JCTC Journal of Chemical Theory and Computation

Article

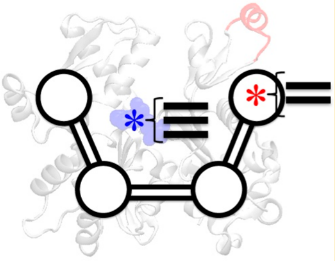
pubs.acs.org/JCTC

The Theory of Ultra-Coarse-Graining. 1. General Principles

James F. Dama,[†,‡,§,#] Anton V. Sinitskiy,[†,‡,§,#] Martin McCullagh,[†,‡,§] Jonathan Weare,[‡,∥] Benoît Roux,[‡,⊥] Aaron R. Dinner,[†,‡,§] and Gregory A. Voth*,[†,‡,§]

[†]Department of Chemistry and Institute for Biophysical Dynamics, [‡]Computation Institute, [§]James Franck Institute, [∥]Department of Mathematics, [⊥]Department of Biochemistry and Molecular Biology, University of Chicago, Chicago, Illinois 60637, United States

ABSTRACT: Coarse-grained (CG) models provide a computationally efficient means to study biomolecular and other soft matter processes involving large numbers of atoms correlated over distance scales of many covalent bond lengths and long time scales. Variational methods based on information from simulations of finer-grained (e.g., all-atom) models, for example the multiscale coarse-graining (MS-CG) and relative entropy minimization methods, provide attractive tools for the systematic development of CG models. However, these methods have important drawbacks when used in the "ultra-coarse-grained" (UCG) regime, e.g., at a resolution level coarser or much coarser than one amino acid residue per effective CG particle in proteins. This is due to the possible existence of multiple metastable states "within" the CG sites for a given UCG model configuration. In this work, systematic variational UCG methods are presented that are specifically designed to CG entire protein domains and subdomains into single effective CG particles. This is accomplished by augmenting existing effective particle CG schemes to allow for discrete state transitions and configuration-dependent resolution. Additionally, certain conclusions of this work connect back to single-state force matching and open up new avenues for method development in that area. These results provide a formal statistical mechanical basis for UCG methods related to force matching and relative entropy CG methods and suggest practical algorithms for constructing optimal approximate UCG models from fine-grained simulation data.



I. INTRODUCTION

Many key biomolecular processes implicated in cellular growth, function, and disease cannot be studied effectively by means of atomistically detailed computer simulation despite impressive recent progress in computational hardware, software, and methodology.[1−9] Furthermore, even when these processes can be simulated at the all-atom resolution level, it is often unclear whether these simulations are necessary or efficient uses of computational resources and whether or not the resulting data are so detailed that they obscure the essential physics of the process being studied. For these reasons, the computational biophysics community has devoted considerable effort to develop efficient coarse-grained (CG) models that describe the processes of interest in reduced detail and at lower computational cost.[6,10−20]

An important subclass of these CG models is developed by using systematic CG methods to create reduced-detail approximations of systems by direct comparison to existing fine-grained (FG), for example, all-atom, models.[21−36] These methods have the goal of reducing the computational cost of obtaining CG statistics from the FG model and of elucidating the essential details needed to reproduce the physics of the FG model at a CG level; they can be contrasted to methods in which CG models are developed to reproduce experimentally known properties directly rather than through agreement with a FG reference model.[37−40] The two approaches each have complementary merits, with one particular advantage of systematic CG methods being that improvements in FG models can be used to improve systematically derived CG models and that this class of CG models can be validated against reference models more thoroughly and transparently. A disadvantage is that the connection of these models to experiments is more indirect, and they can inherit the systematic errors of their base FG models. However, FG models contain a wealth of experimental data and accumulated physical insight that cannot always be incorporated in a CG model through direct fits to experiments. Therefore, while the connection to experiments for systematic CG models may be more indirect, it can also be more extensive and varied, especially in systems for which detailed experimental data are sparse. Which approach should be applied to test a particular hypothesis will depend on the relative availability of reliable FG models and directly relevant experimental data and cannot be judged without expert knowledge of the particular system in question. Previous researchers have argued that systematic CG methods can provide quantitatively accurate, computationally efficient models for a wide range of complex molecular phenomena including biological processes.[6,14,19,23,28,41]

Two very different systematic coarse-graining approaches show promise and have attained significant early successes. In the first of these, CG proceeds by collapsing groups of atoms into single effective particles, also called CG sites, and then using statistically averaged collective forces or histograms from the atomistic model to define interactions between these effective particles.[23,28,29,42] This approach is inspired by a structural biology view of biomolecular function in which the most important fluctuations occur as smooth deformations

Received: January 17, 2013
Published: March 21, 2013

such as twisting, compression, and stretching of common structural motifs. Systematic methods of this type applied to biomolecular simulation include, among others, reverse Monte Carlo,[43,44] Boltzmann inversion,[22,45,46] elastic network modeling (ENM),[47−49] multiscale coarse-graining (MS-CG),[23,24,28,29,50] and relative entropy minimization.[27,51] Most CG models of this type described in the literature have the resolution level of one effective CG particle per several heavy atoms, while more coarse models with one CG site per several amino acid or nucleotide residues, on the order of a hundred atoms per effective particle,[42,52−57] are not as numerous.

In the second approach, a "coarse-graining" of sorts proceeds by collapsing sets of atomistic configurations into single effective states. Statistics from the atomistic model are then used to define transition rates and free energy differences between the discrete states.[25,34,58−60] This approach is inspired by a more biochemical view of biomolecular function in which the most important changes, such as binding events, chemical reactions, and protein domains folding or unfolding, occur in the form of discrete kinetic transitions between states with particular characteristics. Often an additional assumption that the system retains no memory of previous transitions is introduced so that the CG models are Markov state models (MSM).[25,34,36,58,61−66]

The effective particle methods are well-suited to modeling systems in which all events of interest at the CG scale can be naturally explained in terms of smooth structural transitions, whereas effective state models are better suited to modeling systems that seem to evolve exclusively through more abrupt transitions between metastable states when viewed at the CG level. However, neither of these methods is generally applicable to all possible systems.

In key, paradigmatic biomolecular systems, for any level of coarse-graining beyond a few CG particles per residue, there also exists the possibility that some effective CG particles should be able to undergo protonation or other reactions (e.g., ATP hydrolysis), or major conformational changes "within" the CG site. These internal CG site state transitions should be connected to the effective CG particle configurations. Though systematic CG methods for combining different metastable protein conformations in a single model have been previously employed in the literature, no previously published systematic CG method appears capable of constructing models that can represent the essential physics at an "ultra-coarse-grained" (UCG), tens to hundreds of atoms per CG degree of freedom, scale as just described. This article aims to fill that gap, which is an important step in multiscale theory and simulation that must first be taken to link molecular scale interactions with cellular scale phenomena involving many interacting biomolecules.

This article will demonstrate that the effective state and effective particle coarse-graining can be combined naturally within a single UCG framework, enabling the development of models that combine discrete transitions with more continuous motion. The methods presented here promise to combine the strength of these two types of models, by allowing CG particles to have internal states and by allowing state transitions to couple to continuous structural deformations within individual models, using a similar level of rigor to that currently attainable in CG strategies at a higher level of resolution. Our strategy is to split configuration space into discrete volumes based on the values of collective variables either within or between CG particles and then simultaneously optimize a CG model for each volume using volume-restricted FG particle distributions as individual reference models. A CG model similar in spirit has been developed for lipids in the past,[44] but to our knowledge this is the first general, systematic derivation of such state-dependent CG models from generic atomistic FG reference systems in the canonical ensemble.

The remainder of this article is organized as follows. Section II will discuss the motivation of the UCG methodology in the case of a specific and particularly illustrative example, the actin filament. Section III will introduce notation for FG models, CG models, and UCG schemes. Section IV will define consistency between the models and derive useful mathematical formulations of consistency in terms of effective free energy surfaces, forces, and masses. Section V will then illustrate the theory by providing examples of applying UCG consistency to simple models that do not explicitly require an approximate treatment. Section VI will next describe variational approximations to consistency between free energy surfaces, while section VII will complete the body of the article with a summary of the requirements of the method and a discussion of its relationship to previously introduced approaches. Finally, section VIII will conclude the paper.

II. MOTIVATION

As a paradigmatic example motivating UCG methodology, consider the problem of constructing a model of actin filaments (F-actin). These filaments are a key component of the cellular cytoskeleton and play a crucial role in the dynamics of the cell.[67,68] A single monomer (Figure 1a) contains 2883 atoms,[69] and at least 13 monomers are required to approximate an infinite filament using periodic boundary conditions (Figure 1b).[70] As a result, standard atomistic molecular dynamics (MD) simulations of F-actin with explicit water represent a significant challenge even with the use of modern supercomputing resources; the physical time scale of these simulations is currently limited to on the order of 100 ns.[71,72] This time scale is insufficient for full sampling of the filament dynamics,[71,72] which motivates the use of enhanced sampling techniques with biophysically motivated collective CG variables.

Previous work has considered a highly CG model of actin monomer with four CG sites (Figure 1c).[42] In this model, each of the four CG beads was placed at the center of mass of a corresponding domain, and assignment of amino acid residues to one of the four domains was performed by essential dynamics coarse-graining (ED-CG)[42,54,55] based on the spatial correlation of thermal fluctuations of the atoms in the protein. This model reflects many (though not all) important aspects of F-actin filament dynamics.[71,72] One possible way to improve the performance of the model is to add more CG sites, i.e., to introduce more continuous CG variables; these may be defined as the centers of mass of specific fragments of the actin monomer.[71,72]

However, an alternative approach is possible. Experimental studies have established that polymerization of G-actin monomers leading to F-actin, depolymerization of F-actin releasing back G-actin, and interactions of F-actin with other proteins are modulated by the states of the nucleotides bound by the monomers in the filament (ATP, ADP, or ADP + phosphate) and, probably, by the states of the D-loops and other residue conformations in the monomers (folded or unfolded).[67,70,73−75] This behavior suggests that successful UCG models of F-actin should explicitly distinguish between different states of the bound nucleotides and between different states of the D-loops within the CG sites. To improve the four-

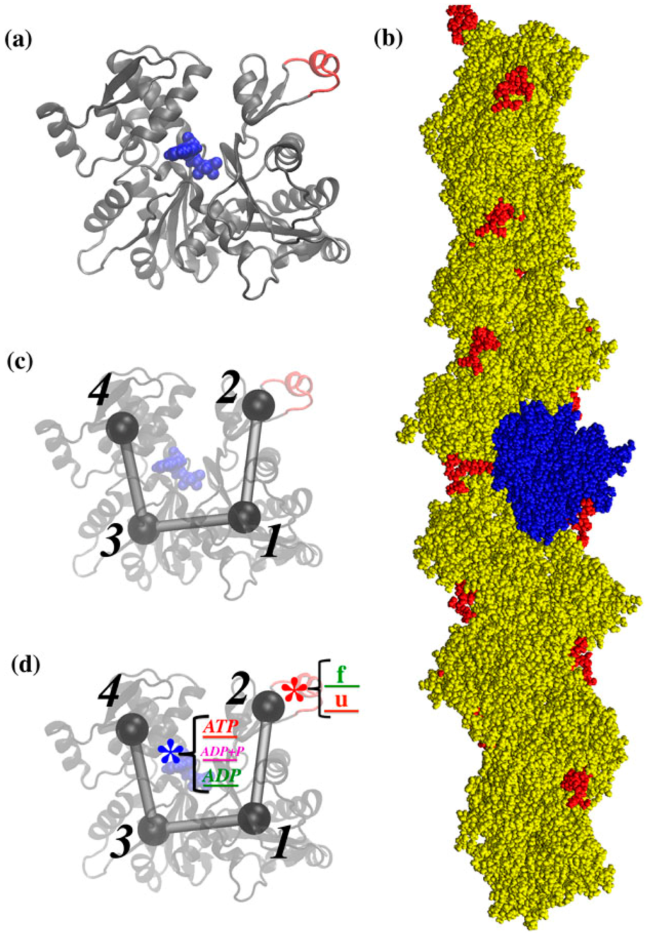


Figure 1. (a) Actin is a 375-residue protein of vital importance for living cells (D-loop shown in red, nucleotide in blue). (b) To simulate F-actin filament dynamics, one must use periodic boundary conditions with 13 monomers in a unit cell, which makes straightforward MD simulations computationally expensive (D-loops are shown in red, one monomer in blue). (c) A simple CG model of actin with four structureless CG sites per monomer as studied in previous work.[42] (d) Addition of two discrete CG variables indicating the states of the nucleotide and the D-loop is expected to significantly improve the performance of the resulting UCG model at the expense of only a slight increase in the number of collective variables.

site CG model without increasing the number of sites, one could, for example, add two discrete CG variables per monomer to indicate the states of the nucleotide and the D-loop in the monomer.

In the resulting model, CG bead number 2 has two internal states defined by whether the D-loop in it is folded or not. As for the other discrete CG variable, it indicates different states of the actin monomer as a whole rather than a specific CG site in it, since the nucleotide was not explicitly included in any of the four CG domains in the original model.[42] The resulting heuristic UCG model (Figure 1d) better fits with biochemical and biophysical reality as compared to the original model with just four structureless CG beads. Given the intuitive attractiveness of this approach and its ability to still scale up to much larger length and time scales relative to all-atom or more highly resolved CG models, the question arises of how to build such an UCG model in a systematic way, strictly from the viewpoint of statistical mechanics. This problem is discussed in the subsequent sections.

III. MODEL DEFINITIONS

The goal of this paper is to describe an UCG methodology suited to creating simple, intuitive highly CG models that blend discrete state changes with continuous structural motion from detailed, complex all-atom models with many continuous degrees of freedom. Our starting point is to formalize the intuitive processes that structural biologists and computational biochemists use in practice. It is worthwhile to sketch out this process here before proceeding to the formal description.

The intuitive process proceeds as follows. First, a researcher selects a detailed FG model, for example an all-atom model[76] or perhaps a more highly resolved CG model such as a Gō model,[77] to simulate and then visualize the molecular system under investigation. Next, the researcher looks for typical large-scale motions and other interesting transitions using principal component analysis, clustering, foreknowledge of the system, or any of many other tools to aid this search.[74,78−80] Once the large-scale motions and states are identified, the researcher will then examine their coupling with more subtle motions, perhaps matching a mixed ENM[81] to their all-atom data or analyzing their Markov chain model states[58] in terms of which CG schemes work best in each state. In this way, the important motions within each state and between each linked group of states are characterized separately. The purpose of the present section will be to describe each step in this process mathematically in a way that treats the states and models for each state together to form a single unified UCG model.

The first step is to describe an FG model. As in previous work on the MS-CG method,[23,24,28] an FG model is a system of n structureless point particles in three-dimensional Euclidean space with a set of positions $\mathbf{r}^n \overset{\text{def}}{=} \{\mathbf{r}_i\}$ and velocities $\dot{\mathbf{r}}^n \overset{\text{def}}{=} \{\dot{\mathbf{r}}_i\}$, where i enumerates individual particles, held at constant temperature and evolving ergodically on an energy surface defined by the Hamiltonian function $h(\mathbf{r}^n, \dot{\mathbf{r}}^n)$:

$$h(\mathbf{r}^n, \dot{\mathbf{r}}^n) = t(\dot{\mathbf{r}}^n) + u(\mathbf{r}^n) \tag{1}$$

where $t(\dot{\mathbf{r}}^n)$ and $u(\mathbf{r}^n)$ are kinetic and potential energies, respectively. Then, the classical partition function for the system can be written (ideal gas normalization is understood) as

$$z = \int \mathrm{d}\mathbf{r}^n \, \mathrm{d}\dot{\mathbf{r}}^n \, \mathrm{e}^{-\beta h(\mathbf{r}^n, \dot{\mathbf{r}}^n)} \tag{2}$$

and the probability density at a given point in phase space is separable in the following sense

$$\begin{aligned} p(\mathbf{r}^n, \dot{\mathbf{r}}^n) &= z^{-1} \, \mathrm{e}^{-\beta h(\mathbf{r}^n, \dot{\mathbf{r}}^n)} \\ &= z^{-1} \, \mathrm{e}^{-\beta u(\mathbf{r}^n)} \, \mathrm{e}^{-\beta t(\dot{\mathbf{r}}^n)} \\ &= p_r(\mathbf{r}^n) \, p_{\dot{r}}(\dot{\mathbf{r}}^n) \end{aligned} \tag{3}$$

The second stage is the process of identifying and assigning discrete states. In this process, one first examines a numerically simulated (or otherwise acquired) ensemble of FG configurations and then chooses a set of distinct states by picking regions of FG configuration space that exemplify qualitatively different behavior, for instance in the case of actin, a region in which a transient helix forms in the flexible D-loop and a region where it is completely disordered. For these regions, the state can be determined unambiguously. However, in most other regions, say when the transient helix is only half-formed, the configuration will appear to be in between well-defined discrete configurational states. The problem of setting precise boundaries on gradual qualitative transitions has a long history as the sorites paradox;[82] in practice, these configurations between well-defined states are assigned somewhat fuzzily between the likely states, with slightly different assignments being possible. Instead of trying to avoid this fuzziness, the present approach formalizes it in the following way. All possible states are described as members ν of a discrete set Σ of latent (hidden) states, where ν is a label, and every configuration $\mathbf{r}^n$ of the FG model has a probability $p_\Sigma(\nu;\, \mathbf{r}^n)$ to be assigned to each state ν. Such probabilities are also known as membership functions. This latent variable distribution varies with both ν and $\mathbf{r}^n$ and is normalized for every value of $\mathbf{r}^n$. In the case of the actin D-loop described above, this could be accomplished by using two states labeled "folded" and "unfolded" and a latent state probability written as a Fermi switching function of the path collective variable "$s(\mathbf{r}^n)$" described in a previous study by Pfaendtner et al.[83]

$$p_{\{\text{folded,unfolded}\}}(\text{folded};\, \mathbf{r}^n) = \frac{1}{e^{(s(\mathbf{r}^n)-s_t)/\sigma} + 1} \tag{4}$$

$$p_{\{\text{folded,unfolded}\}}(\text{unfolded};\, \mathbf{r}^n) = \frac{1}{e^{-(s(\mathbf{r}^n)-s_t)/\sigma} + 1} \tag{5}$$

where s_t is the value of the path collective variable at the midpoint of the transition and σ is the width of the transition in collective variable coordinates. As another example, a protein that may be ligand bound or unbound ({b,u}) and protonated or not protonated ({p,np}) would have four states {bp,up,bnp,unp} for bound and protonated, unbound and protonated, bound and not protonated, and unbound and not protonated, respectively, and if the state changes reflect local changes that are far apart on the protein, the latent state variables might be independent so that

$$p_{\{\text{bp,up,bnp,unp}\}}(\text{bp};\, \mathbf{r}^n) = p_{\{\text{b,u}\}}(\text{b};\, \mathbf{r}^n)\, p_{\{\text{p,np}\}}(\text{p};\, \mathbf{r}^n) \tag{6}$$

which can make specifying and reasoning about the model much simpler. The probability for classifying a configuration as protonated might be a function of the distance of the closest proton from the protonation site, and the probability for classifying a ligand as bound might be a function of the number of hydrogen bonds between the protein and the ligand. In any example, the probability of finding each state is the expectation over the FG configurations of the probability of finding it in any configuration,

$$P_\Sigma(\nu) = \int \mathrm{d}\mathbf{r}^n\, p_r(\mathbf{r}^n)\, p_\Sigma(\nu;\, \mathbf{r}^n) \tag{7}$$

This quantified-uncertainty approach to mapping from a FG description to discrete states has a long precedent in both biophysical modeling (see, e.g., ref 58) and the general field of statistical inference (e.g., in logistic regression[84]), and taking advantage of smoothness in $p_\Sigma(\nu;\, \mathbf{r}^n)$ can be important for developing robust, practical methods for parametrizing UCG models. However, much of what will follow is also valid when $p_\Sigma(\nu;\, \mathbf{r}^n)$ is not smooth, so the use of state maps with no fuzziness can be seen as a special case of this framework.

The state-specific FG distribution $p_r(\mathbf{r}^n)\ p_\Sigma(\nu;\, \mathbf{r}^n)/P_\Sigma(\nu)$ plays an important conceptual role in this work. As stated in the Introduction, our strategy is to split up the FG configuration space into a multiplicity of state volumes, then simultaneously optimize a CG model for each of those volumes using the volume-restricted FG models as individual reference models. The normalized distribution $p_r(\mathbf{r}^n)\ p_\Sigma(\nu;\, \mathbf{r}^n)/P_\Sigma(\nu)$ is exactly a volume-restricted FG model's probability distribution, and the approach in this article is very closely related to force matching or relative entropy minimizing each of these state-specific FG models separately. The potential associated with these state-specific models is $u(\mathbf{r}^n) - \beta^{-1} \ln p_\Sigma(\nu;\, \mathbf{r}^n)$, consisting of the usual energy term and an additional state-dependent entropic term; this quantity will be very important in UCG force matching expressions discussed later. For the example of the folded or unfolded actin D-loop, this state-specific potential is

$$u(\mathbf{r}^n) + \beta^{-1} \ln(e^{(s(\mathbf{r}^n)-s_t)/\sigma} + 1) \tag{8}$$

which has a simple interpretation along its asymptotes; for the unfolded state ($s(\mathbf{r}^n) \gg s_t$), the state-specific potential is approximately $u(\mathbf{r}^n) + \beta^{-1}\sigma^{-1}(s(\mathbf{r}^n) - s_t)$, which corresponds to the usual system plus a linear restraint pulling the system back to the state transition region, while for the folded state ($s(\mathbf{r}^n) \ll s_t$) the state-specific potential is approximately $u(\mathbf{r}^n)$, which is the usual, unperturbed potential of the system. A good state-specific CG model will have similar behavior—it will match the reference system's behavior in the state it is intended to, and it will keep the system in that state until there is a state transition.

Next, with states and state map defined, each state must be analyzed in terms of a reduced set of continuous variables specific to that state. In terms of the previous development, for every value of ν the FG coordinates $\mathbf{r}^n$ are mapped onto a set of reduced variables. Though the reduced set of variables may differ between states, it is not necessary that they do for any of the results that follow. For convenience and clarity, in this paper only linear reductions from the FG variables to the CG variables for each state are considered,[28] though there are reasons to expect that generalizations to nonlinear maps will prove feasible.[85,86]

To make sure that the final UCG model is physically intuitive, each CG degree of freedom must behave as either a particle or an internal mode of a particle under translations of the FG system. The mapped degrees of freedom must rotate properly under rotations of the FG coordinate system, and the model must contain no fully redundant degrees of freedom. Under these limitations, the CG configuration for each atomistic configuration and state assignment will be a system of N_ν linear combinations of FG degrees of freedom, with values $\{\mathbf{R}_{I,\nu}\} \overset{\text{def}}{=} \mathbf{R}^{N_\nu}$, $\mathbf{R}_{I,\nu} = \sum_i c_{Ii,\nu}\mathbf{r}_i$ for some scalars $c_{Ii,\nu}$, with the translation condition requiring that $\sum_i c_{Ii,\nu} = 1$ or $\sum_i c_{Ii,\nu} = 0$ for each continuous degree of freedom index I and the condition that no variables are redundant requiring that all the $c_{Ii,\nu}$ row vectors for each I for a given state are linearly independent. The rotation condition insures that the $c_{Ii,\nu}$ are scalars rather than tensors.

At the UCG level, the physical meaning and importance of inertia, and thus CG momenta, become far less clear. In many cases of biophysical relevance, motions at the CG level are damped, frustrated, caged, and otherwise perturbed by fast bath interactions enough that momentum becomes a less useful variable than simple rate of change, i.e., velocity. For this reason, we consider models with mapped velocities $\dot{\mathbf{R}}_{I,s} = \sum_i c_{Ii,\nu}\dot{\mathbf{r}}_i = \sum_i c_{Ii,\nu}\mathbf{p}_i/m_i$ instead of mapped momenta $\mathbf{P}_{I,\nu} = \sum_i c_{Ii,\nu}\mathbf{p}_i$ as in MS-CG. Following the notation used for MS-CG, both of these transformations are written in terms of mapping matrices

$$M_{R,\nu}^{N_\nu}(\mathbf{r}^n) \overset{\text{def}}{=} \{\sum_i c_{Ii,\nu}\mathbf{r}_i\} \tag{9}$$

$$M_{R_{I,\nu}}(\mathbf{r}^n) \overset{\text{def}}{=} \sum_i c_{Ii,\nu}\mathbf{r}_i \tag{10}$$

$$M_{R,\nu}^{N_\nu}(\dot{\mathbf{r}}^n) \overset{\text{def}}{=} \{\sum_i c_{Ii,\nu}\dot{\mathbf{r}}_i\} \tag{11}$$

$$M_{R_{I,\nu}}(\dot{\mathbf{r}}^n) \overset{\text{def}}{=} \sum_i c_{Ii,\nu}\dot{\mathbf{r}}_i \tag{12}$$

Though a map must be defined for each state in this scheme, it is not necessary that the maps differ between states for the results of this article to hold.

Finally, for the given system, define an apparent (effective) UCG energy function

$$H(\nu, \mathbf{R}^{N_\nu}, \dot{\mathbf{R}}^{N_\nu}) = T(\nu, \dot{\mathbf{R}}^{N_\nu}) + U(\nu, \mathbf{R}^{N_\nu}) \tag{13}$$

that depends on both the state and a point in corresponding CG phase space and can be used to define a Boltzmann distribution on the full CG model space. The partition function for the full UCG model is

$$Z = \sum_\nu \int \mathrm{d}\mathbf{R}^{N_\nu}\, \mathrm{d}\dot{\mathbf{R}}^{N_\nu}\, \mathrm{e}^{-\beta H(\nu,\mathbf{R}^{N_\nu},\dot{\mathbf{R}}^{N_\nu})} \tag{14}$$

with state-specific partition functions

$$Z_\nu = \int \mathrm{d}\mathbf{R}^{N_\nu}\, \mathrm{d}\dot{\mathbf{R}}^{N_\nu}\, \mathrm{e}^{-\beta H(\nu,\mathbf{R}^{N_\nu},\dot{\mathbf{R}}^{N_\nu})} \tag{15}$$

and the probability distribution for the full model is

$$P(\nu, \mathbf{R}^{N_\nu}, \dot{\mathbf{R}}^{N_\nu}) = Z^{-1}\, \mathrm{e}^{-\beta H(\nu,\mathbf{R}^{N_\nu},\dot{\mathbf{R}}^{N_\nu})} \tag{16}$$

with state-specific probability distributions

$$P_\nu(\mathbf{R}^{N_\nu}, \dot{\mathbf{R}}^{N_\nu}) = Z_\nu^{-1}\, \mathrm{e}^{-\beta H(\nu,\mathbf{R}^{N_\nu},\dot{\mathbf{R}}^{N_\nu})} = P(\nu, \mathbf{R}^{N_\nu}, \dot{\mathbf{R}}^{N_\nu})/P_\Sigma(\nu) \tag{17}$$

The coordinate and velocity distributions are separable given the form of the energy, allowing the definition of probabilities $P_{\nu,R}(\boldsymbol{R}^{N_\nu})$ and $P_{\nu,\dot{R}}(\dot{\mathbf{R}}^{N_\nu})$ such that

$$\begin{aligned} P(\nu, \mathbf{R}^{N_\nu}, \dot{\mathbf{R}}^{N_\nu}) &= P_\Sigma(\nu)\, P_\nu(\mathbf{R}^{N_\nu}, \dot{\mathbf{R}}^{N_\nu}) \\ &= P_\Sigma(\nu)\, P_{\nu,R}(\mathbf{R}^{N_\nu})\, P_{\nu,\dot{R}}(\dot{\mathbf{R}}^{N_\nu}) \end{aligned} \tag{18}$$

Finally, it is also useful to define state-dependent delta functions

$$\delta(M_{R,\nu}^{N_\nu}(\mathbf{r}^n) - \mathbf{R}^{N_\nu}) \overset{\text{def}}{=} \prod_I \delta(M_{R_{I,\nu}}(\mathbf{r}^n) - \mathbf{R}_{I,\nu}) \tag{19}$$

$$\delta(M_{R,\nu}^{N_\nu}(\dot{\mathbf{r}}^n) - \dot{\mathbf{R}}^{N_\nu}) \overset{\text{def}}{=} \prod_I \delta(M_{R_{I,\nu}}(\dot{\mathbf{r}}^n) - \dot{\mathbf{R}}_{I,\nu}) \tag{20}$$

which allows for the shorthand notation

$$\begin{aligned} P(\nu, \mathbf{R}^{N_\nu}, \dot{\mathbf{R}}^{N_\nu}) = \int &\mathrm{d}\mathbf{r}^n\, \mathrm{d}\dot{\mathbf{r}}^n\, p_r(\mathbf{r}^n)\, p_{\dot{r}}(\dot{\mathbf{r}}^n)\, p_\Sigma(\nu;\, \mathbf{r}^n) \\ &\times \delta(M_{R,\nu}^{N_\nu}(\mathbf{r}^n) - \mathbf{R}^{N_\nu})\, \delta(M_{R,\nu}^{N_\nu}(\dot{\mathbf{r}}^n) - \dot{\mathbf{R}}^{N_\nu}) \end{aligned} \tag{21}$$

In summary, each point in the FG phase space $(\mathbf{r}^n, \dot{\mathbf{r}}^n)$ is assigned a latent state variable according to $p_\Sigma(\nu;\, \mathbf{r}^n)$ and then mapped to a point in CG phase space $(\nu, \mathbf{R}^{N_\nu}, \dot{\mathbf{R}}^{N_\nu})$ using the state-specific map $M_{R,\nu}^{N_\nu}$. The CG potential must depend on state as well as the other mapped variables. The distribution $p_\Sigma(\nu;\, \mathbf{r}^n)$ must be a discrete probability distribution for the latent variable ν normalized for each $\mathbf{r}^n$, though it need not be continuous as a function of $\mathbf{r}^n$, while the mapping transformations $M_{R,\nu}^{N_\nu}(\mathbf{r}^n) \overset{\text{def}}{=} \{\sum_i c_{Ii,\nu}\mathbf{r}_i\}$ must be full-rank and such that $\sum_i c_{Ii,\nu} \in \{0,1\}$ for each I. Many other UCG schemes are possible, but this subset of them is particularly convenient for analytical investigation while still allowing the examination of important and practical cases. As with the MS-CG method before, we fully expect that this initial foray into the theory of UCG can be followed up with significant generalizations.[87,88]

IV. CONSISTENCY

Once the FG and CG models and the maps between them have been defined, one can begin to consider the multiscale relationship between the two models. One particularly important relationship is consistency, or indistinguishability. Put simply, one would ideally like to know how to make a CG model so robust that no sequence of experiments separated by long times that measure only functions of the CG variables in a single snapshot in time could ever distinguish between the FG and CG models. (The subject of dynamic consistency is beyond the scope of this paper.) A necessary and sufficient criterion for this indistinguishability is that the probability of seeing any given set of CG variables in the CG model should be the same as the probability of seeing that same set of CG variables mapped from the FG model; in other words, the equilibrium joint probability densities for the variables $(\nu,\mathbf{R}^{N_\nu},\dot{\mathbf{R}}^{N_\nu})$ must be identical in both models. In terms of simulation output, it means that any histograms of any combinations of $(\nu,\mathbf{R}^{N_\nu},\dot{\mathbf{R}}^{N_\nu})$ variables from the CG and FG models should be equal between them.

The probability of seeing a given set of CG configuration variables in the atomistic model is just the combination of the probabilities of seeing that CG configuration given whatever FG configuration it is currently in, so that equality of the equilibrium joint probability densities for configuration can be written as the integral

$$P_{\nu,R}(\mathbf{R}^{N_\nu}) = \int \mathrm{d}\mathbf{r}^n\, \delta(M_{R,\nu}^{N_\nu}(\mathbf{r}^n) - \mathbf{R}^{N_\nu})\, p_\Sigma(\nu;\, \mathbf{r}^n)\, p_r(\mathbf{r}^n) \tag{22}$$

Equation 22 is a necessary and sufficient criterion for the two models to be consistent written in terms of configuration distribution data; it will be referred to as the distributional consistency condition. It can also be written in terms of energies as

$$\begin{aligned} U(\nu, \mathbf{R}^{N_\nu}) = -\beta^{-1} \ln\Big[\int &\mathrm{d}\mathbf{r}^n\, \delta(M_{R,\nu}^{N_\nu}(\mathbf{r}^n) - \mathbf{R}^{N_\nu}) \\ &p_\Sigma(\nu;\, \mathbf{r}^n)\, \mathrm{e}^{-\beta u(\mathbf{r}^n)}\Big] + C \end{aligned} \tag{23}$$

where C is an arbitrary constant. Using the state-specific or volume-restricted potential form introduced earlier, it can also be written as

$$\begin{aligned} U(\nu, \mathbf{R}^{N_\nu}) = -\beta^{-1} \ln\Big[\int &\mathrm{d}\mathbf{r}^n\, \delta(M_{R,\nu}^{N_\nu}(\mathbf{r}^n) - \mathbf{R}^{N_\nu}) \\ &\mathrm{e}^{-\beta(u(\mathbf{r}^n) + kT \ln p_\Sigma(\nu;\mathbf{r}^n))}\Big] + C \end{aligned} \tag{24}$$

which is intuitively related to previous CG consistency expressions—one now chooses to CG a biased FG model restricted to a specific region with bias $kT \ln p_\Sigma(\nu;\, \mathbf{r}^n)$ instead of the full FG model. Satisfying this consistency principle requires that the CG potential reproduce the FG probability distribution at every point in configuration space, which may require the CG potential to be quite complex. There is no general guarantee that a CG potential must be of any specific functional form,[89] or even that it must be smooth, without knowing some details of the FG model. Equation 23 for the apparent CG potential will be used as the basis for a relative-entropy-based variational approach in section IV.

Equation 23 can also be converted to an equation in terms of forces by taking the gradient of the logarithm of each side. In the case that the model only has one discrete state, the resulting equation would be exactly the MS-CG force-matching equation.[23,28] However, unlike in the MS-CG approach, in UCG there are potentially many distinct mappings, up to one per discrete state. Each state's CG free energy surface can be found by force matching; however, because there is no gradient across state boundaries in the CG model, the constant differences between the surfaces must be estimated by another method, for instance state-only distribution matching. To define each state-dependent force field, hold the state fixed, take the logarithm of both sides, and then take the gradient with respect to any of the continuous CG variables defined in that state; the force on a CG variable $R_{I,\nu}$ in state ν is described by

$$\begin{aligned}&\frac{\partial}{\partial \mathbf{R}_{I,\nu}}\ln[p(\nu,\mathbf{R}^{N_\nu})]\\&=\frac{\partial}{\partial \mathbf{R}_{I,\nu}}\ln\left[\int \mathrm{d}\mathbf{r}^n\,\delta(M_{R,\nu}^{N_\nu}(\mathbf{r}^n)-\mathbf{R}^{N_\nu})\,p_\Sigma(\nu;\mathbf{r}^n)\,p_r(\mathbf{r}^n)\right]\end{aligned} \tag{25}$$

The right-hand side can be expressed in terms of atomistic forces by expressing the partial derivative with respect to the CG variable applied to the delta function in eq 25 as a linear combination of partial derivatives with respect to atomistic coordinates and integrating by parts so that the derivatives act on the potential $u(\mathbf{r}^n)$ after applying the chain rule to the Boltzmann factor $p_r(\mathbf{r}^n)$.

To simplify the above expression, we start with the consistency equation with probabilities in terms of the free energies in each model

$$Z^{-1}\,\mathrm{e}^{-\beta U(\nu,\mathbf{R}^{N_\nu})}=z^{-1}\int \mathrm{d}\mathbf{r}^n\,\delta(M_{R,\nu}^{N_\nu}(\mathbf{r}^n)-\mathbf{R}^{N_\nu})\,p_\Sigma(\nu;\mathbf{r}^n)\,\mathrm{e}^{-\beta u(\mathbf{r}^n)} \tag{26}$$

Next, take the logarithm of each side to convert the multiplicative partition functions into an additive free energy offset

$$\begin{aligned}U(\nu,\mathbf{R}^{N_\nu})=-\beta^{-1}\ln&\left[\int \mathrm{d}\mathbf{r}^n\,\delta(M_{R,\nu}^{N_\nu}(\mathbf{r}^n)-\mathbf{R}^{N_\nu})\right.\\&\left.p_\Sigma(\nu;\mathbf{r}^n)\,\mathrm{e}^{-\beta u(\mathbf{r}^n)}\right]+\beta^{-1}\ln\left(\frac{z}{Z}\right)\end{aligned} \tag{27}$$

then take partial derivatives of both sides with respect to the continuous variables for this state

$$\begin{aligned}&\frac{\partial}{\partial \mathbf{R}_{I,\nu}}U(\nu,\mathbf{R}^{N_\nu})=\\&-\beta^{-1}\left(\frac{\int \mathrm{d}\mathbf{r}^n\,\frac{\partial}{\partial \mathbf{R}_{I,\nu}}\delta(M_{R,\nu}^{N_\nu}(\mathbf{r}^n)-R^{N_\nu})\,p_\Sigma(\nu;\mathbf{r}^n)\,\mathrm{e}^{-\beta u(\mathbf{r}^n)}}{\int \mathrm{d}\mathbf{r}^n\,\delta(M_{R,\nu}^{N_\nu}(\mathbf{r}^n)-\mathbf{R}^{N_\nu})\,p_\Sigma(\nu;\mathbf{r}^n)\,\mathrm{e}^{-\beta u(\mathbf{r}^n)}}\right)\end{aligned} \tag{28}$$

From the mapping transformation for this state

$$\frac{\partial}{\partial \mathbf{R}_{I,\nu}}\delta(M_{R_{I,\nu}}(\mathbf{r}^n)-\mathbf{R}_I)=-\frac{1}{c_{Ii,\nu}}\frac{\partial}{\partial \mathbf{r}_i}\delta(M_{R_{I,\nu}}(\mathbf{r}^n)-\mathbf{R}_{I,\nu}) \tag{29}$$

for any nonzero $c_{Ii,s}$, and more generally,

$$\begin{aligned}&\frac{\partial}{\partial \mathbf{R}_{I,\nu}}\delta(M_{R_{I,\nu}}(\mathbf{r}^n)-\mathbf{R}_I)\\&=-\sum_i\frac{d_{Ii,\nu}}{c_{Ii,\nu}}\frac{\partial}{\partial \mathbf{r}_i}\delta(M_{R_{I,\nu}}(\mathbf{r}^n)-\mathbf{R}_{I,\nu})\end{aligned} \tag{30}$$

for any set of nonzero $c_{Ii,\nu}$ if $\sum_i d_{Ii,\nu}=1$, which implies

$$\frac{\partial}{\partial \mathbf{R}_{I,\nu}}U(\nu,\mathbf{R}^{N_\nu})=\beta^{-1}\left(\frac{\int \mathrm{d}\mathbf{r}^n\,\sum_i\frac{d_{Ii,\nu}}{c_{Ii,\nu}}\frac{\partial}{\partial \mathbf{r}_i}\delta(M_{R_{I,\nu}}(\mathbf{r}^n)-\mathbf{R}_{I,\nu})\prod_{J\neq I}^{N_\nu-1}\delta(M_{R_{J,\nu}}(\mathbf{r}^n)-\mathbf{R}_{J,\nu})\,p_\Sigma(\nu;\mathbf{r}^n)\,\mathrm{e}^{-\beta u(\mathbf{r}^n)}}{\int \mathrm{d}\mathbf{r}^n\,\delta(M_{R,\nu}^{N_\nu}(r^n)-\mathbf{R}^{N_\nu})\,p_\Sigma(\nu;r^n)\,\mathrm{e}^{-\beta u(\mathbf{r}^n)}}\right) \tag{31}$$

which can be integrated by parts to

$$\begin{aligned}&\frac{\partial}{\partial \mathbf{R}_{I,\nu}}U(\nu,\mathbf{R}^{N_\nu})=\\&-\beta^{-1}\left(\frac{\int \mathrm{d}\mathbf{r}^n\,\delta(M_{R,\nu}^{N_\nu}(\mathbf{r}^n)-\mathbf{R}^{N_\nu})\sum_i\frac{d_{Ii,\nu}}{c_{Ii,\nu}}\frac{\partial}{\partial \mathbf{r}_i}p_\Sigma(\nu;\mathbf{r}^n)\,\mathrm{e}^{-\beta u(\mathbf{r}^n)}}{\int \mathrm{d}\mathbf{r}^n\,\delta(M_{R,\nu}^{N_\nu}(\mathbf{r}^n)-\mathbf{R}^{N_\nu})\,p_\Sigma(\nu;\mathbf{r}^n)\,\mathrm{e}^{-\beta u(\mathbf{r}^n)}}\right)\end{aligned} \tag{32}$$

so long as

$$\sum_i\frac{d_{Ii,\nu}}{c_{Ii,\nu}}\frac{\partial}{\partial \mathbf{r}_i}\delta(M_{R_{I,\nu}}(\mathbf{r}^n)-\mathbf{R}_{J,\nu})=0\text{ for }I\neq J \tag{33}$$

or, using the delta function identities again,

$$\sum_i\frac{d_{Ii,\nu}}{c_{Ii,\nu}}c_{Ji,\nu}=\delta_{IJ} \tag{34}$$

This system of equations implies that the transpose of the $d_{Ii,\nu}/c_{Ii,\nu}$ matrix must be a right inverse of the matrix $M_{R,\nu}^{N_\nu}$; so as long as $M_{R,\nu}^{N_\nu}$ has a right inverse, then the integration by parts can be done. This inverse exists for every UCG model considered in this article due to the earlier requirement that the rows of each $M_{R,\nu}^{N_\nu}$ be linearly independent. Representing the matrix with the elements $d_{Ii,\nu}/c_{Ii,\nu}$ with a new force map symbol $M_{R,\nu}^{N_\nu+}$, the result can be written compactly as

$$\begin{aligned}&\frac{\partial}{\partial \mathbf{R}^{N_\nu}}U(\nu,\mathbf{R}^{N_\nu})=\\&-\beta^{-1}\left(\frac{\int \mathrm{d}\mathbf{r}^n\,\delta(M_{R,\nu}^{N_\nu}(\mathbf{r}^n)-\mathbf{R}^{N_\nu})\,{M_{R,\nu}^{N_\nu}}^{+}\left(\frac{\partial}{\partial \mathbf{r}^n}(p_\Sigma(\nu;\mathbf{r}^n)\,\mathrm{e}^{-\beta u(\mathbf{r}^n)})\right)}{\int \mathrm{d}\mathbf{r}^n\,\delta(M_{R,\nu}^{N_\nu}(\mathbf{r}^n)-\mathbf{R}^{N_\nu})\,p_\Sigma(\nu;\mathbf{r}^n)\,\mathrm{e}^{-\beta u(\mathbf{r}^n)}}\right)\end{aligned} \tag{35}$$

and this can in turn be simplified into a fixed-state, fixed-CG-configuration expectation:

$$\frac{\partial}{\partial \mathbf{R}^{N_\nu}} U(\nu, \mathbf{R}^{N_\nu}) = \left\langle {M_{R,\nu}^{N_\nu}}^{+} \left(\frac{\partial}{\partial \mathbf{r}^n} (u(\mathbf{r}^n) - \beta^{-1} \ln p_\Sigma(\nu; \mathbf{r}^n)) \right) \right\rangle_{\mathbf{R}^{N_\nu},\nu} \tag{36}$$

where the notation

$$\langle f(\mathbf{r}^n) \rangle_{\mathbf{R}^{N_\nu,\nu}} \overset{\text{def}}{=} \frac{\int \mathrm{d}\mathbf{r}^n \, \delta(M_{R,\nu}^{N_\nu}(\mathbf{r}^n) - \mathbf{R}^{N_\nu}) \, p_\Sigma(\nu; \mathbf{r}^n) \, p_r(\mathbf{r}^n) \, f(\mathbf{r}^n)}{\int \mathrm{d}\mathbf{r}^n \, \delta(M_{R,\nu}^{N_\nu}(\mathbf{r}^n) - \mathbf{R}^{N_\nu}) \, p_\Sigma(\nu; \mathbf{r}^n) \, p_r(\mathbf{r}^n)} \tag{37}$$

is used for the UCG configuration specific expectation for any function f of the atomistic configurational variables.

This set of equations, one for the force on each CG degree of freedom defined in each state ν, is satisfied if and only if the structural consistency equations specific to each state are also satisfied, making the two equally valid criteria for consistency between two state-dependent UCG models much as in the case of MS-CG. However, two important changes between these equations and the MS-CG equations deserve comment. First, there is a different force field for each state, as discussed earlier, and the force matching cannot specify the model force field completely because it cannot specify the state-to-state energy differences. Second, the matched forces are no longer the usual atomistic forces; instead, those forces are derivatives of an atomistic free energy that incorporates the entropy of the UCG state distribution. The emergence of this term is the reason why it is useful to allow $p_\Sigma(\nu; \mathbf{r}^n)$ to be smooth, as stated in section III. If it is smooth, that term can be estimated like any other force from simulation as a new entropic contribution; otherwise, the term contains divergences and cannot be estimated directly from points sampled in simulation. Instead, it must be evaluated as a surface integral on the boundary separating the regions in the FG conformational space with different values of discrete CG variables (details not shown). Though this complicates the implementation of the case where $p_\Sigma(\nu; \mathbf{r}^n)$ is not smooth, the mathematics still holds so long as $p_\Sigma(\nu; \mathbf{r}^n)$ is a distribution. In the section that follows, this consistency equation will be used as the basis for a force-matching variational principle.

Consistency in velocity space can be defined and guaranteed using similar logic. In this case, state-dependence is important for defining which velocities exist in each state, but because the state-map is independent of momentum by construction, the equations for consistency are closer to the MS-CG results than the configurational consistency equations are. However, the results are not exactly the same, because for these low-resolution models allowing overlaps in the definitions of effective particles is expected to become increasingly useful,[52,90] and internal breathing modes can also be expected to become physically important.[91,92]

The consistency equation in velocity space can be written as the Gaussian integral

$$P_{\nu,\dot{R}}(\dot{\mathbf{R}}^{N_\nu}) \propto \int \mathrm{d}\dot{\mathbf{r}}^n \, \delta(M_{R,\nu}^{N_\nu}(\dot{\mathbf{r}}^n) - \dot{\mathbf{R}}^{N_\nu}) \, \mathrm{e}^{-\beta \sum_i (1/2) m_i \dot{r}_i^2} \tag{38}$$

which can be solved straightforwardly to obtain the result

$$P_{\nu,\dot{R}}(\dot{\mathbf{R}}^{N_\nu}) \propto \mathrm{e}^{-(1/2)\beta \sum_I \sum_J \mu_{IJ,\nu} \dot{\mathbf{R}}_I \cdot \dot{\mathbf{R}}_J} \tag{39}$$

where μ_ν is the inverse of the matrix with I,J elements equal to $\sum_i c_{Ii,\nu} c_{Ji,\nu} m_i^{-1}$. This derivation recaptures the single-state, no-map-overlap case originally derived for MS-CG; in that case, there is no mixing between sites and $\mu_{IJ,\nu} = \delta_{IJ}(\sum_i {c_{Ii,\nu}}^2 {m_i}^{-1})^{-1}$.

While this kinetic energy is still quadratic no matter how complex the linear mapping, the site velocities can become coupled if their definitions overlap. This behavior need not be taken into account when configuration-space consistency is all that is needed, but in cases where kinetics is of interest, these couplings must be included in a consistent model. They can, however, be made to disappear by designing the map such that the site definitions are orthogonal under the coordinate system with a dot product weighted by the inverse of the atomistic masses, i.e., by specifying that

$$\sum_i c_{Ii,\nu} c_{Ji,\nu} m_i^{-1} = 0 \text{ for all } \nu \text{ and for all } I \neq J \tag{40}$$

Until widely used CG model simulation codes advance to allow for more complex kinetic energy terms, designing maps in a way so these couplings disappear will be necessary for practical studies. However, this challenge has been addressed before in the internal coordinate molecular dynamics community,[93] and there is little reason to believe it will remain impractical in the CG community for long.

The results of this subsection define consistency between a FG model and a CG model, but satisfying the configurational consistency equations may require the CG model to be highly complex. No CG method that requires potential terms that depend on each CG variable simultaneously will be of practical use; even CG models with three-body potentials are presently of limited usage.[94−96] For this reason, full consistency between models is almost never achieved or even intended. Instead, CG practitioners use approximations to consistency and reduced sets of consistency equations in place of the exact, full consistency equations. Deriving approximate variational methods based on these equations for exact consistency is the focus of section VI. However, before presenting approximation strategies, it makes sense to examine a few simple models in which no approximations are needed; these will demonstrate the flexibility and power of this new approach to UCG consistency without forcing us from theory into detailed numerics.

V. EXAMPLES OF CONSISTENT UCG MODELS

At this point, with consistency defined, it appears worthwhile to provide two examples to illustrate how to use these UCG consistent models to clarify essential physics of systems that could be obscured using either effective particle CG or effective state CG alone. (See technical details on these examples in the Appendix.) The first example mimics a system in which the conceptual integrity of a unique effective particle CG breaks down because each CG configuration hyperplane slices through multiple metastable basin manifolds. The second example imitates a system in which the conceptual integrity of a unique effective particle CG breaks down because each metastable basin has different slow and fast variables; the basins are each best described by different CG variables. For the sake of simplicity and to make graphical representation possible, each example starts with a two-dimensional FG model and constructs one-dimensional CG models. Only qualitative behavior of the two examples will be described in this section; the quantitative details of each are available in the Appendix.

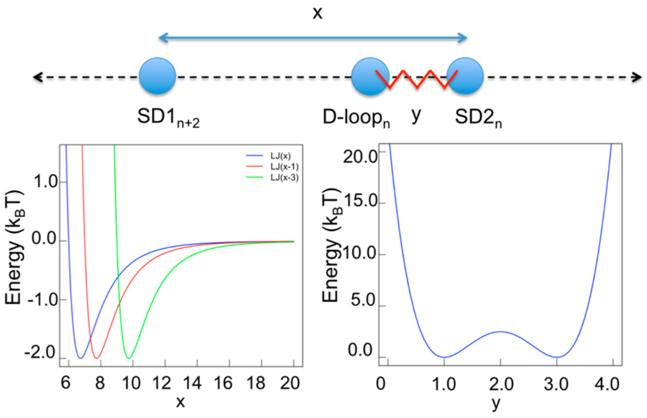


Figure 2. A three-particle system interacting through Lennard-Jones nonbonded potentials and a double well bond potential (top) and the potentials operative between them (below). The particles are labeled to suggest similarity to a conformational selection mechanism suspected to influence filamentous actin dynamics. SD1 and SD2 indicate actin subdomains 1 and 2, while subscripts indicate actin monomer position along the filament.

As a first example, consider a model with three particles constrained to a line with two bonded and the third not bonded to either of the others as shown in Figure 2. This model can be considered a very simple model of the conformational selection ligand binding mechanism, since it involves calculating the approach potential of a nonbonded species to a system with multiple thermally populated metastable states with different configurations. In this example, the two bonded particles interact with the third through a Lennard-Jones potential, while the bonded interaction is a double well potential so that there are two stable bond lengths between the particles. There is no limit to how far the third particle can be from the two bonded particles, while the bonded particles are highly constrained, so it is intuitive to CG the two bonded particles into a single effective particle, especially at long-range. However, if this CG is performed, it obscures the two-state nature of the bonded subsystem (Figure 3), and the resulting nonbonded potential is a counterintuitive double well with complex enthalpic and entropic contributions that may not give reasonable kinetic barriers. A UCG model without these drawbacks is easy to conceive; using the bonded particles' separation distance at the midpoint of the double-well potential as a strict dividing line between two states is a particularly natural choice. In the resulting multistate CG model, one can more easily see that the approach of the third particle to the bonded particles proceeds by a barrierless approach in either state, and that a state transition to the smaller bond length is promoted as the third particle approaches the other two. In addition, one expects that the UCG force field will converge more quickly than an MS-CG force field for this example because the convergence of UCG force matching depends on convergence within states rather than between states, whereas MS-CG depends on convergence of both, and the transition between states is the slowest relaxation in the system. While this example is very simple and low-dimensional, it indicates the advantages of UCG over effective particle CG for representing the physics of a simplistic model of a key mechanism in protein−protein and protein−ligand interaction, conformational selection.

The state assignment probability in the this first example is independent of the CG variable, so the state map entropic force is always mapped to zero, and thus one is free to choose a discontinuous, zero-one state assignment with no loss of convenience. The second example will consider a situation in which that entropic force term is not trivial and demonstrate how each state may have different CG variables. For this example, consider two long thin wells with major axes at right angles to one another, as shown in the upper left panel of Figure 4. In each well, motion along the minor axis is fast, and motion along the major axis is slow, so each seems to suggest CG to a variable orthogonal to the slowest CG variable in the other well. If either is chosen outright, then the CG model will appear to have a single shallow broad well and a single deep narrow well. Furthermore, the depth of the narrow well will be set by entropy and thus highly temperature dependent and poorly transferable. Introducing a state for each well with the line separating the states as the mirror line bisecting the right angle between the wells and then coarse-graining to the major axis variable for each state results in physically more reasonable and transferable models. In this example, the state separation is not independent of the CG variable in either state, so a state map dependent force should be included in the force matching. The bottom left panel of Figure 4 illustrates what happens when this term is left out in the naïve two-state model. Simply put, models that neglect the entropic state assignment weight force show an unphysical stabilization of configurations in the state transition region. As before, though this model is simple, the intuitive picture corresponds to important, paradigmatic models, for instance inactive smooth muscle heavy meromyosin attached to an actin filament, one head bound to actin and the other blocked, with each head fluctuating differently based on its binding state.[97,98]

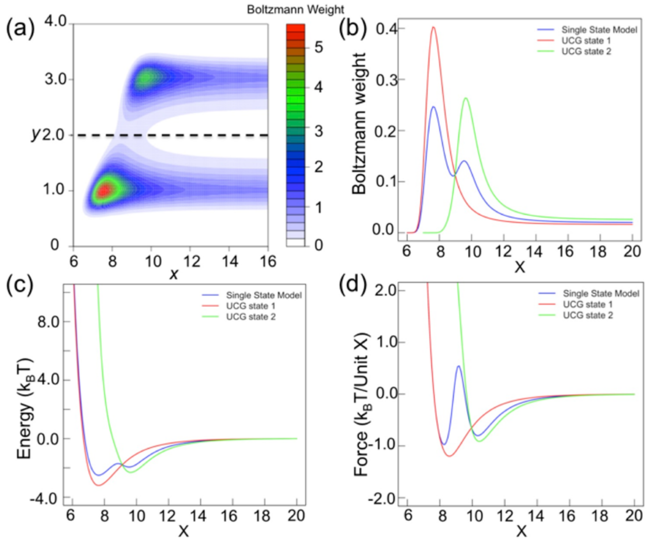


Figure 3. Single-state CG and multistate UCG models corresponding to the three-particle model in Figure 2. On the left are the FG probability distribution (top, a), with corresponding consistent CG and UCG distributions (top, b), forces (below, c), and potentials (below, d).

VI. VARIATIONAL METHODS

Section IV defined consistency between FG and state-dependent UCG models and derived two equivalent conditions for consistency, eqs 23 and 36. The first was written in terms of equality of distributions and the second in terms of equality of forces. This section will describe two variational approaches to choosing CG models based on FG data, one based on

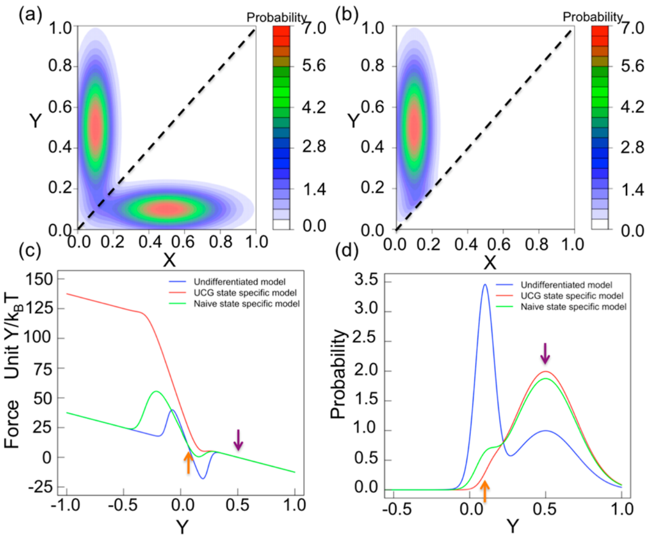


Figure 4. Correct description of the forces acting on UCG sites require taking into consideration the entropy contribution arising from the latent probabilities of various discrete states. Given are the FG probability distribution (top, a) and state-specific FG probability distribution (top, b), with corresponding consistent CG forces (c) and probability distributions (d) for two different CG strategies to the right. Both CG by integrating out a well-transverse variable, but one uses a single, undifferentiated state while the second uses two states. Panels c and d show the consistent CG model as the UCG state-specific model and the CG model derived without the state map entropic force term as the naïve state-specific model; arrows indicate the locations of the peak maxima in the FG probability distributions.

distributional equality and minimizing relative entropies and the other based on force equality and minimizing mean-squared error.

The distributional consistency equations define an optimal choice of UCG distribution given data for the FG system it represents. By measuring how different the distributions in trial state-dependent UCG models are from this optimal choice, one can determine their relative merits. A natural choice for measuring differences between probability distributions is the relative entropy

$$S_{\mathrm{rel},g}[W] = \sum_{\nu\in\Sigma}\int p_{\Sigma}(\nu;\, \mathbf{r}^n)\, p_r(\mathbf{r}^n) \ln\frac{p_{\Sigma}(\nu;\, \mathbf{r}^n)\, p_r(\mathbf{r}^n)}{P_{\nu,R}(\nu,\, M_{R,\nu}^{N_\nu}(\mathbf{r}^n))\, g(\mathbf{r}^n;\, \nu,\, M_{R,\nu}^{N_\nu}(\mathbf{r}^n))}\, \mathrm{d}\mathbf{r}^n \tag{41}$$

where $P_{\nu,R}(\nu, M_{R,\nu}^{N_\nu}(\mathbf{r}^n))$ is the probability under trial CG force field $W(\nu,\mathbf{R}^{N_\nu})$ and $g(\mathbf{r}^n;\, \nu, M_{R,\nu}^{N_\nu}(\mathbf{r}^n))$ is a weighting function for $\mathbf{r}^n$ given $(\nu, M_{R,\nu}^{N_\nu}(\mathbf{r}^n))$; in previous work g has been chosen to be either uniform[27] or equal to the conditional probability of seeing a FG configuration given the CG configuration,[99] and it amounts to no more than a choice of a FG-model-specific constant term. The relative entropy has been applied as the basis of a variational principle for single-state models by Shell[27] and demonstrated to be a powerful tool for the CG of biomolecular systems.[33,51] Shell's approach has initially been used to develop CG models with only continuous degrees of freedom but naturally generalizes to the case of multiple latent states. The generalization is simply a variational method centered on minimizing the quantity $S_{\mathrm{rel},g}$ as written above rather than the single-state version. This quantity differs conceptually from Shell's relative entropy in that the latent state variables must be introduced into the outer sum, so that the relative entropy written as a sum over fine-grained states is converted to a sum over fine-grained states distinguished according to imputed coarse-grained states. It is a weighted sum of the relative entropies for each of the models of the separate volumes of configuration space split out of the FG model using the state map.

The above approach to UCG shares many of the appealing qualities of the relative entropy minimization method in the single-state case. In particular, following the work of Shell and Chaimovich,[27,33] one can readily show that this relative entropy UCG obeys the same minimization conditions and has the same attractive convexity properties when the set of trial CG potentials consists of linear combinations of basis functions.

An equally rigorous and potentially more computationally convenient variational method can be built around force matching equations.[23,24,28,99] Here, optimal CG forces are defined by the consistency condition, and difference from the optimal forces is measured using a weighted mean-squared error

$$\chi_\rho^2[W] = \sum_{\nu\in\Sigma}\int \mathrm{d}\mathbf{R}^{N_\nu}\, \rho(\nu,\, \mathbf{R}^{N_\nu}) \times \Big\| \frac{\partial}{\partial \underline{R}^{N_\nu}} W(\nu,\, \mathbf{R}^{N_\nu}) - \left\langle {M_{R,\nu}^{N_\nu}}^{+}\left(\frac{\partial}{\partial \mathbf{r}^n}(u(\mathbf{r}^n) - \beta^{-1}\ln p_{\Sigma}(\nu;\, \mathbf{r}^n))\right)\right\rangle_{\mathbf{R}^{N_\nu},\nu} \tag{42}$$

for any arbitrary everywhere-positive normalized probability distribution $\rho(\nu,\mathbf{R}^{N_\nu})$ and trial potential $W(\nu,\mathbf{R}^{N_\nu})$. By construction, this is uniquely minimized by the consistent UCG force field. The single-state version with canonical weights in place of $\rho(\nu,\mathbf{R}^{N_\nu})$ is a natural choice and has attained notable success as the multiscale coarse-graining (MS-CG) variational principle.[23,24,28]

The residual χ^2 in eq 42 is written in terms of restrained expectations but can easily be converted to an expectation over the entire phase space without restraints and then to an expectation over a FG simulation trajectory much like the MS-CG residual. To see this, we perform the following manipulations, first defining CG and FG force field variables for the sake of brevity

$$\mathbf{G}^{N_\nu}(\nu,\, \mathbf{R}^{N_\nu}) \overset{\mathrm{def}}{=} \frac{\partial}{\partial \mathbf{R}^{N_\nu}} W(\nu,\, \mathbf{R}^{N_\nu}) \tag{43}$$

$$\mathbf{f}^n(\nu,\, \mathbf{r}^n) \overset{\mathrm{def}}{=} \frac{\partial}{\partial \mathbf{r}^n}(u(\mathbf{r}^n) - \beta^{-1}\ln p_{\Sigma}(\nu;\, \mathbf{r}^n)) \tag{44}$$

Then

$$\chi_\rho^2[W] = \sum_{\nu\in\Sigma}\int \mathrm{d}\mathbf{R}^{N_\nu}\, \rho(\nu,\, \mathbf{R}^{N_\nu})\, \| \mathbf{G}^{N_\nu}(\nu,\, \mathbf{R}^{N_\nu}) - \langle {M_{R,\nu}^{N_\nu}}^{+}\mathbf{f}^n(\nu,\, \mathbf{r}^n)\rangle_{\mathbf{R}^{N_\nu},\nu} \|^2 \tag{45}$$

$$\chi_\rho^2[W] = \sum_{\nu\in\Sigma}\int d\mathbf{R}^{N_\nu}\rho(\nu,\, \mathbf{R}^{N_\nu})(\| \mathbf{G}^{N_\nu}(\nu,\, \mathbf{R}^{N_\nu})\|^2 - 2\mathbf{G}^{N_\nu}(\nu,\, \mathbf{R}^{N_\nu})\cdot \langle {M_{R,\nu}^{N_\nu}}^{+}\mathbf{f}^n(\nu,\, \mathbf{r}^n)\rangle_{\mathbf{R}^{N_\nu},\nu} + \| {M_{R,\nu}^{N_\nu}}^{+}\mathbf{f}^n(\nu,\, \mathbf{r}^n)_{\mathbf{R}^{N_\nu},\nu} \|^2) \tag{46}$$

$$\chi_\rho^2[W] = \sum_{\nu\in\Sigma} \int \mathrm{d}\mathbf{R}^{N_\nu}\, \rho(\nu, \mathbf{R}^{N_\nu})(\|\, \mathbf{G}^{N_\nu}(\nu, \mathbf{R}^{N_\nu})\|^2 - 2\mathbf{G}^{N_\nu}(\nu, \mathbf{R}^{N_\nu})\cdot\langle M_{R,\nu}^{N_\nu\,+}\mathbf{f}^n(\nu, \mathbf{r}^n)\rangle_{R^{N_\nu},\nu} + \langle \|\, M_{R,\nu}^{N_\nu\,+}\mathbf{f}^n(\nu, \mathbf{r}^n)\|^2\rangle_{R^{N_\nu},\nu}) + C' \quad (47)$$

$$\chi_\rho^2[W] = \sum_{\nu\in\Sigma} \int \mathrm{d}\mathbf{R}^{N_\nu}\, \rho(\nu, \mathbf{R}^{N_\nu})(\langle \|\, \mathbf{G}^{N_\nu}(\nu, \mathbf{R}^{N_\nu})\|^2\rangle_{R^{N_\nu},\nu} - \langle 2\mathbf{G}^{N_\nu}(\nu, \mathbf{R}^{N_\nu})\cdot M_{R,\nu}^{N_\nu\,+}\mathbf{f}^n(\nu, \mathbf{r}^n)\rangle_{\mathbf{R}^{N_\nu},\nu} + \langle \|\, M_{R,\nu}^{N_\nu\,+}\mathbf{f}^n(\nu, \mathbf{r}^n)\|^2\rangle_{\mathbf{R}^{N_\nu},\nu}) + C' \quad (48)$$

$$\chi_\rho^2[W] = \sum_{\nu\in\Sigma} \int \mathrm{d}\mathbf{R}^{N_\nu}\, \rho(\nu, \mathbf{R}^{N_\nu}) \times \langle \|\, \mathbf{G}^{N_\nu}(\nu, \mathbf{R}^{N_\nu}) - M_{R,\nu}^{N_\nu\,+}\mathbf{f}^n(\nu, \mathbf{r}^n)\|^2\rangle_{\mathbf{R}^{N_\nu},\nu} + C' \quad (49)$$

$$\chi_\rho^2[W] = \sum_{\nu\in\Sigma} \int \mathrm{d}\mathbf{r}^n\, \rho(\nu, M_{R,\nu}^{N_\nu}(\mathbf{r}^n))\, p(\mathbf{r}^n;\, \nu, M_{R,\nu}^{N_\nu}(\mathbf{r}^n)) \times \|\, \mathbf{G}^{N_\nu}(\nu, M_{R,\nu}^{N_\nu}(\mathbf{r}^n)) - M_{R,\nu}^{N_\nu\,+}\mathbf{f}^n(\nu, \mathbf{r}^n)^2 \,\|^2 + C' \quad (50)$$

$$\chi_\rho^2[W] = \langle \|\, \mathbf{G}^{N_\nu}(\nu, M_{R,\nu}^{N_\nu}(\mathbf{r}^n)) - M_{R,\nu}^{N_\nu\,+}\mathbf{f}^n(\nu, \mathbf{r}^n)\|^2\rangle_\rho + C' \quad (51)$$

The constant term C' is irrelevant for optimization and can safely be ignored.

This variational principle is written in terms of an arbitrary everywhere-positive weight distribution. When FG configurations are sampled from a simulation generating a FG distribution that is canonical except for possible bias in the distribution of the CG variables, $\rho_r(\mathbf{r}^n) = \rho_r(M_{R,\nu}^{N_\nu}(\mathbf{r}^n))\ p(\mathbf{r}^n; M_{R,\nu}^{N_\nu}(\mathbf{r}^n))$, for instance in constant *NVT* molecular dynamics or umbrella sampling using a CG variable as a collective variable, then this expectation can be written as the time expectation

$$\chi_\rho^2[W] = \lim_{\tau\to\infty} \sum_{\tau}^{t=0} \sum_{\nu\in\Sigma} \rho_\Sigma(\nu;\, \mathbf{r}^n) \times \|\, \mathbf{G}^{N_\nu}(M_{R,\nu}^{N_\nu}(\mathbf{r}_{t,\rho}^n)) - M_{R,\nu}^{N_\nu\,+}\mathbf{f}^n(\nu, \mathbf{r}_{t,\rho}^n)\|^2 \quad (52)$$

where $\rho_\Sigma(\nu;\, \mathbf{r}^n) = \rho(\nu,\mathbf{R}^{N_\nu})/\rho_r(\mathbf{R}^{N_\nu})$. In the case that $\rho(\nu,\mathbf{R}^{N_\nu})$ is the consistent $P_{\nu,R}(\nu,\mathbf{R}^{N_\nu})$, a natural choice corresponding to the MS-CG method, $\rho_\Sigma(\nu;\, \mathbf{r}^n) = p_\Sigma(\nu;\, \mathbf{r}^n)$. This has a notable similarity to the residual of MS-CG,[23,24,28] and it can be treated much the same way algorithmically, though the choice of weight is novel and has significant implications that will be commented on further in section VII. Using a set of trial force fields that are linear combinations of state-dependent basis functions, it becomes a linear least-squares problem amenable to the same numerical techniques used in implementations of MS-CG. However, as mentioned in section IV, it is important to remember that the variance of this expression may be infinite if the state assignment probabilities are not smooth functions of atomistic configuration.

The two error metrics $S_{\mathrm{rel},g}[W]$ and $\chi_\rho{}^2[W]$ are each uniquely minimized by the consistent UCG force field with no other global minima, and each therefore forms the basis for a rigorous variational method for deriving state-dependent UCG force fields from FG simulation data. Each can be seen as a weighted sum of metrics for the CG of individual state-specific FG models with potentials $u(\mathbf{r}^n) - \beta^{-1} \ln p_\Sigma(\nu;\, \mathbf{r}^n)$. Furthermore, for trial force fields that can be written as linear combinations of basis functions with the coefficients the only free parameters, each of these functionals is convex so that the global minimum is the only minimum. The two methods are complementary, with each based on a different concept of error with different implications for application and each requiring a different minimization procedure with its own advantages and disadvantages. Relative entropy minimization is often more expensive but better able to fit structural properties, even when the trial potentials do not match the underlying model's essential physics, while force matching is often less expensive but will not fit structural properties well when the trial potentials cannot represent the physics of a model. Readers interested in choosing between the two approaches are directed to the excellent analysis provided by Rudzinski and Noid[99] and Chaimovich and Shell.[33]

VII. DISCUSSION

Systematic CG of atomistic models has become an increasingly important method for analyzing and simulating complex biological processes that are infeasible to study with state-of-the-art atomistic methods. Current approaches based on grouping atoms together into effective particles become unwieldy as many amino acid residues are collapsed into single particles, while approaches based on modeling conformational space as a network of discrete states represent smooth structural deformations jaggedly. This work has proposed a systematic approach to incorporating both continuous and discrete degrees of freedom in a single model for UCG. It has introduced definitions of maps, consistency, and figures of merit for such CG models. Conveniently, the two methods we have introduced, one method based on relative entropy minimization and the other on force matching, retain many of the strengths of each respective method. Both lead to the same optimal, consistent, unique potential when a complete basis is used to represent the set of trial potentials, and both residuals are fully convex when the set of trial potentials is a space of linear combinations of a set of fixed-form basis functions. This work also described methods to develop models in which effective particle definitions overlap and in which some continuous degrees of freedom in the CG model transform as internal modes of a particle rather than particles; these can be applied whether or not the model incorporates discrete state transitions. Finally, the derivation of the force matching residual for this type of model makes it clear that while canonical sampling of the FG model can be used for model development, certain types of biased sampling can also be used with equal algorithmic difficulty.

According to the development in this work, constructing a systematic UCG model should proceed according to the following steps:

- Substantial analysis of the biomolecular peculiarities of the system under investigation, and typically a certain amount of preliminary MD simulation as well, must be performed to decide on the number and the definitions of discrete CG variables to be used.
- Each configuration in the FG model must be assigned a set of normalized weights describing its similarity to each state present in the UCG model.
- The groups of atoms that should be used to define each continuous degree of freedom in each state of the UCG model must be chosen. Contrary to previous work,[28] it is

not necessary that at least one atom be specific to each site in a given state.

- The coefficients of each atom in each definition of a continuous degree of freedom in each state must be assigned so that the new degree of freedom transforms like a particle or an internal mode.
- The sets of coefficients comprising each definition of each continuous degree of freedom must be linearly independent in each state.
- If the instantaneous dynamics of the system is of interest, then kinetic energy must be assigned for each state according to the velocity consistency equation. To obtain a kinetic energy with no coupling between particles and modes, the mapped degrees of freedom must also satisfy orthogonality relations under the inverse-mass-weighted dot product.
- An estimator for either figure of merit for UCG potentials, relative entropy or force residual, must be devised using data from the FG model.
- The state and configuration dependent potential of the UCG model must be chosen by variationally minimizing the estimator for the chosen figure of merit over an appropriate set of trial potentials.

In principle, if these steps are followed using any set of trial functions that contains the exact potential and the estimators for the optimized figure of merit are sufficiently accurate, then the resulting UCG model will be consistent with the FG model in the sense defined in section III, eq 21. However, obtaining accurate estimators for the variational minimization and choosing an appropriate set of trial force fields are major challenges for both the existing relative entropy and force matching methods.

While this method has been presented in terms of equilibrium statistics, we have been intentionally silent on the topic of finite time dynamics. The status of time in effective particle CG simulations is generally unclear except where memory functions are explicitly matched between the FG and CG models.[100,101] Also, for the UCG models proposed in this paper, molecular dynamics or even generalized Langevin dynamics will not be a feasible simulation scheme; to achieve a dynamic simulation of the full model a state-transition scheme such as kinetic Monte Carlo may perhaps be included in the integration of the equations of motion. Models that change in effective particle number will prove a further challenge to implement, since these may require dynamic particle insertion and deletion.

The behavior of the UCG method is precisely analogous to previous methods, but the set of feasible models in this UCG method is considerably larger than in its MS-CG precursor. Not only do UCG models contain both continuous and discrete degrees of freedom, the definitions of continuous degrees of freedom may overlap far more than in the development of Noid et al.,[28] the models may include both particle-like and mode-like combinations of atomic coordinates, and they can vary throughout configuration space through their state dependence. Each of these new features can be assigned to a paradigmatic application that currently resists study by existing CG models. State-dependent mappings can be motivated by the idea of using a single CG site to represent an ATP-hydrolyzing domain, which is feasible only if a single bead can correspond to each combination of domain+ATP, domain+ADP+phosphate, domain+ADP, and the domain without nucleotide bound.[67] Overlapping site definitions arise naturally in coarse representations of proteins without clear separation between dynamic units and have appeared in work on the optimal CG of elastic network models.[57] Internal modes, meanwhile, play an important role in the polarization of protein domains, and their inclusion may be necessary to describe fluctuations in subdomains neighboring redox active sites.[91,102]

However, it is not necessarily the case that all of the above effects will be needed in every UCG model, and the scheme put forward here scales back gracefully so that models can be only as complex as they need to be. If desired, the configuration map for each state may be the same as every other, the maps may not have any overlapping effective particle definitions, or a model may have overlapping effective particle definitions and internal mode-like coordinates, but no states. Similarly, the state-assignment probabilities can be zero or one everywhere, with no uncertainty, though that may make force field development more challenging unless the states are chosen carefully. Each of those special cases is derived in this work implicitly as well, including rederivation of the original MS-CG and relative entropy methods. In fact, the rederivation of MS-CG contains a notable change that is worth discussing on its own.

In the original MS-CG method,[28] the variational residual was defined as a squared difference between force fields at many individual points sampled from canonical simulation. In this work, by deriving a new method *ab initio* rather than trying to find a justification for an existing method, it has been shown that it is equally valid to use a squared difference between force fields at many individual points sampled from biased simulations as well. The only restrictions on this claim are that the bias must be a function of only CG variables and that it must be finite everywhere. Thus many biased-ensemble enhanced sampling methods, from grid and windowing schemes like adaptive biasing force[103] and umbrella sampling[104] to extended Lagrangian schemes such as temperature-accelerated molecular dynamics,[105] are entirely compatible with MS-CG and can be used to generate the samples at which CG and FG models should be compared without reweighting. Because the cost of sampling the FG canonical ensemble has consistently been the highest cost of MS-CG to date,[29,106] this has significant implications for the potential of MS-CG; investigations into its application will be described in a forthcoming paper.

Any UCG model with multiple states will define a specific number of effective CG potential surfaces. However, it is not necessary to define each surface separately in the variational minimization. For instance, it will often be interesting to define force fields so that one transition affects the force field for one protein domain while leaving the others unchanged, and another independent transition affects a different domain exclusively. In that specific case, the two transitions define four total states, but most of the trial force field will be the same in all four states. No portions of the trial force field will have more than two possible forms, so the complexity of the potential does not grow as quickly as the number of states. Since the complexity of the UCG potential is a major determinant of the cost of the variational minimization, it is important to design the state-dependence of the force field to take advantage of as many symmetries and invariances as possible.

As the examples in section V demonstrate, the methods introduced in this paper provide elegant descriptions of systems that neither effective-particle-CG models nor effective-state-CG models can represent as intuitively. At present, it is not yet clear whether these techniques for systematic CG will be applicable to the study of all biological processes on the UCG scale.[56,107,108] In particular, the approach in this paper may be unable to describe certain fractal-like structures hypothesized in many proteins,[56,107,108] since it is restricted to particle-, mode-, and state-like degrees of freedom. However, pure effective particle models and pure state models have each proven useful for studying substantial, complementary sets of problems, and the ways in which each loses efficacy in the UCG regime suggest that hybrid models will be able to address key open challenges in the field.

VIII. CONCLUSIONS

The UCG methods presented in this work describe means to systematically derive CG state-and-particle models from FG data, bridging two major existing research efforts into the systematic derivation of state-only and particle-only CG models. The UCG hybrid models can represent phenomena expected to play a major role in biomolecular, nanotechnological, and other supramolecular systems related to the existence of multiple metastable FG wells underlying single CG particle configurations and the existence of different natural particle-CG mappings in different regions of configuration space. Neither particle-only nor state-only models represent this physics intuitively. The preceding section has summarized the proposed methods and discussed them in depth. The sections preceding that summary motivate, describe, and formally justify the new UCG methods, as well as provide low dimensional examples as intuitive models for the failure of the state-only and particle-only models and the relative success of hybrid state-and-particle models. Furthermore, the derivation of the formal justifications for the new methods indicates the possibility of significant performance improvements for the MS-CG method based on a new, CG-configuration-weighted force-matching variational residual. In sum, this work has provided a hypothesis as to the type of physical model that can efficiently represent complex chemical and biochemical systems at the UCG level and derived formulas for developing the models necessary to test this hypothesis. In forthcoming papers, we will demonstrate practical implementations of these methods, their numerical performance, and their implications for the hypothesis that combining state and particle degrees of freedom together in a single CG model is necessary and sufficient to describe a wide range of key biological phenomena at the level of hundreds of atoms per CG degree of freedom.

■ APPENDIX

Example 1

One of the major interactions leading to actin filament stability is suggested to be the interaction of the D-loop of monomer n with subdomain one (SD1) of the $n + 2$ monomer. The conformation of the D-loop has been found to contain many different states including two limiting cases, folded and unfolded (see references for actin in main text). One can envision a drastic simplification to this system as the D-loop and bound subdomain two (SD2) interacting by a double well bonded potential (to mimic the folded and unfolded states) and a co-linear SD1 interacting with both the D-loop and SD2 via a Lennard-Jones potential. This system and the energy functions are depicted in Figure 2.

The total energy of the system is given as

$$u(x, y) = u_{\mathrm{dw}}(y) + u_{\mathrm{LJ}}(x) + u_{\mathrm{LJ}}(x - y) \tag{A1}$$

where the individual energy functions are given by (in units of $k_{\mathrm{B}}T$)

$$u_{\mathrm{dw}}(y) = \frac{5}{2}[(y - 2)^2 - 1]^2 \tag{A2}$$

$$u_{\mathrm{LJ}}(x) = 4\left[\left(\frac{6}{x}\right)^{12} - \left(\frac{6}{x}\right)^{6}\right] \tag{A3}$$

Given that the total energy is a function of only two variables, one can compute the two-dimensional fine-grained (FG) Boltzmann weights of different conformations using the standard definition

$$p(x, y) \sim \mathrm{e}^{-u(x,y)} \tag{A4}$$

This can be computed numerically over a finite range of y for the specific energy functions given in eqs A2 and A3, yielding the Boltzmann weights in Figure 3.

A natural coarse-graining of this system can be achieved by grouping the D-loop and SD1 of monomer n into a single particle. Integrating the FG probability function over y yields the desired one-dimensional single state CG probability distribution

$$\begin{aligned} P(X) &= \int_{\infty}^{-\infty}\int_{\infty}^{-\infty} p(x, y)\, \delta(x - X)\, \mathrm{d}x\, \mathrm{d}y \\ &= \int_{\infty}^{-\infty} p(X, y)\, \mathrm{d}y \end{aligned} \tag{A5}$$

$$U(X) = -k_{\mathrm{B}}T \ln[P(X)] \tag{A6}$$

$$F(X) = -\frac{dU(X)}{dX} \tag{A7}$$

In practice, this integration is done over a finite range and yields the one-dimensional Boltzmann weights, energy, and forces depicted as blue lines in Figures 3b,c,d, respectively. This procedure, however, leads to at least two issues. The first issue is the misleadingly low energy barrier of approximately 1 $k_{\mathrm{B}}T$ between energy minima at $X = 8$ and $X = 10$. This barrier, though technically correct, is the barrier along a poorly chosen reaction coordinate and therefore may not be very informative as to the mechanism of the transition. In order to go from a minimum at $X = 8$ to a minimum at $X = 10$, the FG system would have to cross the double-well barrier of 2.5 $k_{\mathrm{B}}T$. This is compensated for by the nonbonded interaction in the single-state model, conflating the effects of the bonded and nonbonded interactions obscurely. The second issue is the peak in the single state model force (blue line in Figure 3d) at $X = 9$. This type of peak will be very difficult to resolve in a numerical force matching procedure for more elaborate systems.

A more physically intuitive CGing of this system is achieved by separating the system based on states and then integrating over y. Here, a state separation at $y = 2$ seems like a natural choice due to the double well potential in y having minima at $y = 1$ and $y = 3$. State 1 will consist of all configurations with $y < 2$; state 2 will be $y > 2$. The latent state probability functions for this choice are

$$p_{\Sigma}(1; X, y) = H(2 - y) \tag{A8}$$

$$p_\Sigma(2;\, X,\, y) = H(y - 2) \tag{A9}$$

where H is the Heaviside step function.

Given these, we can obtain state specific CG probabilities

$$P(1,\, X) = \int_{-\infty}^{\infty} p(X,\, y)\, p_\Sigma(1;\, X,\, y)\, \mathrm{d}y = \int_{-\infty}^{2.0} p(X,\, y)\, \mathrm{d}y \tag{A10}$$

$$\begin{aligned} P(2,\, X) &= \int_{-\infty}^{\infty} p(X,\, y)\, p_\Sigma(2;\, X,\, y)\, \mathrm{d}y \\ &= \int_{2.0}^{-\infty} p(X,\, y)\, \mathrm{d}y \end{aligned} \tag{A11}$$

and, respectively, state specific free energies $U(\nu,X)$ and forces $F(\nu,X)$

$$U(\nu,\, X) = -\ln[P(\nu,\, X)] \text{ where } \nu \in \{1,\, 2\} \tag{A12}$$

$$F(\nu,\, X) = -\frac{dU(\nu,\, X)}{dX} \text{ where } \nu \in \{1,\, 2\} \tag{A13}$$

The CG energies for both states are plotted in Figure 3c. The energy difference between the minimum at $X = 8$ and $X = 10$ is approximately 2.0 $k_\mathrm{B}T$ in this model, which is closer to the original double well barrier of 2.5 $k_\mathrm{B}T$ and qualitatively much more understandable. Similarly, the CG forces are plotted in Figure 3d, and the peak in the single state model at $X = 9$ is not present in either state. Instead, the state specific CG model leads to more smoothly varying forces that should converge more quickly in a force matching procedure.

Example 2

As a second example, consider a two-dimensional probability distribution in which the state separation is not independent of the CG variable as it was in example 1. One such distribution is composed of two similarly shaped but orthogonally oriented anisotropic Gaussians (Figure 4a); in our example, it is described analytically as

$$\begin{aligned} p(x,\, y) = {} & N{\cdot}\exp\left[-0.5\left(\frac{(x - 0.5)^2}{0.2^2} + \frac{(y - 0.1)^2}{0.06^2}\right)\right] \\ & + N{\cdot}\exp\left[-0.5\left(\frac{(x - 0.1)^2}{0.06^2} + \frac{(y - 0.5)^2}{0.2^2}\right)\right] \end{aligned} \tag{A14}$$

where N is a normalization constant ($N = 0.15$). As in the previous example, it may be useful to coarse-grain these distributions by integrating out the fast degree of freedom. Due to the orthogonal nature of the two Gaussians, however, using only a single CG dimension does not allow us to remove the high frequency motion from both Gaussians. We can, however, split the distribution by introducing state probabilities of the form

$$p_\Sigma(1;\, x,\, y) = \frac{1}{\mathrm{e}^{100(x - y)} + 1} \tag{A15}$$

$$p_\Sigma(2;\, x,\, y) = \frac{1}{\mathrm{e}^{100(y - x)} + 1} \tag{A16}$$

On the right-hand side of eq A15, there is a smoothly varying function that goes from a value of 1 in the upper triangle made by the dashed line in Figure 4a to zero in the lower triangle, as shown in Figure 4b. Multiplying eqs A14 and A15 yields a state-specific probability that is pictured in Figure 4c. We can then integrate out the fast degree of freedom by integrating along the x direction. Similarly, we can introduce a latent state probability distribution that selects only the lower Gaussian in Figure 4a, but due to the symmetric nature of this example we will not consider this second state. The results are identical.

Now consider three ways of coarse-graining this probability distribution using force matching. The first step in each will be to compute the state independent FG force along the y direction

$$f(y) = \frac{\partial}{\partial y}\ln(p(x,\, y)) \tag{A17}$$

For a single state CG model, one can simply integrate this force over all x values to give the CG force

$$F_{1\text{-state}}(Y) = \frac{\int_{-\infty}^{\infty} f(y)\, p(x,\, y)\, \delta(y - Y)\, \mathrm{d}x\, \mathrm{d}y}{\int_{-\infty}^{\infty} p(x,\, y)\, \delta(y - Y)\, \mathrm{d}x\, \mathrm{d}y} \tag{A18}$$

This approach leads to the force plotted in blue in Figure 4d. An improvement to the single state model is to determine the expectation value of the force in the y direction for the UCG (state-dependent) probability distributions

$$\begin{aligned} & F_{\text{naive-2-state}}(1,\, Y) \\ & = \frac{\int_{-\infty}^{\infty} f(y){\cdot}p(x,\, y){\cdot}p_\Sigma(1;\, x,\, y)\, \delta(y - Y)\, \mathrm{d}x\, \mathrm{d}y}{\int_{-\infty}^{\infty} p(x,\, y){\cdot}p_\Sigma(1;\, x,\, y)\, \delta(y - Y)\, \mathrm{d}x\, \mathrm{d}y} \end{aligned} \tag{A19}$$

The resulting force is depicted as a green curve in Figure 4d. However, this approach does not lead to a consistent CG model. Instead, this naïve approach leads to an artificial attraction to the overlap region that can be seen clearly as a shoulder around $Y = 0$ of the yellow line entitled "Naive State Specific Model" in Figure 4e. The reason is that eq A19 neglects the entropic force contribution due to the dependence of the state probability on the CG variable. The correct state-specific CG force can be determined using the UCG approach

$$F_{\text{UCG-2-state}}(1,\, Y) = \frac{\int_{-\infty}^{\infty}\left(f(y) - \frac{d}{dy}\ln(p_\Sigma(1;\, x,\, y))\right){\cdot}p(x,\, y){\cdot}p_\Sigma(1;\, x,\, y)\, \delta(y - Y)\, \mathrm{d}x\, \mathrm{d}y}{\int_{-\infty}^{\infty} p(x,\, y){\cdot}p_\Sigma(1;\, x,\, y)\, \delta(y - Y)\, \mathrm{d}x\, \mathrm{d}y} \tag{A20}$$

From Figure 4d, it can be seen that the three resulting forces are drastically different. Converting back to probabilities helps to make these differences clear; these probabilities are given in Figure 4e. From the one-dimensional distributions, one can see that the single state model predicts a large narrow peak center at $Y = 0.1$ and a broad peak centered at $Y = 0.5$. The naive two-state model removes most of the narrow peak centered at $Y = 0.1$ but has a small shoulder in this region corresponding to attraction to the latent state boundary. The UCG method compensates for this as can be seen by the removal of the shoulder at $Y = 0.1$ in the red curve of Figure 4e, yielding a fully consistent state dependent model.

AUTHOR INFORMATION

Corresponding Author

*E-mail: gavoth@uchicago.edu.

Author Contributions

#These authors contributed equally to this work.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation through the Center for Multiscale Theory and Simulation (grant CHE-1136709). The authors are grateful to Dr. Joseph Baker for comments on the manuscript.

REFERENCES

(1) Allen, F.; Almasi, G.; Andreoni, W.; Beece, D.; Berne, B. J.; Bright, A.; Brunheroto, J.; Cascaval, C.; Castanos, J.; Coteus, P.; et al. *IBM Syst. J.* **2001**, *40*, 310−327.

(2) Karplus, M.; McCammon, J. A. *Nat. Struct. Biol.* **2002**, *9*, 646−652.

(3) Stone, J. E.; Phillips, J. C.; Freddolino, P. L.; Hardy, D. J.; Trabuco, L. G.; Schulten, K. *J. Comput. Chem.* **2007**, *28*, 2618−2640.

(4) Anderson, J. A.; Lorenz, C. D.; Travesset, A. *J. Comput. Phys.* **2008**, *227*, 5342−5359.

(5) Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; et al. *J. Comput. Chem.* **2009**, *30*, 1545−1614.

(6) Voth, G. A. *Coarse-Graining of Condensed Phase and Biomolecular Systems*; CRC Press: Boca Raton, FL, 2009.

(7) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; et al. *Science (Washington, DC, U. S.)* **2010**, *330*, 341−346.

(8) Zuckerman, D. M. *Annu. Rev. Biophys.* **2011**, *40*, 41−62.

(9) Dror, R. O.; Dirks, R. M.; Grossman, J. P.; Xu, H.; Shaw, D. E. *Annu. Rev. Biophys.* **2012**, *41*, 429−452.

(10) Tozzini, V. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144−150.

(11) Ayton, G. S.; Noid, W. G.; Voth, G. A. *Curr. Opin. Struct. Biol.* **2007**, *17*, 192−198.

(12) Murtola, T.; Bunker, A.; Vattulainen, I.; Deserno, M.; Karttunen, M. *Phys. Chem. Chem. Phys.* **2009**, *11*, 1869−1892.

(13) Rühle, V.; Junghans, C.; Lukyanov, A.; Kremer, K.; Andrienko, D. *J. Chem. Theory Comput.* **2009**, *5*, 3211−3223.

(14) Tozzini, V. *Q. Rev. Biophys.* **2010**, *43*, 333−371.

(15) de Pablo, J. J. *Annu. Rev. Phys. Chem.* **2011**, *62*, 555−574.

(16) Takada, S. *Curr. Opin. Struct. Biol.* **2012**, *22*, 130−137.

(17) Saunders, M. G.; Voth, G. A. *Curr. Opin. Struct. Biol.* **2012**, *22*, 144−150.

(18) Davtyan, A.; Schafer, N. P.; Zheng, W.; Clementi, C.; Wolynes, P. G.; Papoian, G. A. *J. Phys. Chem. B* **2012**, *116*, 8494−8503.

(19) Riniker, S.; Allison, J. R.; van Gunsteren, W. F. *Phys. Chem. Chem. Phys.* **2012**, *14*, 12423−12430.

(20) Saunders, M. G.; Voth, G. A. *Annu. Rev. Biophys.* **2013**, *42*, in press.

(21) Lyubartsev, A. P.; Laaksonen, A. *Phys. Rev.* **1995**, *52*, 3730−3737.

(22) Reith, D.; Putz, M.; Muller-Plathe, F. *J. Comput. Chem.* **2003**, *24*, 1624−1636.

(23) Izvekov, S.; Voth, G. A. *J. Phys. Chem. B* **2005**, *109*, 2469−2473.

(24) Izvekov, S.; Voth, G. A. *J. Chem. Phys.* **2005**, *123*, 134105.

(25) Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K. A.; Swope, W. C. *J. Chem. Phys.* **2007**, *126*, 155101.

(26) Buchete, N.-V.; Hummer, G. *J. Phys. Chem. B* **2008**, *112*, 6057−6069.

(27) Shell, M. S. *J. Chem. Phys.* **2008**, *129*, 144108.

(28) Noid, W. G.; Chu, J.-W.; Ayton, G. S.; Krishna, V.; Izvekov, S.; Voth, G. A.; Das, A.; Andersen, H. C. *J. Chem. Phys.* **2008**, *128*, 244114.

(29) Noid, W. G.; Liu, P.; Wang, Y.; Chu, J.-W.; Ayton, G. S.; Izvekov, S.; Andersen, H. C.; Voth, G. A. *J. Chem. Phys.* **2008**, *128*, 244115.

(30) Savelyev, A.; Papoian, G. A. *Biophys. J.* **2009**, *96*, 4044−4052.

(31) Savelyev, A.; Papoian, G. A. *J. Phys. Chem. B* **2009**, *113*, 7785−7793.

(32) Mullinax, J.; Noid, W. G. *Phys. Rev. Lett.* **2009**, *103*, 1−4.

(33) Chaimovich, A.; Shell, M. S. *J. Chem. Phys.* **2011**, *134*, 094112.

(34) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. *J. Chem. Phys.* **2011**, *134*, 174105.

(35) Buchner, G. S.; Murphy, R. D.; Buchete, N.-V.; Kubelka, J. *Biochim. Biophys. Acta* **2011**, *1814*, 1001−1020.

(36) Beauchamp, K. A.; McGibbon, R.; Lin, Y. S.; Pande, V. S. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 17807−17813.

(37) Shinoda, W.; DeVane, R.; Klein, M. L. *Mol. Simul.* **2007**, *33*, 27−36.

(38) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. *J. Phys. Chem. B* **2007**, *111*, 7812−7824.

(39) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S.-J. *J. Chem. Theory Comput.* **2008**, *4*, 819−834.

(40) Devane, R.; Shinoda, W.; Moore, P. B.; Klein, M. L. *J. Chem. Theory Comput.* **2009**, *5*, 2115−2124.

(41) Ravikumar, K. M.; Huang, W.; Yang, S. *Biophys. J.* **2012**, *103*, 837−845.

(42) Zhang, Z.; Lu, L.; Noid, W. G.; Krishna, V.; Pfaendtner, J.; Voth, G. A. *Biophys. J.* **2008**, *95*, 5073−5083.

(43) Murtola, T.; Falck, E.; Karttunen, M.; Vattulainen, I. *J. Chem. Phys.* **2007**, *126*, 075101.

(44) Murtola, T.; Karttunen, M.; Vattulainen, I. *J. Chem. Phys.* **2009**, *131*, 055101.

(45) Tschöp, W.; Kremer, K.; Batoulis, J.; Bürger, T.; Hahn, O. *Acta Polym.* **1998**, *49*, 61−74.

(46) Ashbaugh, H. S.; Patel, H. A.; Kumar, S. K.; Garde, S. *J. Chem. Phys.* **2005**, *122*, 104908.

(47) Yang, L.; Song, G.; Jernigan, R. L. *Biophys. J.* **2007**, *93*, 920−929.

(48) Lyman, E.; Pfaendtner, J.; Voth, G. A. *Biophys. J.* **2008**, *95*, 4183−4192.

(49) Bahar, I.; Lezon, T. R.; Yang, L.-W.; Eyal, E. *Annu. Rev. Biophys.* **2010**, *39*, 23−42.

(50) Zhou, J.; Thorpe, I. F.; Izvekov, S.; Voth, G. A. *Biophys. J.* **2007**, *92*, 4289−4303.

(51) Carmichael, S. P.; Shell, M. S. *J. Phys. Chem. B* **2012**, *116*, 8383−8393.

(52) Doruker, P.; Jernigan, R. L.; Bahar, I. *J. Comput. Chem.* **2002**, *23*, 119−127.

(53) Arkhipov, A.; Freddolino, P. L.; Schulten, K. *Structure (Oxford, U. K.)* **2006**, *14*, 1767−1777.

(54) Zhang, Z.; Pfaendtner, J.; Grafmuller, A.; Voth, G. A. *Biophys. J.* **2009**, *97*, 2327−2337.

(55) Zhang, Z.; Voth, G. A. *J. Chem. Theory Comput.* **2010**, *6*, 2990−3002.

(56) Sinitskiy, A. V.; Saunders, M. G.; Voth, G. A. *J. Phys. Chem. B* **2012**, *116*, 8363−8374.

(57) Sinitskiy, A. V.; Voth, G. A. *Chem. Phys.* 2013, in press; DOI: 10.1016/j.chemphys.2013.01.024.

(58) Pande, V. S.; Beauchamp, K.; Bowman, G. R. *Methods (Amsterdam, Neth.)* **2010**, *52*, 99−105.

(59) Lau, A. Y.; Roux, B. *Nat. Chem. Biol.* **2011**, *7*, 130−131.

(60) Vitalis, A.; Caflisch, A. *J. Chem. Theory Comput.* **2012**, *8*, 1108−1120.

(61) Jayachandran, G.; Vishal, V.; Pande, V. S. *J. Chem. Phys.* **2006**, *124*, 164902.

(62) Noé, F.; Horenko, I.; Schütte, C.; Smith, J. C. *J. Chem. Phys.* **2007**, *126*, 155102.

(63) Pan, A. C.; Roux, B. *J. Chem. Phys.* **2008**, *129*, 064107.

(64) Rains, E. K.; Andersen, H. C. *J. Chem. Phys.* **2010**, *133*, 144113.

(65) Lane, T. J.; Bowman, G. R.; Beauchamp, K. A.; Voelz, V. A.; Pande, V. S. *J. Am. Chem. Soc.* **2011**.

(66) Kellogg, E. H.; Lange, O. F.; Baker, D. *J. Phys. Chem. B* **2012**, *116*, 11405−11413.

(67) Pollard, T. D.; Borisy, G. G. *Cell (Cambridge, MA, U. S.)* **2003**, *112*, 453−465.

(68) Pollard, T. D.; Cooper, J. A. *Science (Washington, DC, U. S.)* **2009**, *326*, 1208−1212.

(69) Otterbein, L. R.; Graceffa, P.; Dominguez, R. *Science (Washington, DC, U. S.)* **2001**, *293*, 708−711.

(70) Chu, J. W.; Voth, G. A. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 13111−13116.

(71) Saunders, M. G.; Voth, G. A. *Structure (Oxford, U. K.)* **2012**, *20*, 641−653.

(72) Fan, J.; Saunders, M. G.; Voth, G. A. *Biophys. J.* **2012**, *103*, 1334−1342.

(73) Wegner, A. *J. Mol. Biol.* **1976**, *108*, 139−150.

(74) Ceriotti, M.; Tribello, G. A.; Parrinello, M. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 13023−13028.

(75) Durer, Z. A.; Kudryashov, D. S.; Sawaya, M. R.; Altenbach, C.; Hubbell, W.; Reisler, E. *Biophys. J.* **2012**, *103*, 930−939.

(76) Frenkel, D.; Smit, B. *Understanding Molecular Simulation*, 2nd ed.; Academic Press: New York, 2001.

(77) Gō, N. *Annu. Rev. Biophys. Bioeng.* **1983**, *12*, 183−210.

(78) Qi, B.; Muff, S.; Caflisch, A.; Dinner, A. R. *J. Phys. Chem. B* **2010**, *114*, 6979−6989.

(79) Zheng, W.; Qi, B.; Rohrdanz, M. A.; Caflisch, A.; Dinner, A. R.; Clementi, C. *J. Phys. Chem. B* **2011**, *115*, 13065−13074.

(80) Benson, N. C.; Daggett, V. *J. Phys. Chem. B* **2012**.

(81) Zheng, W.; Brooks, B. R.; Hummer, G. *Proteins* **2007**, *69*, 43−57.

(82) Laërtius, D. *Lives of the Eminent Philosophers*, Book II; f 108; ancient manuscript.

(83) Pfaendtner, J.; Branduardi, D.; Parrinello, M.; Pollard, T. D.; Voth, G. A. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 12723−12728.

(84) Menard, S. *Logistic Regression: From Introductory to Advanced Concepts and Applications*; SAGE Publications: Los Angeles, 2010.

(85) Ciccotti, G.; Kapral, R.; Vanden-Eijnden, E. *Chem. Phys. Chem.* **2005**, *6*, 1809−1814.

(86) Wong, K.-y.; York, D. M. *J. Chem. Theory Comput.* **2012**, *8*, 3998−4003.

(87) Mullinax, J. W.; Noid, W. G. *J. Chem. Phys.* **2009**, *131*, 104110.

(88) Das, A.; Andersen, H. C. *J. Chem. Phys.* **2010**, *132*, 164106.

(89) Das, A.; Andersen, H. C. *J. Chem. Phys.* **2012**, *136*, 194113.

(90) Chennubhotla, C.; Bahar, I. *J. Comput. Biol.* **2007**, *14*, 765−776.

(91) Jha, S. K.; Ji, M.; Gaffney, K. J.; Boxer, S. G. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 16612−16617.

(92) Makowski, L.; Rodi, D. J.; Mandava, S.; Minh, D. D. L.; Gore, D. B.; Fischetti, R. F. *J. Mol. Biol.* **2008**, *375*, 529−546.

(93) Jain, A.; Park, I.-H.; Vaidehi, N. *J. Chem. Theory Comput.* **2012**, *8*, 2581−2587.

(94) Larini, L.; Lu, L.; Voth, G. A. *J. Chem. Phys.* **2010**, *132*, 164107.

(95) Das, A.; Andersen, H. C. *J. Chem. Phys.* **2012**, *136*, 194114.

(96) Molinero, V.; Moore, E. B. *J. Phys. Chem. B* **2009**, *113*, 4008−4016.

(97) Wendt, T.; Taylor, D.; Trybus, K. M.; Taylor, K. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 4361−4366.

(98) Baumann, B. J.; Taylor, D. W.; Huang, Z.; Tama, F.; Fagnant, P. M.; Trybus, K. M.; Taylor, K. *J. Mol. Biol.* **2012**, *415*, 274−287.

(99) Rudzinski, J. F.; Noid, W. G. *J. Chem. Phys.* **2011**, *135*, 214101.

(100) Chang, X.-Y.; Freed, K. F. *Chem. Eng. Sci.* **1994**, *49*, 2821−2832.

(101) Izvekov, S.; Voth, G. A. *J. Chem. Phys.* **2006**, *125*, 151101.

(102) Bandaria, J. N.; Dutta, S.; Nydegger, M. W.; Rock, W.; Kohen, A.; Cheatum, C. M. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 17974−17979.

(103) Darve, E.; Rodríguez-Gómez, D.; Pohorille, A. *J. Chem. Phys.* **2008**, *128*, 144120.

(104) Kästner, J. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 932−942.

(105) Maragliano, L.; Vanden-Eijnden, E. *Chem. Phys. Lett.* **2006**, *426*, 168−175.

(106) Lu, L.; Izvekov, S.; Das, A.; Andersen, H. C.; Voth, G. A. *J. Chem. Theory Comput.* **2010**, *6*, 954−965.

(107) Banerji, A.; Ghosh, I. *Cell. Mol. Life Sci.* **2011**, *68*, 2711−2737.

(108) Reuveni, S.; Klafter, J.; Granek, R. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **2012**, *85*, 011906.