address biomolecular structure and dynamics with a resolution of a few Angstroms. In this case, it is not obvious that such a separation of time scales exists; the atomic interactions that are averaged over may be intimately coupled to the degrees of freedom that are treated in the CG model. Top-down models address this challenge by focusing on the emergent phenomena rather than on its microscopic origin. Nevertheless, it may prove challenging for top-down models to provide quantitative, predictive insight for specific systems and especially for those with complex biomolecular structure. In contrast, statistical mechanics and the PMF, in particular, provide an exact framework for addressing this challenge. Consequently, bottom-up modeling approaches that harness this framework hold great promise.

In closing, this perspective attempts to (1) summarize some leading approaches for coarse-graining biomolecular systems, (2) clarify their fundamental philosophies and assumptions, (3) illuminate their various inter-relationships, and (4) identify certain outstanding challenges in the field. In so doing, it is hoped that this review will equip and focus researchers to address these fundamental challenges, so that highly efficient CG models will realize their promise of providing not only basic, transparent insight, but also quantitatively accurate predictions for complex biomolecular systems.

ACKNOWLEDGMENTS

This work has been financially supported by a CAREER award (Grant No. MCB 1053970) from the National Science Foundation (NSF), an Alfred P. Sloan Research Fellowship, and a Camille Dreyfus Teacher-Scholar Award. W.G.N. gratefully acknowledges the support and hospitality of the Kavli Institute for Theoretical Physics (KITP) at the University of California, Santa Barbara, as well as many helpful conversations with participants in the KITP program on multiscale modeling of soft matter systems. W.G.N. particularly thanks Qiang Cui, Markus Deserno, Lasse Jensen, Ed O'Brien, Christine Peter, M. Scott Shell, and John Straub for many very helpful comments regarding the contents and style of this manuscript. W.G.N. also expresses his appreciation to Chris Ellis, Sushant Kumar, Nick Dunn, Tommy Foley, and especially J. Wayne Mullinax and Joe Rudzinski for helpful comments on this manuscript and for their various contributions leading to the perspective expressed herein. Finally, W.G.N. gratefully acknowledges Markus Deserno for many very formative discussions, especially regarding philosophies of coarse-graining, and John Straub for his patience and support in drafting this review. Figures 1, 2, and 4 were made with VMD. VMD is developed with NIH support by the Theoretical and Computational Biophysics group at the Beckman Institute, University of Illinois at Urbana-Champaign. Images of bonds between CG sites were introduced using scripts provided by the Marrink group.

[1]R. W. Hamming, *Numerical Methods for Scientists and Engineers* (McGraw-Hill, New York, 1962).
[2]P. W. Anderson, Science **177**, 393 (1972).
[3]*Coarse-Graining of Condensed Phase and Biomolecular Systems*, edited by G. A. Voth (CRC Press, Boca Raton, FL, 2008).
[4]D. Fritz, C. R. Herbers, K. Kremer, and N. F. A. van der Vegt, Soft Matter **5**, 4556 (2009).
[5]T. Schlick, R. Collepardo-Guevara, L. A. Halvorsen, S. Jung, and X. Xiao, Q. Rev. Biophys. **44**, 191 (2011).
[6]S. O. Nielsen, C. F. Lopez, G. Srinivas, and M. L. Klein, J. Phys.: Condens. Matter **16**, R481 (2004).
[7]G. S. Ayton, W. G. Noid, and G. A. Voth, Curr. Opin. Struct. Biol. **17**, 192 (2007).
[8]C. Clementi, Curr. Opin. Struct. Biol. **18**, 10 (2008).
[9]S. Matysiak and C. Clementi, Arch. Biochem. Biophys. **469**, 29 (2008).
[10]C. Peter, L. Delle Site, and K. Kremer, Soft Matter **4**, 859 (2008).
[11]P. Sherwood, B. R. Brooks, and M. S. P. Sansom, Curr. Opin. Struct. Biol. **18**, 630 (2008).
[12]T. Murtola, A. Bunker, I. Vattulainen, M. Deserno, and M. Karttunen, Phys. Chem. Chem. Phys. **11**, 1869 (2009).

[13]C. Peter and K. Kremer, Faraday Discuss. **144**, 9 (2010).
[14]J. Trylska, J. Phys.: Condens. Matter **22**, 453101 (2010).
[15]C. Hyeon and D. Thirumalai, Nat. Commun. **2**, 487 (2011).
[16]S. C. L. Kamerlin, S. Vicatos, A. Dryga, and A. Warshel, Annu. Rev. Phys. Chem. **62**, 41 (2011).
[17]S. Takada, Curr. Opin. Struct. Biol. **22**, 130 (2012).
[18]S. Riniker, J. R. Allison, and W. F. van Gunsteren, Phys. Chem. Chem. Phys. **14**, 12423 (2012).
[19]M. G. Saunders and G. A. Voth, Annu. Rev. Biophys. **42**, 73 (2013).
[20]J. W. Mullinax and W. G. Noid, J. Chem. Phys. **131**, 104110 (2009).
[21]Q. Shi, S. Izvekov, and G. A. Voth, J. Phys. Chem. B **110**, 15045 (2006).
[22]C. F. Abrams, J. Chem. Phys. **123**, 234101 (2005).
[23]M. Praprotnik, L. Delle Site, and K. Kremer, J. Chem. Phys. **123**, 224106 (2005).
[24]M. Praprotnik, L. Delle Site, and K. Kremer, Annu. Rev. Phys. Chem. **59**, 545 (2008).
[25]A. Heyden and D. G. Truhlar, J. Chem. Theory Comput. **4**, 217 (2008).
[26]B. Ensing, S. O. Nielsen, P. B. Moore, M. L. Klein, and M. Parrinello, J. Chem. Theory Comput. **3**, 1100 (2007).
[27]S. O. Nielsen, R. E. Bulo, P. B. Moore, and B. Ensing, Phys. Chem. Chem. Phys. **12**, 12401 (2010).
[28]T. Murtola, M. Karttunen, and I. Vattulainen, J. Chem. Phys. **131**, 055101 (2009).
[29]J. Zhang and M. Muthukumar, J. Chem. Phys. **130**, 035102 (2009).
[30]J. F. Dama, A. V. Sinitskiy, M. McCullagh, J. Weare, B. Roux, A. R. Dinner, and G. A. Voth, J. Chem. Theory Comput. **9**, 2466 (2013).
[31]J. G. Gay and B. J. Berne, J. Chem. Phys. **74**, 3316 (1981).
[32]M. Babadi, R. Everaers, and M. R. Ejtehadi, J. Chem. Phys. **124**, 174708 (2006).
[33]A. Das, L. Lu, H. C. Andersen, and G. A. Voth, J. Chem. Phys. **136**, 194115 (2012).
[34]W. Tschöp, K. Kremer, O. Hahn, J. Batoulis, and T. Bürger, Acta Polym. **49**, 75 (1998).
[35]A. A. Canutescu, A. A. Shelenkov, and R. L. Dunbrack, Jr., Protein Sci. **12**, 2001 (2003).
[36]J. Maupetit, R. Gautier, and P. Tufféry, Nucleic Acids Res. **34**, W147 (2006).
[37]P. Rotkiewicz and J. Skolnick, J. Comput. Chem. **29**, 1460 (2008).
[38]S. M. Gopal, S. Mukherjee, Y.-M. Cheng, and M. Feig, Proteins **78**, 1266 (2010).
[39]M. Maciejczyk, A. Spasic, A. Liwo, and H. A. Scheraga, J. Comput. Chem. **31**, 1644 (2010).
[40]G. S. Ayton and G. A. Voth, Curr. Opin. Struct. Biol. **19**, 138 (2009).
[41]M. G. Saunders and G. A. Voth, Curr. Opin. Struct. Biol. **22**, 144 (2012).
[42]H. Gohlke and M. F. Thorpe, Biophys. J. **91**, 2115 (2006).
[43]Z. Y. Zhang, L. Y. Lu, W. G. Noid, V. Krishna, J. Pfaendtner, and G. A. Voth, Biophys. J. **95**, 5073 (2008).
[44]A. V. Sinitskiy, M. G. Saunders, and G. A. Voth, J. Phys. Chem. B **116**, 8363 (2012).
[45]R. Potestio, F. Pontiggia, and C. Micheletti, Biophys. J. **96**, 4993 (2009).
[46]M. Stepanova, Phys. Rev. E **76**, 051918 (2007).
[47]S. Kundu, D. C. Sorensen, and G. N. Phillips, Proteins **57**, 725 (2004).
[48]V. A. Harmandaris, D. Reith, N. F. A. Van der Vegt, and K. Kremer, Macromol. Chem. Phys. **208**, 2109 (2007).
[49]A. Arkhipov, P. Freddolino, and K. Schulten, Structure **14**, 1767 (2006).
[50]A. Arkhipov, Y. Yin, and K. Schulten, Biophys. J. **95**, 2806 (2008).
[51]Z. Y. Zhang and G. A. Voth, J. Chem. Theory Comput. **6**, 2990 (2010).
[52]B. Smit, P. A. J. Hilbers, K. Esselink, L. A. M. Rupert, N. M. van Os, and A. G. Schlijper, Nature (London) **348**, 624 (1990).
[53]R. Goetz and R. Lipowsky, J. Chem. Phys. **108**, 7397 (1998).
[54]G. Brannigan, L. C.-L. Lin, and F. L. H. Brown, Eur. Biophys. J. **35**, 104 (2006).
[55]M. Muller, K. Katsov, and M. Schick, Phys. Rep. **434**, 113 (2006).
[56]M. Venturoli, M. M. Sperotto, M. Kranenburg, and B. Smit, Phys. Rep. **437**, 1 (2006).
[57]M. Deserno, Macromol. Rapid Commun. **30**, 752 (2009).
[58]S. J. Marrink, A. H. de Vries, and D. P. Tieleman, Biochim. Biophys. Act. **1788**, 149 (2009).
[59]V. Molinero and W. A. Goddard, "Molecular modeling of carbohydrates with no charges, no hydrogen bonds, and no atoms," in *NMR Spectroscopy and Computer Modeling of Carbohydrates*, ACS Symposium Series, Vol. 930 (ACS, 2006), Chap. 15, pp. 271–284.
[60]P. Liu, S. Izvekov, and G. A. Voth, J. Phys. Chem. B **111**, 11566 (2007).
[61]C. A. Lopez, A. J. Rzepiela, A. H. de Vries, L. Dijkhuizen, P. H. Hunenberger, and S. J. Marrink, J. Chem. Theory Comput. **5**, 3195 (2009).
[62]H. M. Cho, A. S. Gross, and J.-W. Chu, J. Am. Chem. Soc. **133**, 14033 (2011).
[63]A.-P. Hynninen, J. F. Matthews, G. T. Beckham, M. F. Crowley, and M. R. Nimlos, J. Chem. Theory Comput. **7**, 2137 (2011).
[64]J. Wohlert and L. A. Berglund, J. Chem. Theory Comput. **7**, 753 (2011).
[65]G. Srinivas, X. Cheng, and J. C. Smith, J. Chem. Theory Comput. **7**, 2539 (2011).
[66]S. Markutsya, Y. A. Kholod, A. Devarajan, T. L. Windus, M. S. Gordon, and M. H. Lamm, Theor. Chem. Acc. **131**, 1162 (2012).
[67]M. Levitt and A. Warshel, Nature (London) **253**, 694 (1975).
[68]M. Levitt, J. Mol. Biol. **104**, 59 (1976).
[69]V. Tozzini, W. Rocchia, and J. A. McCammon, J. Chem. Theory Comput. **2**, 667 (2006).
[70]J. D. Honeycutt and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **87**, 3526 (1990).
[71]S. Brown, N. J. Fawzi, and T. Head-Gordon, Proc. Natl. Acad. Sci. U.S.A. **100**, 10712 (2003).
[72]D. Thirumalai and D. K. Klimov, Curr. Opin. Struct. Biol. **9**, 197 (1999).
[73]J. M. Sorenson and T. Head-Gordon, Proteins **46**, 368 (2002).
[74]M. Friedel, D. J. Sheeler, and J.-E. Shea, J. Chem. Phys. **118**, 8106 (2003).
[75]M. Enciso and A. Rey, J. Chem. Phys. **136**, 215103 (2012).
[76]I. Bahar, T. R. Lezon, L. W. Yang, and E. Eyal, Annu. Rev. Biophys. **39**, 23 (2010).
[77]P. C. Whitford, K. Y. Sanbonmatsu, and J. N. Onuchic, Rep. Prog. Phys. **75**, 076601 (2012).
[78]A. Rey and J. Skolnick, Chem. Phys. **158**, 199 (1991).
[79]B. M. Messer, M. Roca, Z. T. Chu, S. Vicatos, A. V. Kilshtain, and A. Warshel, Proteins **78**, 1212 (2010).
[80]A. Liwo, S. Oldziej, M. R. Pincus, R. J. Wawak, S. Rackovsky, and H. A. Scheraga, J. Comput. Chem. **18**, 849 (1997).
[81]A. Liwo, C. Czaplewski, J. Pillardy, and H. A. Scheraga, J. Chem. Phys. **115**, 2323 (2001).
[82]A. Liwo, Y. He, and H. A. Scheraga, Phys. Chem. Chem. Phys. **13**, 16890 (2011).
[83]M. Zacharias, Protein Sci. **12**, 1271 (2003).
[84]N. Basdevant, D. Borgis, and T. Ha-Duong, J. Phys. Chem. B **111**, 9390 (2007).
[85]S. Takada, Z. Luthey-Schulten, and P. G. Wolynes, J. Chem. Phys. **110**, 11616 (1999).
[86]A. Irbäck, F. Sjunnesson, and S. Wallin, Proc. Natl. Acad. Sci. U.S.A. **97**, 13614 (2000).
[87]C. Guo, M. S. Cheung, H. Levine, and D. A. Kessler, J. Chem. Phys. **116**, 4353 (2002).
[88]H. D. Nguyen and C. K. Hall, Proc. Natl. Acad. Sci. U.S.A. **101**, 16180 (2004).
[89]A. E. van Giessen and J. E. Straub, J. Chem. Phys. **122**, 024904 (2005).
[90]F. Ding, S. V. Buldyrev, and N. V. Dokholyan, Biophys. J. **88**, 147 (2005).
[91]N. Y. Chen, Z. Y. Su, and C. Y. Mou, Phys. Rev. Lett. **96**, 078103 (2006).
[92]Y. Mu and Y. Q. Gao, J. Chem. Phys. **127**, 105102 (2007).
[93]T. Bereau and M. Deserno, J. Chem. Phys. **130**, 235106 (2009).
[94]M. Cheon, I. Chang, and C. K. Hall, Proteins **78**, 2950 (2010).
[95]Y. Chebaro, S. Pasquali, and P. Derreumaux, J. Phys. Chem. B **116**, 8741 (2012).
[96]P. Doruker, R. L. Jernigan, and I. Bahar, J. Comput. Chem. **23**, 119 (2002).
[97]J.-W. Chu and G. A. Voth, Biophys. J. **90**, 1572 (2006).
[98]M. Hagan and D. Chandler, Biophys. J. **91**, 42 (2006).
[99]H. D. Nguyen, V. S. Reddy, and C. L. Brooks III, Nano Lett. **7**, 338 (2007).
[100]Q. Wang and M. S. Cheung, Biophys. J. **102**, 2353 (2012).
[101]D. Sept and F. C. MacKintosh, Phys. Rev. Lett. **104**, 018101 (2010).
[102]T. Head-Gordon and S. Brown, Curr. Opin. Struct. Biol. **13**, 160 (2003).
[103]V. Tozzini, Curr. Opin. Struct. Biol. **15**, 144 (2005).
[104]V. Tozzini, Quart. Rev. Biophys. **43**, 333 (2010).
[105]C. Wu and J.-E. Shea, Curr. Opin. Struct. Biol. **21**, 209 (2011).
[106]M. Rubinstein and R. Colby, *Polymer Physics* (Oxford University Press, 2003).
[107]T. Schlick, Curr. Opin. Struct. Biol. **5**, 245 (1995).
[108]W. K. Olson, Curr. Opin. Struct. Biol. **6**, 242 (1996).
[109]W. K. Olson and V. B. Zhurkin, Curr. Opin. Struct. Biol. **10**, 286 (2000).
[110]H. L. Tepper and G. A. Voth, J. Chem. Phys. **122**, 124906 (2005).
[111]E. Villa, A. Balaeff, L. Mahadevan, and K. Schulten, Multiscale Model. Simul. **2**, 527 (2004).
[112]F. W. Starr and F. Sciortino, J. Phys.: Condens. Matter **18**, L347 (2006).

[113]F. Trovato and V. Tozzini, J. Phys. Chem. B **112**, 13197 (2008).
[114]K. Doi, T. Haga, H. Shintaku, and S. Kawano, Philos. Trans. R. Soc. London, Ser. A **368**, 2615 (2010).
[115]A. Savelyev and G. A. Papoian, Biophys. J. **96**, 4044 (2009).
[116]A. Savelyev and G. A. Papoian, Proc. Natl. Acad. Sci. U.S.A. **107**, 20340 (2010).
[117]M. Sayar, B. c. Avşaroğlu, and A. Kabakçıoğlu, Phys. Rev. E **81**, 041916 (2010).
[118]K. Drukker, G. Wu, and G. C. Schatz, J. Chem. Phys. **114**, 579 (2001).
[119]M. Sales-Pardo, R. Guimerà, A. A. Moreira, J. Widom, and L. A. N. Amaral, Phys. Rev. E **71**, 051902 (2005).
[120]E. Sambriski, D. Schwartz, and J. de Pablo, Biophys. J. **96**, 1675 (2009).
[121]T. E. Ouldridge, A. A. Louis, and J. P. K. Doye, Phys. Rev. Lett. **104**, 178101 (2010).
[122]T. E. Ouldridge, A. A. Louis, and J. P. K. Doye, J. Chem. Phys. **134**, 085101 (2011).
[123]T. A. Knotts, IV, N. Rathore, D. C. Schwartz, and J. J. de Pablo, J. Chem. Phys. **126**, 084901 (2007).
[124]M. C. Linak, R. Tourdot, and K. D. Dorfman, J. Chem. Phys. **135**, 205102 (2011).
[125]A. Morriss-Andrews, J. Rottler, and S. S. Plotkin, J. Chem. Phys. **132**, 035105 (2010).
[126]B. Mergell, M. R. Ejtehadi, and R. Everaers, Phys. Rev. E **68**, 021911 (2003).
[127]P. Sulc, F. Romano, T. E. Ouldridge, L. Rovigatti, J. P. K. Doye, and A. A. Louis, J. Chem. Phys. **137**, 135101 (2012).
[128]F. Zhang and M. A. Collins, Phys. Rev. E **52**, 4217 (1995).
[129]S. Khalid, P. J. Bond, J. Holyoake, R. W. Hawtin, and M. S. Sansom, J. R. Soc., Interface **5**, 241 (2008).
[130]P. D. Dans, A. Zeida, M. R. Machado, and S. Pantano, J. Chem. Theory Comput. **6**, 1711 (2010).
[131]P. Poulain, A. Saladin, B. Hartmann, and C. Prévost, J. Comput. Chem. **29**, 2582 (2008).
[132]P. Setny, R. P. Bahadur, and M. Zacharias, BMC Bioinf. **13**, 228 (2012).
[133]C. Hyeon and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **102**, 6789 (2005).
[134]M. Paliy, R. Melnik, and B. A. Shapiro, Phys. Biol. **7**, 036001 (2010).
[135]F. Ding, S. Sharma, P. Chalasani, V. V. Demidov, N. E. Broude, and N. V. Dokholyan, RNA **14**, 1164 (2008).
[136]Z. Xia, D. P. Gardner, R. R. Gutell, and P. Ren, J. Phys. Chem. B **114**, 13497 (2010).
[137]S. Pasquali and P. Derreumaux, J. Phys. Chem. B **114**, 11957 (2010).
[138]P. Setny and M. Zacharias, Nucleic Acids Res. **39**, 9118 (2011).
[139]M. A. Jonikas, R. J. Radmer, A. Laederach, R. Das, S. Pearlman, D. Herschlag, and R. B. Altman, RNA **15**, 189 (2009).
[140]C. Hyeon and D. Thirumalai, Biophys. J. **92**, 731 (2007).
[141]S. C. Harvey, A. S. Petrov, B. Devkota, and M. B. Boz, Methods Enzymol. **487**, 513 (2011).
[142]J. J. de Pablo, Annu. Rev. Phys. Chem. **62**, 555 (2011).
[143]D. A. Potoyan, A. Savelyev, and G. A. Papoian, Wiley Interdiscip. Rev. Comput. Mol. Sci. **3**, 69 (2013).
[144]Z. Xia and P. Ren, "Prediction and coarse-grained modeling of RNA structures," *Biophysics of RNA Folding* (Springer, 2013), Chap. 4.
[145]K. A. Dill, Biochemistry **29**, 7133 (1990).
[146]K. M. Guckian, B. A. Schweitzer, R. X.-F. Ren, C. J. Sheils, D. C. Tahmassebi, and E. T. Kool, J. Am. Chem. Soc. **122**, 2213 (2000).
[147]T. Head-Gordon and F. H. Stillinger, J. Chem. Phys. **98**, 3313 (1993).
[148]S. Garde and H. S. Ashbaugh, J. Chem. Phys. **115**, 977 (2001).
[149]R. M. Lynden-Bell and T. Head-Gordon, Mol. Phys. **104**, 3593 (2006).
[150]M. E. Johnson, T. Head-Gordon, and A. A. Louis, J. Chem. Phys. **126**, 144509 (2007).
[151]A. Chaimovich and M. S. Shell, Phys. Chem. Chem. Phys. **11**, 1901 (2009).
[152]H. Wang, C. Junghans, and K. Kremer, Eur. Phys. J. E **28**, 221 (2009).
[153]Y. Liu and T. Ichiye, J. Phys. Chem. **100**, 2723 (1996).
[154]M.-L. Tan, J. T. Fischer, A. Chandra, B. R. Brooks, and T. Ichiye, Chem. Phys. Lett. **376**, 646 (2003).
[155]V. Molinero and E. B. Moore, J. Phys. Chem. B **113**, 4008 (2009).
[156]E. B. Moore and V. Molinero, Nature (London) **479**, 506 (2011).
[157]F. H. Stillinger and T. A. Weber, Phys. Rev. B **31**, 5262 (1985).
[158]M. Orsi, J. Michel, and J. W. Essex, J. Phys.: Condens. Matter **22**, 155106 (2010).
[159]R. C. DeMille, T. E. Cheatham, and V. Molinero, J. Phys. Chem. B **115**, 132 (2011).
[160]L. Darré, M. R. Machado, P. D. Dans, F. E. Herrera, and S. Pantano, J. Chem. Theory Comput. **6**, 3793 (2010).
[161]T. HaDuong, S. Phan, M. Marchi, and D. Borgis, J. Chem. Phys. **117**, 541 (2002).
[162]J. C. Shelley, M. Y. Shelley, R. C. Reeder, S. Bandyopadhyay, and M. L. Klein, J. Phys. Chem. B **105**, 4464 (2001).
[163]S. J. Marrink, A. H. de Vries, and A. E. Mark, J. Phys. Chem. B **108**, 750 (2004).
[164]B. van Hoof, A. J. Markvoort, R. A. van Santen, and P. A. J. Hilbers, J. Phys. Chem. B **115**, 10001 (2011).
[165]K. R. Hadley and C. McCabe, J. Phys. Chem. B **114**, 4590 (2010).
[166]S. J. Marrink, X. Periole, D. P. Tieleman, and A. H. de Vries, Phys. Chem. Chem. Phys. **12**, 2254 (2010).
[167]E. G. Flekkøy and P. V. Coveney, Phys. Rev. Lett. **83**, 1775 (1999).
[168]A. Eriksson, M. N. Jacobi, J. Nyström, and K. Tunstrøm, Europhys. Lett. **86**, 44001 (2009).
[169]H. Bock, K. E. Gubbins, and S. H. L. Klapp, Phys. Rev. Lett. **98**, 267801 (2007).
[170]M. Winger, D. Trzesniak, R. Baron, and W. F. van Gunsteren, Phys. Chem. Chem. Phys. **11**, 1934 (2009).
[171]S. O. Yesylevskyy, L. V. Schafer, D. Sengupta, and S. J. Marrink, PLOS Comput. Biol. **6**, e1000810 (2010).
[172]S. Riniker and W. F. van Gunsteren, J. Chem. Phys. **134**, 084110 (2011).
[173]M. Masella, D. Borgis, and P. Cuniasse, J. Comput. Chem. **29**, 1707 (2008).
[174]M. Masella, D. Borgis, and P. Cuniasse, J. Comput. Chem. **32**, 2664 (2011).
[175]Z. Wu, Q. A. Cui, and A. Yethiraj, J. Phys. Chem. B **114**, 10524 (2010).
[176]Z. Wu, Q. Cui, and A. Yethiraj, J. Phys. Chem. Lett. **2**, 1794 (2011).
[177]R. Baron and V. Molinero, J. Chem. Theory Comput. **8**, 3696 (2012).
[178]X. Chu, Y. Wang, L. Gan, Y. Bai, W. Han, E. Wang, and J. Wang, PLOS Comput. Biol. **8**, e1002608 (2012).
[179]Y. Levy, J. N. Onuchic, and P. G. Wolynes, J. Am. Chem. Soc. **129**, 738 (2007).
[180]I. R. Cooke, K. Kremer, and M. Deserno, Phys. Rev. E **72**, 011506 (2005).
[181]S. Izvekov and G. A. Voth, J. Phys. Chem. B **113**, 4443 (2009).
[182]Z. J. Wang and M. Deserno, J. Phys. Chem. B **114**, 11207 (2010).
[183]M. Makowski, A. Liwo, and H. A. Scheraga, J. Phys. Chem. B **111**, 2910 (2007).
[184]J. Chen, C. L. Brooks III, and J. Khandogin, Curr. Opin. Struct. Biol. **18**, 140 (2008).
[185]P. Grochowski and J. Trylska, Biopolymers **89**, 93 (2008).
[186]L. Darré, M. R. Machado, and S. Pantano, Wiley Interdiscip. Rev. Comput. Mol. Sci. **2**, 921 (2012).
[187]K. R. Hadley and C. McCabe, Mol. Simul. **38**, 671 (2012).
[188]K. A. Dill, T. M. Truskett, V. Vlachy, and B. Hribar-Lee, Annu. Rev. Biophys. Biomol. Struct. **34**, 173 (2005).
[189]D. Chandler, Nature (London) **437**, 640 (2005).
[190]F. Schmid, Macromol. Rapid Commun. **30**, 741 (2009).
[191]H. B. Callen, *Thermodynamics and an Introduction to Thermostatistics* (Wiley, 1985).
[192]J. Drouffe, A. Maggs, and S. Leibler, Science **254**, 1353 (1991).
[193]F. A. Detcheverry, D. Q. Pike, U. Nagpal, P. F. Nealey, and J. J. de Pablo, Soft Matter **5**, 4858 (2009).
[194]M. Homberg and M. Muller, J. Chem. Phys. **132**, 155104 (2010).
[195]R. Pellarin, E. Guarnera, and A. Caflisch, J. Mol. Biol. **374**, 917 (2007).
[196]G. Bellesia and J.-E. Shea, J. Chem. Phys. **130**, 145103 (2009).
[197]H. S. Chan and K. A. Dill, Proc. Natl. Acad. Sci. U.S.A. **87**, 6388 (1990).
[198]M. J. Stevens, Biophys. J. **80**, 130 (2001).
[199]W. Shinoda, R. Devane, and M. L. Klein, Mol. Simul. **33**, 27 (2007).
[200]S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries, J. Phys. Chem. B **111**, 7812 (2007).
[201]D. H. de Jong, G. Singh, W. F. D. Bennett, C. Arnarez, T. A. Wassenaar, L. V. Schäfer, X. Periole, D. P. Tieleman, and S. J. Marrink, J. Chem. Theory Comput. **9**, 687 (2013).
[202]S. J. Marrink and D. P. Tieleman, Chem. Soc. Rev. **42**, 6801–6822 (2013).
[203]L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S.-J. Marrink, J. Chem. Theory Comput. **4**, 819 (2008).
[204]K. A. Maerzke and J. I. Siepmann, J. Phys. Chem. B **115**, 3452 (2011).
[205]M. Seo, S. Rauscher, R. Pomès, and D. P. Tieleman, J. Chem. Theory Comput. **8**, 1774 (2012).
[206]X. Periole, M. Cavalli, S. J. Marrink, and M. A. Ceruso, J. Chem. Theory Comput. **5**, 2531 (2009).

[207]R. DeVane, W. Shinoda, P. B. Moore, and M. L. Klein, J. Chem. Theory Comput. **5**, 2115 (2009).
[208]R. Zwanzig, Phys. Rev. **124**, 983 (1961).
[209]C. Hijón, P. Español, E. Vanden-Eijnden, and R. Delgado-Buscalioni, Faraday Discuss. **144**, 301 (2010).
[210]J. G. Kirkwood, J. Chem. Phys. **3**, 300 (1935).
[211]A. Liwo, R. Kaźmierkiewicz, C. Czaplewski, M. Groth, S. Ołdziej, R. J. Wawak, S. Rackovsky, M. R. Pincus, and H. A. Scheraga, J. Comput. Chem. **19**, 259 (1998).
[212]C. N. Likos, Phys. Rep. **348**, 267 (2001).
[213]W. G. Noid, J.-W. Chu, G. S. Ayton, V. Krishna, S. Izvekov, G. A. Voth, A. Das, and H. C. Andersen, J. Chem. Phys. **128**, 244114 (2008).
[214]R. L. C. Akkermans and W. J. Briels, J. Chem. Phys. **113**, 6409 (2000).
[215]H. S. Chan and K. A. Dill, Annu. Rev. Biophys. Biophys. Chem. **20**, 447 (1991).
[216]A. A. Louis, J. Phys.: Condens. Matter **14**, 9187 (2002).
[217]F. H. Stillinger, H. Sakai, and S. Torquato, J. Chem. Phys. **117**, 288 (2002).
[218]*Free Energy Calculations: Theory and Applications in Chemistry and Biology*, edited by C. Chipot and A. Pohorille (Springer, 2007), Vol. 86.
[219]C. D. Christ, A. E. Mark, and W. F. van Gunsteren, J. Comput. Chem. **31**, 1569 (2010).
[220]R. K. Pathria, *Statistical Mechanics*, 2nd ed. (Elsevier, 1996).
[221]H. Berman, K. Henrick, and H. Nakamura, Nat. Struct. Mol. Biol. **10**, 980 (2003).
[222]S. Tanaka and H. A. Scheraga, Macromolecules **9**, 945 (1976).
[223]M. M. Tirion, Phys. Rev. Lett. **77**, 1905 (1996).
[224]K. Hinsen, Proteins **33**, 417 (1998).
[225]A. Atilgan, S. Durell, R. Jernigan, M. Demirel, O. Keskin, and I. Bahar, Biophys. J. **80**, 505 (2001).
[226]P. J. Flory, M. Gordon, and N. G. McCrum, Proc. R. Soc. London, Ser. A **351**, 351 (1976).
[227]I. Bahar, A. R. Atilgan, and B. Erman, Folding Des. **2**, 173 (1997).
[228]S. Kundu, J. S. Melton, D. C. Sorensen, and G. N. Phillips, Biophys. J. **83**, 723 (2002).
[229]D. Riccardi, Q. Cui, and G. N. Phillips, Biophys. J. **96**, 464 (2009).
[230]A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, Proteins **17**, 412 (1993).
[231]M. Rueda, P. Chacón, and M. Orozco, Structure **15**, 565 (2007).
[232]Y. Suezaki and N. Go, Int. J. Pept. Protein Res. **7**, 333 (1975).
[233]B. Halle, Proc. Natl. Acad. Sci. U.S.A. **99**, 1274 (2002).
[234]M. Lu and J. Ma, Biophys. J. **89**, 2395 (2005).
[235]K. Hinsen, A.-J. Petrescu, S. Dellerue, M.-C. Bellissent-Funel, and G. R. Kneller, Chem. Phys. **261**, 25 (2000).
[236]J. A. Kovacs, P. Chacón, and R. Abagyan, Proteins **56**, 661 (2004).
[237]K. Hinsen, Proc. Natl. Acad. Sci. U.S.A. **106**, E128 (2009).
[238]J.-W. Chu and G. A. Voth, Proc. Natl. Acad. Sci. U.S.A. **102**, 13111 (2005).
[239]K. Moritsugu and J. C. Smith, Biophys. J. **93**, 3460 (2007).
[240]T. R. Lezon and I. Bahar, PLOS Comput. Biol. **6**, e1000816 (2010).
[241]T. Z. Sen, Y. P. Feng, J. V. Garcia, A. Kloczkowski, and R. L. Jernigan, J. Chem. Theory Comput. **2**, 696 (2006).
[242]O. Kurkcuoglu, O. T. Turgut, S. Cansu, R. L. Jernigan, and P. Doruker, Biophys. J. **97**, 1178 (2009).
[243]K. Eom, S.-C. Baek, J.-H. Ahn, and S. Na, J. Comput. Chem. **28**, 1400 (2007).
[244]E. Lyman, J. Pfaendtner, and G. A. Voth, Biophys. J. **95**, 4183 (2008).
[245]C. Chennubhotla and I. Bahar, J. Comput. Biol. **14**, 765 (2007).
[246]M. K. Kim, G. S. Chirikjian, and R. L. Jernigan, J. Mol. Graphics Modell. **21**, 151 (2002).
[247]O. Miyashita, J. N. Onuchic, and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **100**, 12570 (2003).
[248]P. Maragakis and M. Karplus, J. Mol. Biol. **352**, 807 (2005).
[249]J. W. Chu and G. A. Voth, Biophys. J. **93**, 3860 (2007).
[250]Z. Yang, P. Májek, and I. Bahar, PLOS Comput. Biol. **5**, e1000360 (2009).
[251]W. J. Zheng, B. R. Brooks, and G. Hummer, Proteins **69**, 43 (2007).
[252]R. Soheilifard, D. E. Makarov, and G. J. Rodin, Phys. Biol. **5**, 026008 (2008).
[253]M. Lu and J. Ma, Proc. Natl. Acad. Sci. U.S.A. **105**, 15358 (2008).
[254]G. Song and R. L. Jernigan, J. Mol. Biol. **369**, 880 (2007).
[255]K. Hinsen, Bioinformatics **24**, 521 (2008).
[256]J. Hafner and W. Zheng, J. Chem. Phys. **132**, 014111 (2010).
[257]W. Zheng and B. R. Brooks, Biophys. J. **89**, 167 (2005).
[258]F. Tama, O. Miyashita, and C. L. Brooks III, J. Mol. Biol. **337**, 985 (2004).
[259]C. Gorba, O. Miyashita, and F. Tama, Biophys. J. **94**, 1589 (2008).
[260]B. K. Poon, X. Chen, M. Lu, N. K. Vyas, F. A. Quiocho, Q. Wang, and J. Ma, Proc. Natl. Acad. Sci. U.S.A. **104**, 7869 (2007).
[261]F. Tama and C. L. Brooks III, J. Mol. Biol. **345**, 299 (2005).
[262]W. Zheng, B. R. Brooks, and D. Thirumalai, Biophys. J. **93**, 2289 (2007).
[263]R. Tehver, J. Chen, and D. Thirumalai, J. Mol. Biol. **387**, 390 (2009).
[264]D. Ming and M. E. Wall, Phys. Rev. Lett. **95**, 198103 (2005).
[265]Y. Wang, A. Rader, I. Bahar, and R. L. Jernigan, J. Struct. Biol. **147**, 302 (2004).
[266]F. Tama, M. Valle, J. Frank, and C. L. Brooks III, Proc. Natl. Acad. Sci. U.S.A. **100**, 9319 (2003).
[267]F. Tama and Y.-H. Sanejouand, Protein Eng. **14**, 1 (2001).
[268]W. Zheng and S. Doniach, Proc. Natl. Acad. Sci. U.S.A. **100**, 13253 (2003).
[269]C. N. Cavasotto, J. A. Kovacs, and R. A. Abagyan, J. Am. Chem. Soc. **127**, 9632 (2005).
[270]A. May and M. Zacharias, Proteins **70**, 794 (2008).
[271]D. Schneidman-Duhovny, R. Nussinov, and H. J. Wolfson, Proteins **69**, 764 (2007).
[272]K. Suhre and Y.-H. Sanejouand, Nucleic Acids Res. **32**, W610 (2004).
[273]E. Eyal, L.-W. Yang, and I. Bahar, Bioinformatics **22**, 2619 (2006).
[274]E. Lindahl, C. Azuara, P. Koehl, and M. Delarue, Nucleic Acids Res. **34**, W52 (2006).
[275]M. T. Zimmermann, A. Kloczkowski, and R. L. Jernigan, BMC Bioinf. **12**, 264 (2011).
[276]I. Bahar, T. R. Lezon, A. Bakan, and I. H. Shrivastava, Chem. Rev. **110**, 1463 (2010).
[277]F. Tama and C. L. Brooks, Annu. Rev. Biophys. Biomol. Struct. **35**, 115 (2006).
[278]L.-W. Yang and C.-P. Chng, Bioinf. Biol. Insight **2**, 25 (2008).
[279]H. Taketomi, Y. Ueda, and N. Go, Int. J. Pept. Protein Res. **7**, 445 (1975).
[280]H. Kaya and H. Chan, J. Mol. Biol. **326**, 911 (2003).
[281]H. Nymeyer, A. Garcia, and J. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **95**, 5921 (1998).
[282]N. Go, Annu. Rev. Biophys. Bioeng. **12**, 183 (1983).
[283]J. Bryngelson and P. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **84**, 7524 (1987).
[284]J. Bryngelson, J. Onuchic, N. Socci, and P. Wolynes, Proteins **21**, 167 (1995).
[285]B. C. Gin, J. P. Garrahan, and P. L. Geissler, J. Mol. Biol. **392**, 1303 (2009).
[286]P. E. Leopold, M. Montal, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **89**, 8721 (1992).
[287]N. D. Socci, J. N. Onuchic, and P. G. Wolynes, J. Chem. Phys. **104**, 5860 (1996).
[288]K. A. Dill and H. S. Chan, Nat. Struct. Biol. **4**, 10 (1997).
[289]C. Clementi, P. Jennings, and J. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **97**, 5871 (2000).
[290]C. Clementi, H. Nymeyer, and J. Onuchic, J. Mol. Biol. **298**, 937 (2000).
[291]N. Koga and S. Takada, J. Mol. Biol. **313**, 171 (2001).
[292]J.-E. Shea, J. N. Onuchic, and C. L. Brooks III, Proc. Natl. Acad. Sci. U.S.A. **99**, 16064 (2002).
[293]T. X. Hoang and M. Cieplak, J. Chem. Phys. **113**, 8319 (2000).
[294]S. Takada, Proc. Natl. Acad. Sci. U.S.A. **96**, 11698 (1999).
[295]L. Wroblewska, A. Jagielska, and J. Skolnick, Biophys. J. **94**, 3227 (2008).
[296]A. Raval, S. Piana, M. P. Eastwood, R. O. Dror, and D. E. Shaw, Proteins **80**, 2071 (2012).
[297]K. W. Plaxco, K. T. Simons, and D. Baker, J. Mol. Biol. **277**, 985 (1998).
[298]K. Gunasekaran, S. Eyles, A. Hagler, and L. Gierasch, Curr. Opin. Struct. Biol. **11**, 83 (2001).
[299]E. Miller, K. Fischer, and S. Marqusee, Proc. Natl. Acad. Sci. U.S.A. **99**, 10359 (2002).
[300]L. Mirny and E. Shakhnovich, Annu. Rev. Biophys. Biomol. Struct. **30**, 361 (2001).
[301]J. Shimada and E. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **99**, 11175 (2002).
[302]C. Clementi, A. Garcia, and J. Onuchic, J. Mol. Biol. **326**, 933 (2003).
[303]J. Karanicolas and C. Brooks III, Protein Sci. **11**, 2351 (2002).
[304]M. Cheung, A. Garcia, and J. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **99**, 685 (2002).
[305]C. Clementi and S. S. Plotkin, Protein Sci. **13**, 1750 (2004).
[306]P. Das, S. Matysiak, and C. Clementi, Proc. Natl. Acad. Sci. U.S.A. **102**, 10141 (2005).
[307]Z. Zhang and H. S. Chan, Proc. Natl. Acad. Sci. U.S.A. **107**, 2920 (2010).

[308] M. R. Ejtehadi, S. P. Avall, and S. S. Plotkin, Proc. Natl. Acad. Sci. U.S.A. **101**, 15088 (2004).
[309] A. Badasyan, Z. Liu, and H. S. Chan, J. Mol. Biol. **384**, 512 (2008).
[310] D. Zuckerman, J. Phys. Chem. B **108**, 5127 (2004).
[311] R. Best, Y. Chen, and G. Hummer, Structure **13**, 1755 (2005).
[312] K.-i. Okazaki, N. Koga, S. Takada, J. N. Onuchic, and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **103**, 11844 (2006).
[313] E. P. O'Brien, G. Ziv, G. Haran, B. R. Brooks, and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **105**, 13403 (2008).
[314] A. Linhananta, S. Hadizadeh, and S. S. Plotkin, Biophys. J. **100**, 459 (2011).
[315] Z. Liu, G. Reddy, and D. Thirumalai, J. Phys. Chem. B **116**, 6707 (2012).
[316] S. Matysiak and C. Clementi, J. Mol. Biol. **343**, 235 (2004).
[317] C. Hyeon, G. H. Lorimer, and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **103**, 18939 (2006).
[318] N. Koga, T. Kameda, K.-i. Okazaki, and S. Takada, Proc. Natl. Acad. Sci. U.S.A. **106**, 18237 (2009).
[319] F. Takagi, N. Koga, and S. Takada, Proc. Natl. Acad. Sci. U.S.A. **100**, 11367 (2003).
[320] D. K. Klimov, D. Newfield, and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **99**, 8019 (2002).
[321] M. S. Cheung, D. Klimov, and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **102**, 4753 (2005).
[322] C. Hyeon, R. I. Dima, and D. Thirumalai, Structure **14**, 1633 (2006).
[323] J. I. Sułkowska and M. Cieplak, Biophys. J. **95**, 3174 (2008).
[324] Y. Levy, P. Wolynes, and J. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **101**, 511 (2004).
[325] J. Wang, Q. Lu, and H. P. Lu, PLOS Comput. Biol. **2**, e78 (2006).
[326] A. G. Turjanski, J. S. Gutkind, R. B. Best, and G. Hummer, PLOS Comput. Biol. **4**, e1000060 (2008).
[327] S. G. Estacio, C. S. Fernandes, H. Krobath, P. F. N. Faisca, and E. I. Shakhnovich, J. Chem. Phys. **137**, 085102 (2012).
[328] J. Karanicolas and C. L. Brooks III, J. Mol. Biol. **334**, 309 (2003).
[329] J. K. Noel, P. C. Whitford, K. Y. Sanbonmatsu, and J. N. Onuchic, Nucleic Acids Res. **38**, W657 (2010).
[330] R. D. Hills, Jr. and C. L. Brooks III, Int. J. Mol. Sci. **10**, 889 (2009).
[331] S. Miyazawa and R. L. Jernigan, Macromolecules **18**, 534 (1985).
[332] S. Miyazawa and R. Jernigan, J. Mol. Biol. **256**, 623 (1996).
[333] I. Bahar and R. Jernigan, J. Mol. Biol. **266**, 195 (1997).
[334] H. Li, C. Tang, and N. S. Wingreen, Phys. Rev. Lett. **79**, 765 (1997).
[335] Y. C. Kim and G. Hummer, J. Mol. Biol. **375**, 1416 (2008).
[336] A. Samiotakis, D. Homouz, and M. S. Cheung, J. Chem. Phys. **132**, 175101 (2010).
[337] M. J. Sippl, J. Mol. Biol. **213**, 859 (1990).
[338] M. J. Sippl, J. Comput.-Aided Mol. Des. **7**, 473 (1993).
[339] A. V. Finkelstein, A. Y. Badretdinov, and A. M. Gutin, Proteins **23**, 142 (1995).
[340] R. L. Jernigan and I. Bahar, Curr. Opin. Struct. Biol. **6**, 195 (1996).
[341] F. Melo and E. Feytmans, J. Mol. Biol. **267**, 207 (1997).
[342] H. Lu and J. Skolnick, Proteins **44**, 223 (2001).
[343] H. Zhou and Y. Zhou, Protein Sci. **11**, 2714 (2002).
[344] R. S. DeWitte and E. I. Shakhnovich, J. Am. Chem. Soc. **118**, 11733 (1996).
[345] H. Gohlke, M. Hendlich, and G. Klebe, J. Mol. Biol. **295**, 337 (2000).
[346] M. Y. Shen and A. Sali, Protein Sci. **15**, 2507 (2006).
[347] Y. Dehouck, D. Gilis, and M. Rooman, Biophys. J. **90**, 4010 (2006).
[348] Y. Zhou, Y. Duan, Y. Yang, E. Faraggi, and H. Lei, Theor. Chem. Acc. **128**, 3 (2011).
[349] M. R. Betancourt and J. Skolnick, J. Mol. Biol. **342**, 635 (2004).
[350] M. R. Betancourt, J. Phys. Chem. B **112**, 5058 (2008).
[351] N. V. Buchete, J. E. Straub, and D. Thirumalai, J. Chem. Phys. **118**, 7658 (2003).
[352] N. V. Buchete, J. E. Straub, and D. Thirumalai, Curr. Opin. Struct. Biol. **14**, 225 (2004).
[353] S. Miyazawa and R. L. Jernigan, J. Chem. Phys. **122**, 024901 (2005).
[354] M. Lu, A. D. Dousis, and J. Ma, Protein Sci. **17**, 1576 (2008).
[355] H. Zhou and J. Skolnick, Biophys. J. **101**, 2043 (2011).
[356] P. J. Munson and R. K. Singh, Protein Sci. **6**, 1467 (1997).
[357] B. Krishnamoorthy and A. Tropsha, Bioinformatics **19**, 1540 (2003).
[358] X. Li and J. Liang, Proteins **60**, 46 (2005).
[359] Y. P. Feng, A. Kloczkowski, and R. L. Jernigan, Proteins **68**, 57 (2007).
[360] P. Thomas and K. Dill, Proc. Natl. Acad. Sci. U.S.A. **93**, 11628 (1996).
[361] S.-Y. Huang and X. Zou, J. Comput. Chem. **27**, 1866 (2006).
[362] F. Müller-Plathe, ChemPhysChem **3**, 754 (2002).
[363] E. Kussell, J. Shimada, and E. I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **99**, 5343 (2002).
[364] I. A. Hubner, E. J. Deeds, and E. I. Shakhnovich, Proc. Natl. Acad. Sci. U.S.A. **102**, 18914 (2005).
[365] J. S. Yang, W. W. Chen, J. Skolnick, and E. I. Shakhnovich, Structure **15**, 53 (2007).
[366] A. Godzik, A. Kolinski, and J. Skolnick, Protein Sci. **4**, 2107 (1995).
[367] J. Skolnick, L. Jaroszewski, A. Kolinski, and A. Godzik, Protein Sci. **6**, 676 (1997).
[368] M. R. Betancourt and D. Thirumalai, Protein Sci. **8**, 361 (1999).
[369] D. Mohanty, B. N. Dominy, A. Kolinski, C. L. Brooks, and J. Skolnick, Proteins **35**, 447 (1999).
[370] M. R. Betancourt, Proteins **76**, 72 (2009).
[371] R. Samudrala and J. Moult, J. Mol. Biol. **275**, 895 (1998).
[372] S. H. Bryant and C. E. Lawrence, Proteins **16**, 92 (1993).
[373] J. Moult, Curr. Opin. Struct. Biol. **7**, 194 (1997).
[374] W. A. Koppensteiner and M. J. Sippl, Biochemistry(Mosc.) **63**, 247 (1998).
[375] T. Lazaridis and M. Karplus, Curr. Opin. Struct. Biol. **10**, 139 (2000).
[376] S. H. Bryant and C. E. Lawrence, Proteins **9**, 108 (1991).
[377] K. T. Simons, C. Kooperberg, E. Huang, and D. Baker, J. Mol. Biol. **268**, 209 (1997).
[378] H. B. Burgi and J. D. Dunitz, Acta Crystallogr., Sect. B: Struct. Sci. **44**, 445 (1988).
[379] E. Eyal, S. Gerzon, V. Potapov, M. Edelman, and V. Sobolev, J. Mol. Biol. **351**, 431 (2005).
[380] A. Mozzarelli and G. L. Rossi, Annu. Rev. Biophys. Biomol. Struct. **25**, 343 (1996).
[381] P. D. Thomas and K. A. Dill, J. Mol. Biol. **257**, 457 (1996).
[382] A. Ben-Naim, J. Chem. Phys. **107**, 3698 (1997).
[383] F. Seno, A. Maritan, and J. R. Banavar, Proteins **30**, 244 (1998).
[384] A. A. Kossiakoff, M. Randal, J. Guenot, and C. Eigenbrot, Proteins **14**, 65 (1992).
[385] G. L. Butterfoss and J. Hermans, Protein Sci. **12**, 2719 (2003).
[386] R. B. Best, K. Lindorff-Larsen, M. A. DePristo, and M. Vendruscolo, Proc. Natl. Acad. Sci. U.S.A. **103**, 10901 (2006).
[387] O. F. Lange, N. A. Lakomek, C. Fares, G. F. Schroder, K. F. A. Walter, S. Becker, J. Meiler, H. Grubmuller, C. Griesinger, and B. L. de Groot, Science **320**, 1471 (2008).
[388] J. W. Mullinax and W. G. Noid, Proc. Natl. Acad. Sci. U.S.A. **107**, 19867 (2010).
[389] Z. Li, Y. Yang, J. Zhan, L. Dai, and Y. Zhou, Annu. Rev. Biophys. **42**, 315–335 (2013).
[390] G. Crippen, Biochemistry **30**, 4232 (1991).
[391] V. N. Maiorov and G. M. Crippen, J. Mol. Biol. **227**, 876 (1992).
[392] R. A. Goldstein, Z. A. Luthey-Schulten, and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **89**, 4918 (1992).
[393] L. Mirny and E. Shakhnovich, J. Mol. Biol. **264**, 1164 (1996).
[394] M. H. Hao and H. A. Scheraga, Proc. Natl. Acad. Sci. U.S.A. **93**, 4984 (1996).
[395] F. Seno, C. Micheletti, A. Maritan, and J. Banavar, Phys. Rev. Lett. **81**, 2172 (1998).
[396] J. G. Saven, J. Chem. Phys. **118**, 6133 (2003).
[397] F. Seno, A. Trovato, J. R. Banavar, and A. Maritan, Phys. Rev. Lett. **100**, 078102 (2008).
[398] K. T. Simons, I. Ruczinski, C. Kooperberg, B. A. Fox, C. Bystroff, and D. Baker, Proteins **34**, 82 (1999).
[399] A. Liwo, P. Arłukowicz, C. Czaplewski, S. Ołdziej, J. Pillardy, and H. A. Scheraga, Proc. Natl. Acad. Sci. U.S.A. **99**, 1937 (2002).
[400] Y. Zhang, A. Kolinski, and J. Skolnick, Biophys. J. **85**, 1145 (2003).
[401] S. Ołdziej, J. Łagiewka, A. Liwo, C. Czaplewski, M. Chinchio, M. Nanias, and H. A. Scheraga, J. Phys. Chem. B **108**, 16950 (2004).
[402] J. Skolnick, Curr. Opin. Struct. Biol. **16**, 166 (2006).
[403] R. Samudrala and M. Levitt, Protein Sci. **9**, 1399 (2000).
[404] J. Tsai, R. Bonneau, A. V. Morozov, B. Kuhlman, C. A. Rohl, and D. Baker, Proteins **53**, 76 (2003).
[405] C. Keasar and M. Levitt, J. Mol. Biol. **329**, 159 (2003).
[406] J. Maupetit, P. Tuffery, and P. Derreumaux, Proteins **69**, 394 (2007).
[407] M. Vendruscolo and E. Domany, J. Chem. Phys. **109**, 11101 (1998).
[408] D. Tobi, G. Shafran, N. Linial, and R. Elber, Proteins **40**, 71 (2000).
[409] T. Kortemme, A. V. Morozov, and D. Baker, J. Mol. Biol. **326**, 1239 (2003).
[410] Y. Wu, M. Lu, M. Chen, J. Li, and J. Ma, Protein Sci. **16**, 1449 (2007).
[411] J. Skolnick, A. Kolinski, and A. Ortiz, Proteins **38**, 3 (2000).

[412] J. I. Sułkowska, F. Morcos, M. Weigt, T. Hwa, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **109**, 10340 (2012).
[413] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, Proc. Natl. Acad. Sci. U.S.A. **106**, 67–72 (2008).
[414] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, Proc. Natl. Acad. Sci. U.S.A. **108**, E1293 (2011).
[415] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander, PLoS ONE **6**, e28766 (2011).
[416] P. Bradley, K. M. S. Misura, and D. Baker, Science **309**, 1868 (2005).
[417] L. Kinch, S. Yong Shi, Q. Cong, H. Cheng, Y. Liao, and N. V. Grishin, Proteins **79**(Suppl. 10), 59 (2011).
[418] D. Xu, J. Zhang, A. Roy, and Y. Zhang, Proteins **79**(Suppl. 10), 147 (2011).
[419] D. T. Jones, Proteins **45**(Suppl. 5), 127 (2001).
[420] G. Chikenji, Y. Fujitsuka, and S. Takada, J. Chem. Phys. **119**, 6895 (2003).
[421] S. Y. Lee and J. Skolnick, Biophys. J. **95**, 1956 (2008).
[422] A. Roy, A. Kucukural, and Y. Zhang, Nat. Protoc. **5**, 725 (2010).
[423] M. S. Friedrichs and P. G. Wolynes, Science **246**, 371 (1989).
[424] A. Davtyan, N. P. Schafer, W. Zheng, C. Clementi, P. G. Wolynes, and G. A. Papoian, J. Phys. Chem. B **116**, 8494 (2012).
[425] J. A. Hegler, J. Lätzer, A. Shehu, C. Clementi, and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **106**, 15302 (2009).
[426] A. Shehu, L. E. Kavraki, and C. Clementi, Proteins **76**, 837 (2009).
[427] G. Tóth, J. Phys.: Condens. Matter **19**, 335222 (2007).
[428] W. Li and S. Takada, Biophys. J. **99**, 3029 (2010).
[429] R. L. C. Akkermans and W. J. Briels, J. Chem. Phys. **114**, 1020 (2001).
[430] J.-P. Hansen, C. I. Addison, and A. A. Louis, J. Phys.: Condens. Matter **17**, S3185 (2005).
[431] A. J. Clark, J. McCarty, I. Y. Lyubimov, and M. G. Guenza, Phys. Rev. Lett. **109**, 168301 (2012).
[432] W. G. Noid, Methods Mol. Biol. **924**, 487 (2013).
[433] V. Ruhle, C. Junghans, A. Lukyanov, K. Kremer, and D. Andrienko, J. Chem. Theory Comput. **5**, 3211 (2009).
[434] A. Mirzoev and A. P. Lyubartsev, J. Chem. Theory Comput. **9**, 1512 (2013).
[435] J. F. Rudzinski and W. G. Noid, J. Chem. Phys. **135**, 214101 (2011).
[436] R. Evans, Mol. Sim. **4**, 409 (1990).
[437] J. T. Chayes, L. Chayes, and E. H. Lieb, Commun. Math. Phys. **93**, 57 (1984).
[438] J. T. Chayes and L. Chayes, J. Stat. Phys. **36**, 471 (1984).
[439] N. D. Mermin, Phys. Rev. **137**, A1441 (1965).
[440] R. L. Henderson, Phys. Lett. A **49**, 197 (1974).
[441] C. G. Gray and K. E. Gubbins, *Theory of Molecular Fluids: Fundamentals* (Oxford University Press, 1985), Vol. 1.
[442] H. C. Andersen, D. Chandler, and J. D. Weeks, Adv. Chem. Phys. **34**, 105 (1976).
[443] W. Schommers, Phys. Rev. A **28**, 3599 (1983).
[444] L. Reatto, D. Levesque, and J. J. Weis, Phys. Rev. A **33**, 3451 (1986).
[445] R. L. McGreevy and L. Pusztai, Mol. Simul. **1**, 359 (1988).
[446] D. Reith, H. Meyer, and F. Müller-Plathe, Macromolecules **34**, 2335 (2001).
[447] M. C. Rechtsman, F. H. Stillinger, and S. Torquato, Phys. Rev. Lett. **95**, 228301 (2005).
[448] W. Tschöp, K. Kremer, J. Batoulis, T. Bürger, and O. Hahn, Acta Polym. **49**, 61 (1998).
[449] J. Baschnagel, K. Binder, P. Doruker, A. A. Gusev, O. Hahn, K. Kremer, W. L. Mattice, F. Muller-Plathe, M. Murat, W. Paul, S. Santos, U. W. Suter, and V. Tries, Adv. Polym. Sci. **152**, 41 (2000).
[450] A. Villa, C. Peter, and N. F. A. van der Vegt, Phys. Chem. Chem. Phys. **11**, 2077 (2009).
[451] M. R. Betancourt and S. J. Omovie, J. Chem. Phys. **130**, 195103 (2009).
[452] A. A. Louis, P. G. Bolhuis, J. P. Hansen, and E. J. Meijer, Phys. Rev. Lett. **85**, 2522 (2000).
[453] J. W. Mullinax and W. G. Noid, Phys. Rev. Lett. **103**, 198104 (2009).
[454] J. W. Mullinax and W. G. Noid, J. Phys. Chem. C **114**, 5661 (2010).
[455] T. L. Hill, *Statistical Mechanics: Principles and Selected Applications* (Dover reprint, 1987).
[456] J.-P. Hansen and I. R. McDonald, *Theory of Simple Liquids*, 2nd ed. (Academic Press, San Diego, CA, 1990).
[457] M. P. Taylor and J. E. G. Lipson, J. Chem. Phys. **100**, 518 (1994).
[458] P. Attard, J. Chem. Phys. **102**, 5411 (1995).
[459] Y. G. Chen, C. Kaur, and J. D. Weeks, J. Phys. Chem. B **108**, 19874 (2004).
[460] C. R. Ellis, J. F. Rudzinski, and W. G. Noid, Macromol. Theory Simul. **20**, 478 (2011).
[461] J. F. Rudzinski and W. G. Noid, J. Phys. Chem. B **116**, 8621 (2012).
[462] L. R. Pratt and D. Chandler, J. Chem. Phys. **67**, 3683 (1977).
[463] J. W. Mullinax and W. G. Noid, J. Chem. Phys. **133**, 124107 (2010).
[464] A. Villa, N. F. A. van der Vegt, and C. Peter, Phys. Chem. Chem. Phys. **11**, 2068 (2009).
[465] E. Brini, V. Marcon, and N. F. A. van der Vegt, Phys. Chem. Chem. Phys. **13**, 10468 (2011).
[466] O. Engin, A. Villa, C. Peter, and M. Sayar, Macromol. Theory Simul. **20**, 451 (2011).
[467] A. K. Soper, Chem. Phys. **202**, 295 (1996).
[468] D. Reith, M. Pütz, and F. Müller-Plathe, J. Comput. Chem. **24**, 1624 (2003).
[469] H. A. Karimi-Varzaneh and F. Müller-Plathe, Top. Curr. Chem. **307**, 295 (2012).
[470] P. Májek and R. Elber, Proteins **76**, 822 (2009).
[471] K. R. Hadley and C. McCabe, Biophys. J. **99**, 2896 (2010).
[472] K. R. Hadley and C. McCabe, J. Chem. Phys. **132**, 134505 (2010).
[473] T. Terakawa and S. Takada, Biophys. J. **101**, 1450 (2011).
[474] B. Ni and A. Baumketner, J. Chem. Phys. **138**, 064102 (2013).
[475] S. Jain, S. Garde, and S. K. Kumar, Ind. Eng. Chem. Res. **45**, 5614 (2006).
[476] G. Megariotis, A. Vyrkou, A. Leygue, and D. N. Theodorou, Ind. Eng. Chem. Res. **50**, 546 (2011).
[477] C.-C. Fu, P. M. Kulkarni, M. S. Shell, and L. G. Leal, J. Chem. Phys. **137**, 164106 (2012).
[478] R. H. Swendsen, Phys. Rev. Lett. **42**, 859 (1979).
[479] A. P. Lyubartsev and A. Laaksonen, Phys. Rev. E **52**, 3730 (1995).
[480] A. P. Lyubartsev and A. Laaksonen, Phys. Rev. E **55**, 5689 (1997).
[481] A. Lyubartsev, A. Mirzoev, L. J. Chen, and A. Laaksonen, Faraday Discuss. **144**, 43 (2010).
[482] T. Murtola, E. Falck, M. Patra, M. Karttunen, and I. Vattulainen, J. Chem. Phys. **121**, 9156 (2004).
[483] A. P. Lyubartsev, Eur. Biophys. J. **35**, 53 (2005).
[484] A. Savelyev and G. A. Papoian, J. Phys. Chem. B **113**, 7785 (2009).
[485] M. S. Shell, J. Chem. Phys. **129**, 144108 (2008).
[486] A. Chaimovich and M. S. Shell, J. Chem. Phys. **134**, 094112 (2011).
[487] S. Kullback and R. A. Leibler, Ann. Math. Stat. **22**, 79 (1951).
[488] I. Bilionis and N. Zabaras, J. Chem. Phys. **138**, 044313 (2013).
[489] S. P. Carmichael and M. S. Shell, J. Phys. Chem. B **116**, 8383 (2012).
[490] A. Chaimovich and M. S. Shell, Phys. Rev. E **81**, 060104(R) (2010).
[491] P. Español and I. Zúñiga, Phys. Chem. Chem. Phys. **13**, 10538 (2011).
[492] S. Izvekov and G. A. Voth, J. Phys. Chem. B **109**, 2469 (2005).
[493] S. Izvekov and G. A. Voth, J. Chem. Phys. **123**, 134105 (2005).
[494] F. Ercolessi and J. B. Adams, Europhys. Lett. **26**, 583 (1994).
[495] A. J. Chorin, O. H. Hald, and R. Kupferman, Proc. Natl. Acad. Sci. U.S.A. **97**, 2968 (2000).
[496] A. J. Chorin, Multiscale Model. Simul. **1**, 105 (2003).
[497] W. G. Noid, P. Liu, Y. T. Wang, J.-W. Chu, G. S. Ayton, S. Izvekov, H. C. Andersen, and G. A. Voth, J. Chem. Phys. **128**, 244115 (2008).
[498] W. G. Noid, J.-W. Chu, G. S. Ayton, and G. A. Voth, J. Phys. Chem. B **111**, 4116 (2007).
[499] A. J. Chorin and O. H. Hald, *Stochastic Tools in Mathematics and Science* (Springer, New York, NY, 2006).
[500] H. M. Cho and J. W. Chu, J. Chem. Phys. **131**, 134107 (2009).
[501] L. Lu, J. F. Dama, and G. A. Voth, J. Chem. Phys. **139**, 121906 (2013).
[502] W. G. Noid, G. S. Ayton, S. Izvekov, and G. A. Voth, in *Coarse-Graining of Condensed Phase and Biomolecular Systems*, edited by G. A. Voth (CRC Press, 2008), Chap. 3, pp. 21–40.
[503] L. Lu and G. A. Voth, Adv. Chem. Phys. **149**, 47 (2012).
[504] S. Izvekov and G. A. Voth, J. Chem. Theory Comput. **2**, 637 (2006).
[505] L. Lu and G. A. Voth, J. Phys. Chem. B **113**, 1501 (2009).
[506] Y. Wang and G. A. Voth, J. Phys. Chem. B **114**, 8735 (2010).
[507] L. Deng, Y. Wang, and Z.-C. Ou-Yang, J. Phys. Chem. B **116**, 10135 (2012).
[508] J. Zhou, I. F. Thorpe, S. Izvekov, and G. A. Voth, Biophys. J. **92**, 4289 (2007).
[509] I. F. Thorpe, J. Zhou, and G. A. Voth, J. Phys. Chem. B **112**, 13079 (2008).
[510] I. F. Thorpe, D. P. Goldenberg, and G. A. Voth, J. Phys. Chem. B **115**, 11911 (2011).
[511] R. D. Hills, L. Y. Lu, and G. A. Voth, PLOS Comput. Biol. **6**, e1000827 (2010).
[512] S. Izvekov and G. A. Voth, J. Chem. Phys. **125**, 151101 (2006).
[513] M. C. Villet and G. H. Fredrickson, J. Chem. Phys. **132**, 034109 (2010).

[514]B. J. Alder and T. E. Wainwright, J. Chem. Phys. **31**, 459 (1959).
[515]M. P. Allen and D. P. Tildesley, *Computer Simulation of Liquids* (Oxford Press, New York, NY, 1987).
[516]D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*, 2nd ed. (Academic Press, San Diego, CA, 2002).
[517]A. R. Leach, *Molecular Modelling: Principles and Applications* (Addison-Wesley/Longman, 1998).
[518]W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, J. Am. Chem. Soc. **118**, 11225 (1996).
[519]A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-Mccarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus, J. Phys. Chem. B **102**, 3586 (1998).
[520]J. M. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, J. Comput. Chem. **25**, 1157 (2004).
[521]C. Oostenbrink, A. Villa, A. E. Mark, and W. F. Van Gunsteren, J. Comput. Chem. **25**, 1656 (2004).
[522]V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, Proteins **65**, 712 (2006).
[523]K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, Proteins **78**, 1950 (2010).
[524]R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig, and A. D. MacKerell, Jr., J. Chem. Theory Comput. **8**, 3257 (2012).
[525]A. Liwo, C. Czaplewski, S. Oldziej, and H. A. Scheraga, Curr. Opin. Struct. Biol. **18**, 134 (2008).
[526]M. Christen and W. F. Van Gunsteren, J. Comput. Chem. **29**, 157 (2008).
[527]D. Zuckerman, Annu. Rev. Biophys. **40**, 41 (2011).
[528]J. Mervis, Science **293**, 1235 (2001).
[529]T. Makoto, in *16th IEEE Hot Chips Symposium* (IEEE, 2004).
[530]D. E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R. O. Dror, M. P. Eastwood, J. A. Bank, J. M. Jumper, J. K. Salmon, Y. B. Shan, and W. Wriggers, Science **330**, 341 (2010).
[531]J. Zverina, "XSEDE gaining speed as year two begins," HPC Wire (2012); online at http://www.hpcwire.com/hpcwire/2012-07-25/xsede_gaining_speed_as_year_two_begins.html.
[532]A. W. Götz, M. J. Williamson, D. Xu, D. Poole, S. Le Grand, and R. C. Walker, J. Chem. Theory Comput. **8**, 1542 (2012).
[533]L. C. Pierce, R. Salomon-Ferrer, C. Augusto F. de Oliveira, J. A. McCammon, and R. C. Walker, J. Chem. Theory Comput. **8**, 2997 (2012).
[534]B. R. Brooks, C. L. Brooks III, A. D. Mackerell, Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus, J. Comput. Chem. **30**, 1545 (2009).
[535]S. Plimpton, J. Comput. Phys. **117**, 1 (1995).
[536]J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten, J. Comput. Chem. **26**, 1781 (2005).
[537]B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, J. Chem. Theory Comput. **4**, 435 (2008).
[538]D. Case, T. Darden, I. T. E. Cheatham, C. Simmerling, J. Wang, R. Duke, R. Luo, R. Walker, W. Zhang, K. Merz, B. Roberts, S. Hayik, A. Roitberg, G. Seabra, J. Swails, A. Goetz, I. Kolossváry, K. Wong, F. Paesani, J. Vanicek, R. Wolf, J. Liu, X. Wu, S. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G. Cui, D. Roe, D. Mathews, M. Seetin, R. Salomon-Ferrer, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, and P. Kollman, "Amber12" Techical Report. (University of California, San Francisco, 2012).
[539]P. L. Freddolino, S. Park, B. Roux, and K. Schulten, Biophys. J. **96**, 3772 (2009).
[540]O. F. Lange, D. van der Spoel, and B. L. de Groot, Biophys. J. **99**, 647 (2010).
[541]E. A. Cino, W.-Y. Choy, and M. Karttunen, J. Chem. Theory Comput. **8**, 2725 (2012).
[542]K. A. Beauchamp, Y.-S. Lin, R. Das, and V. S. Pande, J. Chem. Theory Comput. **8**, 1409 (2012).
[543]K. Lindorff-Larsen, P. Maragakis, S. Piana, M. P. Eastwood, R. O. Dror, and D. E. Shaw, PLoS ONE **7**, e32131 (2012).
[544]P. Cieplak, F.-Y. Dupradeau, Y. Duan, and J. Wang, J. Phys.: Condens. Matter **21**, 333102 (2009).
[545]D. Paschek, J. Chem. Phys. **120**, 6674 (2004).
[546]C. Vega and J. L. F. Abascal, Phys. Chem. Chem. Phys. **13**, 19663 (2011).
[547]T. A. Halgren and W. Damm, Curr. Opin. Struct. Biol. **11**, 236 (2001).
[548]A. Warshel, M. Kato, and A. V. Pisliakov, J. Chem. Theory Comput. **3**, 2034 (2007).
[549]J. W. Ponder, C. Wu, P. Ren, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D. S. Lambrecht, R. A. DiStasio, M. Head-Gordon, G. N. I. Clark, M. E. Johnson, and T. Head-Gordon, J. Phys. Chem. B **114**, 2549 (2010).
[550]E. Fadda and R. J. Woods, Drug Discovery Today **15**, 596 (2010).
[551]E. P. Raman, O. Guvench, and A. D. MacKerell, J. Phys. Chem. B **114**, 12981 (2010).
[552]H. S. Hansen and P. H. Hünenberger, J. Comput. Chem. **32**, 998 (2011).
[553]D. L. Mobley, J. Comput.-Aided Mol. Des. **26**, 93 (2012).
[554]S. O. Nielsen, C. F. Lopez, G. Srinivas, and M. L. Klein, J. Chem. Phys. **119**, 7043 (2003).
[555]P. K. Depa and J. K. Maranas, J. Chem. Phys. **123**, 094901 (2005).
[556]P. K. Depa and J. K. Maranas, J. Chem. Phys. **126**, 054903 (2007).
[557]D. Fritz, K. Koschke, V. A. Harmandaris, N. F. A. van der Vegt, and K. Kremer, Phys. Chem. Chem. Phys. **13**, 10412 (2011).
[558]H. Mori, Prog. Theor. Phys. **33**, 423 (1965).
[559]A. N. Naganathan, "Coarse-grained models of protein folding as detailed tools to connect with experiments," Wiley Interdiscip. Rev. Comput. Mol. Sci. (published online).
[560]M. L. Klein and W. Shinoda, Science **321**, 798 (2008).
[561]H. L. Scott, Curr. Opin. Struct. Biol. **12**, 495 (2002).
[562]A. T. Hagler and B. Honig, Proc. Natl. Acad. Sci. U.S.A. **75**, 554 (1978).
[563]C. Wilson and S. Doniach, Proteins **6**, 193 (1989).
[564]F. E. Boas and P. B. Harbury, Curr. Opin. Struct. Biol. **17**, 199 (2007).
[565]L. Larini and J.-E. Shea, J. Phys. Chem. B **116**, 8337 (2012).
[566]S. Fritsch, S. Poblete, C. Junghans, G. Ciccotti, L. Delle Site, and K. Kremer, Phys. Rev. Lett. **108**, 170602 (2012).
[567]R. Potestio, S. Fritsch, P. Español, R. Delgado-Buscalioni, K. Kremer, R. Everaers, and D. Donadio, Phys. Rev. Lett. **110**, 108301 (2013).
[568]D. Alemani, F. Collu, M. Cascella, and M. Dal Peraro, J. Chem. Theory Comput. **6**, 315 (2010).
[569]E. H. Yap, N. L. Fawzi, and T. Head-Gordon, Proteins **70**, 626 (2008).
[570]M. Makowski, E. Sobolewski, C. Czaplewski, S. Ołdziej, A. Liwo, and H. A. Scheraga, J. Phys. Chem. B **112**, 11385 (2008).
[571]V. Krishna, G. S. Ayton, and G. A. Voth, Biophys. J. **98**, 18 (2010).
[572]M. Enciso and A. Rey, J. Chem. Phys. **132**, 235102 (2010).
[573]J. A. Armstrong, C. Chakravarty, and P. Ballone, J. Chem. Phys. **136**, 124503 (2012).
[574]N. Guttenberg, J. F. Dama, M. G. Saunders, G. A. Voth, J. Weare, and A. R. Dinner, J. Chem. Phys. **138**, 094111 (2013).
[575]M. Kroger, Phys. Rep. **390**, 453 (2004).
[576]S. Rackovsky and H. A. Scheraga, J. Biomol. Struct. Dyn. **28**, 593 (2011).
[577]D. J. Wales, *Energy Landscapes* (Cambridge University Press, 2003).
[578]F. H. Stillinger and T. A. Weber, Phys. Rev. A **25**, 978 (1982).
[579]O. M. Becker and M. Karplus, J. Chem. Phys. **106**, 1495 (1997).
[580]S. V. Krivov and M. Karplus, J. Chem. Phys. **117**, 10894 (2002).
[581]G. D'Adamo, A. Pelissetto, and C. Pierleoni, J. Chem. Phys. **138**, 234107 (2013).
[582]T. Murtola, E. Falck, M. Karttunen, and I. Vattulainen, J. Chem. Phys. **126**, 075101 (2007).
[583]A. Das and H. C. Andersen, J. Chem. Phys. **132**, 164106 (2010).
[584]B. Mukherjee, L. Delle Site, K. Kremer, and C. Peter, J. Phys. Chem. B **116**, 8474 (2012).
[585]M. Enciso, C. Schutte, and L. Delle Site, Soft Matter **9**, 6118 (2013).
[586]T. Vettorel and H. Meyer, J. Chem. Theory Comput. **2**, 616 (2006).
[587]J. Ghosh and R. Faller, Mol. Simul. **33**, 759 (2007).
[588]A. Liwo, M. Khalili, C. Czaplewski, S. Kalinowski, S. Ołdziej, K. Wachucik, and H. A. Scheraga, J. Phys. Chem. B **111**, 260 (2007).
[589]H.-J. Qian, P. Carbone, X. Chen, H. A. Karimi-Varzaneh, C. C. Liew, and F. Müller-Plathe, Macromolecules **41**, 9919 (2008).
[590]V. Krishna, W. G. Noid, and G. A. Voth, J. Chem. Phys. **131**, 024103 (2009).
[591]E. Sobolewski, M. Makowski, S. Oldziej, C. Czaplewski, A. Liwo, and H. A. Scheraga, Protein Eng. Des. Sel. **22**, 547 (2009).
[592]K. Farah, A. C. Fogarty, M. C. Böhm, and F. Müller-Plathe, Phys. Chem. Chem. Phys. **13**, 2894 (2011).
[593]L. Lu and G. A. Voth, J. Chem. Phys. **134**, 224107 (2011).
[594]S. Izvekov, J. Chem. Phys. **134**, 034104 (2011).
[595]E. Brini and N. F. A. van der Vegt, J. Chem. Phys. **137**, 154113 (2012).

[596]E. Brini, C. R. Herbers, G. Deichmann, and N. F. A. van der Vegt, Phys. Chem. Chem. Phys. **14**, 11896 (2012).
[597]I. Vorobyov, L. Li, and T. W. Allen, J. Phys. Chem. B **112**, 9588 (2008).
[598]W. D. Bennett and D. P. Tieleman, J. Chem. Theory Comput. **7**, 2981 (2011).
[599]M. Jochum, D. Andrienko, K. Kremer, and C. Peter, J. Chem. Phys. **137**, 064102 (2012).
[600]E. C. Allen and G. C. Rutledge, J. Chem. Phys. **128**, 154115 (2008).
[601]S. Izvekov, P. W. Chung, and B. M. Rice, J. Chem. Phys. **133**, 064109 (2010).