The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models

Cite as: J. Chem. Phys. **128**, 244114 (2008); https://doi.org/10.1063/1.2938860
Submitted: 05 March 2008 . Accepted: 13 May 2008 . Published Online: 27 June 2008

W. G. Noid, Jhih-Wei Chu, Gary S. Ayton, et al.

The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models

W. G. Noid,[1,a)] Jhih-Wei Chu,[1,b)] Gary S. Ayton,[1] Vinod Krishna,[1] Sergei Izvekov,[1] Gregory A. Voth,[1,c)] Avisek Das,[2] and Hans C. Andersen[2]
[1]*Center for Biophysical Modeling and Simulation and Department of Chemistry, University of Utah, Salt Lake City, Utah 84112-0850, USA*
[2]*Department of Chemistry, Stanford University, Stanford, California 94305, USA*

(Received 5 March 2008; accepted 13 May 2008; published online 27 June 2008)

Coarse-grained (CG) models provide a computationally efficient method for rapidly investigating the long time- and length-scale processes that play a critical role in many important biological and soft matter processes. Recently, Izvekov and Voth introduced a new multiscale coarse-graining (MS-CG) method [J. Phys. Chem. B **109**, 2469 (2005); J. Chem. Phys. **123**, 134105 (2005)] for determining the effective interactions between CG sites using information from simulations of atomically detailed models. The present work develops a formal statistical mechanical framework for the MS-CG method and demonstrates that the variational principle underlying the method may, in principle, be employed to determine the many-body potential of mean force (PMF) that governs the equilibrium distribution of positions of the CG sites for the MS-CG models. A CG model that employs such a PMF as a "potential energy function" will generate an equilibrium probability distribution of CG sites that is consistent with the atomically detailed model from which the PMF is derived. Consequently, the MS-CG method provides a formal multiscale bridge rigorously connecting the equilibrium ensembles generated with atomistic and CG models. The variational principle also suggests a class of practical algorithms for calculating approximations to this many-body PMF that are optimal. These algorithms use computer simulation data from the atomically detailed model. Finally, important generalizations of the MS-CG method are introduced for treating systems with rigid intramolecular constraints and for developing CG models whose equilibrium momentum distribution is consistent with that of an atomically detailed model. © *2008 American Institute of Physics*. [DOI: 10.1063/1.2938860]

I. INTRODUCTION

Atomistic molecular dynamics (MD) simulations[1,2] have contributed key insight into the structure, dynamics, and function of many important biomolecular systems by providing a model of molecular motion with angstrom level detail and femtosecond resolution.[3] Enabled by the development of increasingly powerful software[4–8] and hardware,[9,10] MD simulations now routinely model the equilibrium fluctuations of biomolecules, such as proteins,[3,11–13] or bioassemblies, such as lipid bilayers,[14] for tens of nanoseconds across length scales of several nanometers. As one example familiar to us,[15] the remodeling of the cellular plasma membrane has recently been investigated by performing atomistic MD simulations of the N-BAR protein domain[16] interacting with a physiologically relevant model membrane. This simulation study, which combined some of the most powerful software[5,6] and hardware[9,10] currently available for conventional MD simulations, observed the N-BAR-induced membrane bending process over tens of nanometers for on the order of 100 ns. However, many important biological processes such as protein folding,[17,18] signal transduction,[19,20] and the assembly of the HIV-1 viral capsid[21] occur on the microsecond time scale or longer. An extensive investigation of the mechanisms involved in these slowly evolving processes remains well beyond the capability of conventional biomolecular MD simulation methodologies. Consequently, there has been rapidly growing interest in the development of coarse-grained (CG) models for investigating such long time- and length-scale processes that cannot be adequately studied with atomically detailed MD simulations.[22,23]

Many widely disparate CG models have been described in the literature. The present work considers CG models that consist of classical interacting mass points (CG sites) that each correspond to one or more atoms in an atomically detailed simulation of the same system. Because the CG representation often has many fewer degrees of freedom, these low resolution models are usually highly computationally efficient.[24] CG models, therefore, provide a powerful computational tool for rapidly exploring the expansive conformational space relevant to complex biological processes evolving on very long time scales. Accordingly, considerable effort has been expended in developing CG models for studying a range of biological processes (see, for example, Refs. 23–44). Implicit in much of this work is the fundamental underlying assumption that the results observed with the low resolution model are somehow consistent with those that

[a)]Present address: Department of Chemistry, Pennsylvania State University, University Park, PA 16802.
[b)]Present address: Department of Chemical Engineering, University of California, Berkeley, CA 94720.
[c)]Electronic mail: voth@chem.utah.edu.

would be observed using a more detailed and thus more computationally expensive all-atom model. However, although CG modeling may allow an exhaustive investigation of the conformational space involved in long time- or length-scale processes, the results may be misleading unless the ensemble of low resolution structures observed with a CG model is a low resolution representation of the ensemble that would be observed using a more atomically detailed model. Consequently, the development of a formal statistical mechanical theory for obtaining low resolution models that are consistent with accurate high-resolution models will play a critical role in achieving the full promise of CG modeling.

Izvekov and Voth recently introduced the multiscale coarse-graining (MS-CG) method[28,29] that determines a CG interaction potential from atomistic force information through a powerful variational minimization procedure. The MS-CG method has already been applied to develop accurate CG models for simple and ionic liquids,[28,45–47] uniform and mixed lipid bilayers,[29,48] small peptides,[49] carbon nanoparticles,[50] and mixed resolution models of transmembrane proteins.[51] The present work develops a formal theoretic framework for the MS-CG method, expanding upon and generalizing the foundations provided by previous work.[28,29] Furthermore, the present analysis introduces a general statistical mechanical theory for developing CG models that are consistent with atomically detailed (or other high-resolution) models. Moreover, it is demonstrated that the general MS-CG method may, in principle, be employed to determine a many-body potential of mean force (PMF) describing the equilibrium distribution of CG sites observed in simulations of atomically detailed models. The ensemble of structures generated by CG models employing this PMF as an interaction potential will be consistent with the ensemble of high resolution structures generated by the atomistic model. The MS-CG method thus provides a rigorous "multiscale bridge" between atomistic and CG models.[23]

In investigating the statistical mechanical foundations of the MS-CG method, it is convenient to introduce a precise definition of consistent CG models. For the present discussion, a CG model of a system is "consistent" with a particular atomistic model of the same system if (1) each CG coordinate and momentum has been assigned a well defined meaning as a linear combination of the coordinates and momenta of a subset of the atoms in the atomistic model and if (2) the equilibrium distribution of coordinates and momenta of the CG model is equal to the distribution determined by the equilibrium distribution function of the atomistic model. Alternatively, a CG model is consistent in the CG configuration space if each CG coordinate has been assigned a well defined meaning as a linear combination of the coordinates of a subset of the atoms in the atomistic model and the equilibrium distribution of coordinates of the CG model is equal to that implied by the atomistic equilibrium distribution. (In this latter case, either the CG momenta are not explicitly treated, as in Monte Carlo simulations of CG models,[52] or the CG momenta are employed in simulating the CG model, but only the resulting equilibrium distribution of CG coordinates is consistent with that implied by the atomistic model.)

The remaining sections of this paper are organized as follows: In Sec. II, we consider a molecular system (e.g., a biological structure) for which we have both an atomistic model and a CG model such that we can specify the relationship between the two descriptions in terms of a linear mapping operator that maps the atomistic coordinates (and, when appropriate, momenta) onto the CG coordinates (and momenta). Then, we discuss the quantitative implications of the assertion that the two models of the same system are consistent in the sense discussed above. Sufficient conditions for consistency are then derived. The analysis shows that under a wide set of conditions the "potential energy function" (i.e., many-body PMF) of the CG model is completely determined (except for an arbitrary additive constant) by the potential energy function of the atomistic model and the mapping operator if the two models are consistent. In Sec. III, the relationship between the two potential energy functions is expressed in terms of a variational principle. We show that this principle suggests algorithms for calculating the CG potential function from computer simulations of the atomistic system. In particular, Sec. III B demonstrates that this variational principle forms the fundamental basis of the MS-CG method introduced in earlier work.[28,29] Section IV provides an overview and discussion of these results, including a summary of how to construct a CG model that is consistent with a specific atomistic model. Section V provides concluding remarks. Numerical aspects of implementing the construction of consistent CG models are discussed in Paper II.

II. "CONSISTENT" COARSE-GRAINED MODELS

A. Atomistic and coarse-grained descriptions of a system

The present work considers a high resolution and a low resolution model for a given molecular system. The high resolution model will be referred to as an atomistic model, and the fundamental interacting particles for the atomistic model will be referred to as "atoms." The low resolution model will be referred to as a CG model and the fundamental interacting particles for the CG model will be referred to as CG "sites." However, the following framework applies quite generally to relate high and low resolution particle-based models of the same system and applies, e.g., when some of the atoms are actually united atoms with light atoms associated with heavy atoms and/or when some of the sites correspond to individual atoms. (For the moment, we restrict our attention to models that have no rigid intramolecular constraints.)

The instantaneous dynamical state of the atomistic model is specified by the values of the Cartesian coordinates $\mathbf{r}^n=\{\mathbf{r}_1,\ldots,\mathbf{r}_n\}$ and momenta $\mathbf{p}^n=\{\mathbf{p}_1,\ldots,\mathbf{p}_n\}$ of the n atoms of the atomistic model for the system. The atomistic Hamiltonian is

$$h(\mathbf{r}^n,\mathbf{p}^n)=\sum_{i=1}^{n}\frac{1}{2m_i}\mathbf{p}_i^2+u(\mathbf{r}^n). \tag{1}$$

The physical meaning of $\mathbf{p}_i$ is $m_i\dot{\mathbf{r}}_i$. The equilibrium probability density of dynamical states in the canonical ensemble is

$$p_{rp}(\mathbf{r}^n,\mathbf{p}^n) = p_r(\mathbf{r}^n)p_p(\mathbf{p}^n), \tag{2}$$

where

$$p_r(\mathbf{r}^n) \propto \exp(-u(\mathbf{r}^n)/k_BT), \tag{3}$$

$$p_p(\mathbf{p}^n) \propto \exp\left(-\sum_{i=1}^{n} \mathbf{p}_i^2/2m_ik_BT\right). \tag{4}$$

Similarly, the instantaneous dynamical state of the CG model is specified by the values of the Cartesian coordinates $\mathbf{R}^N=\{\mathbf{R}_1,\ldots,\mathbf{R}_N\}$ and momenta $\mathbf{P}^N=\{\mathbf{P}_1,\ldots,\mathbf{P}_N\}$ of the N CG sites of the CG model of the system. (To highlight the similarities between the atomistic and CG models while making the distinctions clear, atomistic phase variables and their functions will be represented with lower case symbols, while CG phase variables and functions will be represented with capitalized symbols.) The CG Hamiltonian is

$$H(\mathbf{R}^N,\mathbf{P}^N) = \sum_{I=1}^{N} \frac{1}{2M_I}\mathbf{P}_I^2 + U(\mathbf{R}^N). \tag{5}$$

The physical meaning of $\mathbf{P}_I$ is $M_I\dot{\mathbf{R}}_I$. The equilibrium probability density of dynamical states in the canonical ensemble is

$$P_{RP}(\mathbf{R}^N,\mathbf{P}^N) = P_R(\mathbf{R}^N)P_P(\mathbf{P}^N), \tag{6}$$

where

$$P_R(\mathbf{R}^N) \propto \exp(-U(\mathbf{R}^N)/k_BT), \tag{7}$$

$$P_P(\mathbf{P}^N) \propto \exp\left(-\sum_{I=1}^{N} \mathbf{P}_I^2/2M_Ik_BT\right). \tag{8}$$

If the CG dynamics does not involve momenta (e.g., Monte Carlo dynamics or Smoluchowski dynamics), the state is specified by the positions only. In this case, the equilibrium probability density of dynamical states is given by Eq. (7) where $U(\mathbf{R}^N)$ is the appropriate CG potential energy function.

The CG dynamical model regards the sites as structureless mass points. We assume that the model has been constructed so that each CG coordinate has a well defined physical meaning in terms of the coordinates of the atomistic model. For example, one specific CG site might correspond to the center of mass of a specific set of atoms on a molecule, another specific CG site might correspond to the center of mass of a specific molecule, and another specific site might correspond to the position of a single atom. The physical meaning of the positions of the CG sites is specified by a linear mapping operator $\mathbf{M}_{\mathbf{R}}^N(\mathbf{r}^n)=\{\mathbf{M}_{\mathbf{R}1}(\mathbf{r}^n),\ldots,\mathbf{M}_{\mathbf{R}N}(\mathbf{r}^n)\}$ of the form

$$\mathbf{M}_{\mathbf{R}I}(\mathbf{r}^n) = \sum_{i=1}^{n} c_{Ii}\mathbf{r}_i \quad \text{for } I=1,\ldots,N. \tag{9}$$

Thus, the physical meaning of $\mathbf{R}_I$ in terms of the atomistic model is $\mathbf{M}_{\mathbf{R}I}(\mathbf{r}^n)$. For CG models that use momentum, it follows that the physical meaning of the CG momentum $\mathbf{P}_I$ is

$$\mathbf{M}_{\mathbf{P}I}(\mathbf{p}^n) = M_I\sum_{i=1}^{n} c_{Ii}\mathbf{p}_i/m_i \quad \text{for } I=1,\ldots,N. \tag{10}$$

The collection of these N functions is a linear mapping operator denoted by $\mathbf{M}_{\mathbf{P}}^N(\mathbf{p}^n)$.

B. Definition of a consistent CG model

Following the definition given in the Introduction, we say that the CG model is consistent with the atomistic model in phase space if the equilibrium joint probability density of CG coordinates and momenta, as given by Eq. (6), is equal to that implied by the atomistic probability density [Eq. (2)] together with the mapping operators Eqs. (9) and (10). Also, a CG model is consistent with the atomistic model in configuration space if the equilibrium probability density of CG coordinates in Eq. (7) is equal to that implied by the atomistic distribution [Eq. (3)] together with Eq. (9).

We are concerned with understanding the conditions under which a CG model for a physical system is consistent with an atomistic model for the same physical system. It should be intuitively clear that in order for a specific CG model to be consistent with a specific atomistic model, there must be a relationship between the CG potential $U(\mathbf{R}^N)$ and the atomistic potential $u(\mathbf{r}^n)$. As we shall see, $u(\mathbf{r}^n)$ and the mapping operator $\mathbf{M}_{\mathbf{R}}^N(\mathbf{r}^n)$ determine $U(\mathbf{R}^N)$ for a consistent CG model uniquely, except for an additive constant. It is less obvious, but in fact true, that consistency is obtained only if the mapping operator satisfies certain restrictions. Moreover, for CG models that are consistent in phase space, the CG masses are determined by the atomic masses and the mapping operator.

In this paper, we will derive sufficient conditions for a specific CG model to be consistent with a specific atomistic model of the same system. We shall proceed by evaluating the equilibrium probability density of CG coordinates and momenta as determined by the equilibrium probability density of atomistic variables and the mapping operator, and by showing that under a well chosen set of conditions the CG model has the same equilibrium probability density.

C. Comments on the mapping operator

Any reasonable mapping must satisfy the condition that if all the atoms in an atomistic system are translated by the same vector displacement $\mathbf{r}$, then all the CG sites are similarly translated by the same displacement. Hence, we impose the condition that

$$\sum_{i=1}^{n} c_{Ii} = 1 \quad \text{for all } I. \tag{11}$$

It is convenient at this point to define two special sets of atoms for each of the N CG sites. For each site I, a set of *involved* atoms, $\mathcal{I}_I$, and a set of *specific* atoms, $\mathcal{S}_I$, may be defined,

$$\mathcal{I}_I = \{i|c_{Ii} \neq 0\}, \tag{12}$$

$$\mathcal{S}_I = \{i | c_{Ii} \neq 0 \text{ and } c_{Ji} = 0 \text{ for all } J \neq I\}. \quad (13)$$

An atom i in the atomistic model is involved in a CG site I if and only if the atom provides a nonzero contribution to the sum in Eq. (9). Similarly, the atom i is specific to site I if and only if the atom is involved in site I and is not involved in the definition of any other site.

D. The consistency conditions

The atomistic equilibrium probability density in Eq. (2) and the mapping operators in Eqs. (9) and (10) imply the following equilibrium probability density for the CG variables:

$$p_{RP}(\mathbf{R}^N, \mathbf{P}^N) = \int d\mathbf{r}^n \int d\mathbf{p}^n p_{rp}(\mathbf{r}^n, \mathbf{p}^n) \times \delta(\mathbf{M}_{\mathbf{R}}^N(\mathbf{r}^n) - \mathbf{R}^N) \delta(\mathbf{M}_{\mathbf{P}}^N(\mathbf{p}^n) - \mathbf{P}^N) \quad (14)$$

$$= p_R(\mathbf{R}^N) p_P(\mathbf{P}^N) \quad (15)$$

where

$$p_R(\mathbf{R}^N) = \int d\mathbf{r}^n p_r(\mathbf{r}^n) \delta(\mathbf{M}_{\mathbf{R}}^N(\mathbf{r}^n) - \mathbf{R}^N),$$

$$p_P(\mathbf{P}^N) = \int d\mathbf{p}^n p_p(\mathbf{p}^n) \delta(\mathbf{M}_{\mathbf{P}}^N(\mathbf{p}^n) - \mathbf{P}^N), \quad (16)$$

and

$$\delta(\mathbf{M}_{\mathbf{R}}^N(\mathbf{r}^n) - \mathbf{R}^N) \equiv \prod_{I=1}^{N} \delta(\mathbf{M}_{\mathbf{R}I}(\mathbf{r}^n) - \mathbf{R}_I),$$

$$\delta(\mathbf{M}_{\mathbf{P}}^N(\mathbf{p}^n) - \mathbf{P}^N) \equiv \prod_{I=1}^{N} \delta(\mathbf{M}_{\mathbf{P}I}(\mathbf{p}^n) - \mathbf{P}_I).$$

The CG model will be consistent with the atomistic model in phase space if and only if P_{RP} in Eq. (6) is equal to p_{RP} in Eq. (14), i.e.,

$$P_{RP}(\mathbf{R}^N, \mathbf{P}^N) = p_{RP}(\mathbf{R}^N, \mathbf{P}^N). \quad (17)$$

Since both expressions factorize into position and momentum dependent parts, the CG model will be consistent with the atomistic model in phase space if and only if the two following relationships hold:

$$P_R(\mathbf{R}^N) = p_R(\mathbf{R}^N), \quad (18)$$

$$P_P(\mathbf{P}^N) = p_P(\mathbf{P}^N). \quad (19)$$

These are equivalent to the following equations:

$$\exp(-U(\mathbf{R}^N)/k_BT) \propto \int d\mathbf{r}^n \exp(-u(\mathbf{r}^n)/k_BT) \times \delta(\mathbf{M}_{\mathbf{R}}^N(\mathbf{r}^n) - \mathbf{R}^N) \quad (20)$$

$$\exp\left(-\sum_{I=1}^{N} \mathbf{P}_I^2/2M_Ik_BT\right) \propto \int d\mathbf{p}^n \exp\left(-\sum_{i=1}^{n} \mathbf{p}_i^2/2m_ik_BT\right) \times \delta(\mathbf{M}_{\mathbf{P}}^N(\mathbf{p}^n) - \mathbf{P}^N). \quad (21)$$

Equations (20) and (21) are sufficient conditions for the CG model to be consistent with the atomistic model in phase space. Similarly, the CG model will be consistent with the atomistic model in configuration space if and only if Eq. (18) holds. This equation is equivalent to Eq. (20). Thus, Eq. (20) is a sufficient condition for the CG model to be consistent with the atomistic model in configuration space.

Equation (20) implies that the CG potential for a consistent CG model is a many-body PMF that is completely determined (except for an undetermined additive constant) by the atomistic potential and the mapping operator. The many-body PMF is, in fact, a conditioned free energy surface in the coordinate space of the CG variable. This may be referred to as a CG "potential energy function," but the distinction from the atomistic potential energy function is clear from Eq. (20). Equation (20) also leads to a practical algorithm for evaluating $U(\mathbf{R}^N)$ from simulations of the atomistic model. This is discussed in Sec. III.

If a CG model is to be consistent in phase space, Eq. (21) must also be satisfied. This equation can be satisfied only for certain choices of mapping operators. Moreover, this equation imposes some conditions on the CG masses M_I. This is discussed in Sec. II F.

E. The coarse-grained force field

Equation (20) determines the CG potential $U(\mathbf{R}^N)$ for a consistent model in terms of the atomistic potential,

$$U(\mathbf{R}^N) = -k_BT \ln z(\mathbf{R}^N) + (\text{const}),$$

where

$$z(\mathbf{R}^N) \equiv \int d\mathbf{r}^n \exp(-u(\mathbf{r}^n)/k_BT) \delta(\mathbf{M}_{\mathbf{R}}^N(\mathbf{r}^n) - \mathbf{R}^N).$$

The gradients of this $U(\mathbf{R}^N)$ determine the "force field" for a consistent CG model. The present subsection derives an expression for this CG force field as a certain type of equilibrium average of linear combinations of the atomistic forces.

The CG force field is given by

$$\mathbf{F}_I(\mathbf{R}^N) = -\frac{\partial U(\mathbf{R}^N)}{\partial \mathbf{R}_I} = \frac{k_BT}{z(\mathbf{R}^N)} \frac{\partial z(\mathbf{R}^N)}{\partial \mathbf{R}_I} = \frac{k_BT}{z(\mathbf{R}^N)} \int d\mathbf{r}^n \exp(-u(\mathbf{r}^n)/k_BT) \times \prod_{J(\neq I)} \delta(\mathbf{M}_{\mathbf{R}J}(\mathbf{r}^n) - \mathbf{R}_J) \frac{\partial}{\partial \mathbf{R}_I} \delta\left(\sum_{i \in \mathcal{I}_I} c_{Ii}\mathbf{r}_i - \mathbf{R}_I\right). \quad (22)$$

This may be expressed in terms of atomistic forces by expressing the partial derivative in Eq. (22) as a linear combination of partial derivatives with regard to atomistic coordi-

nates, $\partial/\partial \mathbf{r}_i$, and integrating by parts, so that the derivatives are now acting on $u(\mathbf{r}^n)$ in the Boltzmann factor. It is straightforward to see that the identity

$$\frac{\partial}{\partial \mathbf{R}_I}\delta\left(\sum_{i\in\mathcal{I}_I} c_{Ii}\mathbf{r}_i - \mathbf{R}_I\right) = -\frac{1}{c_{Ik}}\frac{\partial}{\partial \mathbf{r}_k}\delta\left(\sum_{i\in\mathcal{I}_I} c_{Ii}\mathbf{r}_i - \mathbf{R}_I\right) \tag{23}$$

holds for any $k \in \mathcal{I}_I$ because $c_{Ik} \neq 0$ by the definition in Eq. (12). Although this identity may be employed in Eq. (22), if the atom k is involved in the definition of any other CG site, $J \neq I$, then subsequent integration by parts will be complicated by the dependence of the remaining $N-1$ mapping operators on $\mathbf{r}_k$. This complication may be avoided by defining a set of constant coefficients $\{d_{Ii}\}$ such that $d_{Ii} \neq 0$ only if atom i is specific to CG site I and such that

$$\sum_{j\in\mathcal{S}_I} d_{Ij} = 1 \quad \text{for all } I. \tag{24}$$

It is then easily verified that

$$\frac{\partial}{\partial \mathbf{R}_I}\delta\left(\sum_{i\in\mathcal{I}_I} c_{Ii}\mathbf{r}_i - \mathbf{R}_I\right) = -\sum_{j\in\mathcal{S}_I}\frac{d_{Ij}}{c_{Ij}}\frac{\partial}{\partial \mathbf{r}_j}\delta\left(\sum_{i\in\mathcal{I}_I} c_{Ii}\mathbf{r}_i - \mathbf{R}_I\right). \tag{25}$$

In order to obtain such an equation for each I, we assume that for each CG site there is at least one atom that is specific to the site. Because the partial derivative $\partial/\partial \mathbf{R}_I$ in Eq. (22) has been expressed in terms of partial derivatives with regard to the positions of atoms that are not involved in the definition of any other site, the integration by parts may now be performed without further complication. The result is

$$\mathbf{F}_I(\mathbf{R}^N) = \langle \mathcal{F}_I(\mathbf{r}^n)\rangle_{\mathbf{R}^N}, \tag{26}$$

where

$$\mathcal{F}_I(\mathbf{r}^n) = \sum_{j\in\mathcal{S}_I}\frac{d_{Ij}}{c_{Ij}}\mathbf{f}_j(\mathbf{r}^n), \tag{27}$$

$\mathbf{f}_j(\mathbf{r}^n) = -\partial u(\mathbf{r}^n)/\partial \mathbf{r}_j$, and the angular brackets denote an average of the form

$$\langle g(\mathbf{r}^n)\rangle_{\mathbf{R}^N} \equiv \frac{\int d\mathbf{r}^n \exp(-u(\mathbf{r}^n)/k_BT)\,\delta(\mathbf{M}_{\mathbf{R}}^N(\mathbf{r}^n) - \mathbf{R}^N)\,g(\mathbf{r}^n)}{\int d\mathbf{r}^n \exp(-u(\mathbf{r}^n)/k_BT)\,\delta(\mathbf{M}_{\mathbf{R}}^N(\mathbf{r}^n) - \mathbf{R}^N)} \tag{28}$$

for any continuous function $g(\mathbf{r}^n)$ of the atomistic coordinates. The CG force field $\mathbf{F}(\mathbf{R}^N)$ in Eq. (26) can be regarded as a conditional expectation value of $\mathcal{F}$ for an atomistic system, given that the atomistic system is in a configuration such that $\mathbf{M}_{\mathbf{R}}^N(\mathbf{r}^n) = \mathbf{R}^N$.

Equation (26) is a major result of this subsection. It expresses the force on CG site I in terms of a certain type of equilibrium average, for the atomistic model, of $\mathcal{F}_I(\mathbf{r}^n)$, which is a linear combination of the atomistic forces acting on the atoms that are specific to CG site I. It was derived from Eq. (20) using only one assumption, namely, that every CG site has at least one atom that is specific to it. It provides the basis for a variational principle and a practical algorithm for calculating the many-body PMF (potential) from simulations of the atomistic system, as discussed below in Sec. III.

In addition, the logic of the discussion can be reversed to show that Eq. (26), plus the assumption that every CG site has at least one atom that is specific to it, implies Eq. (20). However, Eq. (20) is precisely the condition for consistency in configuration space and is one of the two conditions for consistency in phase space. Thus, we have the basis for stating sufficient conditions for consistency of a CG model with an atomistic model. A CG model will be consistent in configuration space with a given atomistic model if there is at least one atom specific to each site and if the force on each CG site I in a given CG configuration $\mathbf{R}^N$ is given by Eq. (26). A complete set of sufficient conditions is summarized below in Sec. II G.

F. Consistency in momentum space

For CG models that include momenta, Eq. (21) gives the consistency condition in momentum space. The left side of this equation is a product of separate factors, each of which depends on only one $\mathbf{P}_I$. In order for the right side to have this property, it is necessary and sufficient that no atom i be involved in the definition of more than one CG site.

If this condition is satisfied, then it is straightforward to show that the factor on the right associated with site I has a Gaussian form with zero mean, as does the corresponding factor on the left. If the second moments of the factors on left and right are equal, then the CG masses satisfy

$$M_I = \left(\sum_{i\in\mathcal{I}_I}\frac{c_{Ii}^2}{m_i}\right)^{-1} \quad \text{for all } I. \tag{29}$$

Equation (29) is a major result of this subsection. It expresses the mass of each CG site in terms of the atomistic masses and the coefficients in the mapping function. It was derived from Eq. (21), which is one of the two necessary requirements for consistency in phase space, using only one assumption, namely, that no atom is involved in the definition of more than one site. In addition, the logic of the derivation can be reversed to show that Eq. (29) plus this assumption imply Eq. (21), which is also one of the two sufficient conditions for consistency in phase space. Consequently, a CG model will generate an equilibrium distribution of momenta that is consistent with a given atomistic model if no atom is involved in the definition of more than one CG site and if the mass of each site is defined by Eq. (29). In conjunction with the result of Sec. II E, this allows us to construct a complete set of sufficient conditions for consistency of the CG model with the atomistic model.

G. Summary of sufficient conditions for consistency

We can now summarize a set of sufficient conditions that imply that a particular CG model is consistent with a particular atomistic model. These conditions are sufficient for atomistic models that have no rigid intramolecular constraints.

(1) The physical meaning of the location of each CG site in terms of the locations of the atoms is expressed in linear equations of the form of Eq. (9).
(2) Each CG site has at least one atom that is specific to that site.

(3) The CG forces satisfy Eq. (26), which relates them to equilibrium averages in the atomistic canonical ensemble.

These three are a set of sufficient conditions for consistency in configuration space. For consistency in phase space, we have additional conditions:

(4) No atom is involved in the definition of more than one CG site.

(5) The CG masses satisfy Eq. (29).

H. The effect of rigid intramolecular constraints on the problem of consistency

The previous discussion applies only to the case in which the atomistic model has no rigid intramolecular constraints. In this situation, each Cartesian coordinate of each atom corresponds to a mechanical degree of freedom of the system. Atomically detailed molecular models frequently employ rigid mechanical constraints between atoms to increase the efficiency of MD simulations.[53] The presence of such constraints has a significant impact upon the equilibrium statistical mechanics of a model because the number of mechanical degrees of freedom is less than the number of Cartesian coordinates.[53–60] As a result, there are additional considerations that arise in specifying sufficient conditions for a CG model to be consistent with an atomistic model. Here, we will just summarize the results for a case that is likely to be appropriate for most instances in which CG models are used. The analysis follows the work of Ciccotti and co-workers,[53,59,60] which shows how to use Cartesian coordinates of an atomistic system with rigid constraints in describing the mechanics and statistical mechanics of the system. A detailed discussion of other cases will be presented in a subsequent paper.

Suppose the atomistic model has l translationally invariant holonomic constraints,[55] each enforcing a fixed relative geometry among constrained atoms within a single molecule,

$$\sigma_t(\mathbf{r}^n) = 0, \tag{30}$$

where t labels the constraint. In order for each constraint to be satisfied at all times, the velocities of the constrained atoms must satisfy the condition

$$\dot{\sigma}_t(\mathbf{r}^n, \mathbf{p}^n) = 0. \tag{31}$$

The set of l constraints, each of which affects only a particular set of atoms, allows us to partition the atoms into nonintersecting "constrained sets," such that (1) no constraints exist between atoms in distinct constrained sets and (2) the atoms within a constrained set are all either directly or indirectly connected by constraints. It will be convenient if we specify that any atom that is not constrained by other atoms is regarded as a constrained set containing one atom. Given this partitioning of the atoms, the equilibrium probability density for the coordinates and momenta of the atomistic model is

$$p^c_{rp}(\mathbf{r}^n, \mathbf{p}^n) \propto \exp(-h(\mathbf{r}^n, \mathbf{p}^n)/k_B T) \prod_{\eta=1}^{n_c} \mathcal{P}(\{\mathbf{r}\}_\eta, \{\mathbf{p}\}_\eta; \mathcal{R}_\eta). \tag{32}$$

Here, η is a label that denotes constrained sets, n_c is the number of constrained sets, $\{\mathbf{r}\}_\eta$ and $\{\mathbf{p}\}_\eta$ denote the coordinates and momenta of the atoms in constrained set $\mathcal{C}_\eta$, and $\mathcal{R}_\eta$ stands for the particular set of constraints that act on the constrained set $\mathcal{C}_\eta$. Here, $h(\mathbf{r}^n, \mathbf{p}^n)$ is of the usual form [Eq. (1)] for unconstrained systems and $\mathcal{P}(\{\mathbf{r}\}_\eta, \{\mathbf{p}\}_\eta; \mathcal{R}_\eta)$ describes the correlations among the atoms in constrained set $\mathcal{C}_\eta$ that are induced by the constraints. For constrained sets that have only one atom, $\mathcal{P}$ is, of course, equal to unity. An explicit formula for $\mathcal{P}$ for nontrivial constrained sets involving more than one atom can be obtained from the work of Ciccotti and Ryckaert.[53] The detailed form depends on the number and nature of the various constraints acting on the constrained set of atoms.

We consider the case in which the atomistic model has rigid intramolecular constraints while the CG model has none. In this situation, the statistical mechanics of the CG system is described by the same equations as in the previous discussion, but that of the atomistic system is more complicated as discussed just above.

For this case, however, the derivation of Eq. (26), which is associated with consistency in configuration space, proceeds as in the case without constraints, and the same result is obtained, provided that condition (6) holds. This condition allows the integration by parts to get the analog of Eq. (26) despite the additional factors in the integrand derived from the $\mathcal{P}$ functions. That is,

(6) if a member i of a constrained set $\mathcal{C}_\eta$ of atoms is specific to site I and $d_{Ii} \neq 0$, then all members of the constrained set are specific to site I, and d_{Ii}/c_{Ii} has the same value for all $i \in C_\eta$.

Condition (6) implies that if the atomistic force $\boldsymbol{\mathcal{F}}_I$ associated with site I contains a contribution from the atomistic force on any member of a constrained set, then $\boldsymbol{\mathcal{F}}_I$ contains contributions from every member of the constrained set, and the sum of all these contributions is equal to a numerical factor times the sum of all the atomistic forces on the constrained set that is derived from gradients of the atomistic potential u [see Eq. (27)]. In the atomistic model, there are also fluctuating constraint forces acting among the members of the constrained set, but their sum is always zero because of momentum conservation. Thus, when condition (6) is satisfied, the atomistic force $\boldsymbol{\mathcal{F}}_I$ associated with site I depends on the forces on the atoms in the constrained set only as a function of the total force on the atoms in the constrained set.

To prove consistency in momentum space, it is necessary to show that the analog of Eq. (15) holds, i.e., that p_{RP} for the situation with constraints factorizes into position and momentum dependent factors and that the analog of Eq. (21) holds. [The analogs of both Eqs. (15) and (21) contain factors resulting from the $\mathcal{P}$ functions and hence are more complicated than Eqs. (15) and (21), respectively.] This can be proven if the following conditions holds:

(7) If one member of a constrained set $\mathcal{C}_\eta$ of atoms is involved with site I, then all members of the constrained set are involved with the same site, and c_{Ii}/m_i has the same value for all $i \in C_\eta$.

Condition (7) implies that if atoms in a constrained set are involved in a particular CG site, the mapping operator for that site depends on the coordinates of the atoms in the constrained set only as a function of the center of mass of the constrained set.

Thus, a sufficient set of conditions for consistency in phase space, in the case that the atomistic model has rigid intramolecular constraints while the CG model has none, is conditions (1)–(5) from Sec. II G plus conditions (6) and (7) above. A sufficient set of conditions for consistency in configuration space is conditions (1)–(3) plus condition (6). More complicated constraint scenarios required additional analysis and conditions. These will be the focus of future research.

III. VARIATIONAL PRINCIPLE FOR MULTISCALE COARSE-GRAINING

The analysis of the previous section identified sufficient conditions for a particular CG model to be consistent with a particular atomistic model. In a consistent model, the CG potential energy function, $U(\mathbf{R}^N)$, is a many-body PMF determined by the atomistic interaction potential, $u(\mathbf{r}^n)$, according to Eq. (20). The CG force field determined by gradients of the PMF is then related to the atomistic force field according to Eq. (26). The present section discusses the MS-CG variational principle for determining this CG force field and demonstrates that this variational principle forms the fundamental basis for the MS-CG method orginally introduced by Izvekov and Voth.[28,29] Therefore, the MS-CG method may be used, in principle, for systematically developing CG models that will be consistent with a given atomistic model.

A. Variational principle for the many-body PMF

A CG force field is a set of real continuous functions, $\mathbf{G}_I(\mathbf{R}^N)$, of the CG configuration, $\mathbf{R}^N$, for each site, $I=1,\ldots,N$. It will be convenient in the following discussion to consider this set as a single function whose arguments include both I and $\mathbf{R}^N$, in which case a CG force field can be denoted as $\mathbf{G}$. We consider the vector space of real functions of this type, which we shall refer to as the vector space of CG force fields. This space includes the (atomistically consistent) CG force field $\mathbf{F}$ that is determined by Eq. (22), as well as all possible approximations to this force field.

For an arbitrary member $\mathbf{G}$ in this vector space of CG force fields, we define the functional

$$\chi^2[\mathbf{G}] = \frac{1}{3N}\left\langle \sum_{I=1}^{N} |\boldsymbol{\mathcal{F}}_I(\mathbf{r}^n) - \mathbf{G}_I(\mathbf{M}_\mathbf{R}^N(\mathbf{r}^n))|^2 \right\rangle, \tag{33}$$

where the angular brackets denote an equilibrium canonical ensemble average for the atomistic model. Then, defining $\boldsymbol{\Delta}=\mathbf{G}-\mathbf{F}$, it follows that

$$\begin{aligned}\chi^2[\mathbf{G}] &= \chi^2[\mathbf{F}] + \frac{1}{3N}\left\langle \sum_{I=1}^{N} |\boldsymbol{\Delta}_I(\mathbf{M}_\mathbf{R}^N(\mathbf{r}^n))|^2 \right\rangle \\ &\quad - \frac{2}{3N}\left\langle \sum_{I=1}^{N} \boldsymbol{\Delta}_I(\mathbf{M}_\mathbf{R}^N(\mathbf{r}^n)) \cdot (\boldsymbol{\mathcal{F}}_I(\mathbf{r}^n) - \mathbf{F}_I(\mathbf{M}_\mathbf{R}^N(\mathbf{r}^n))) \right\rangle \\ &= \chi^2[\mathbf{F}] + \frac{1}{3N}\left\langle \sum_{I=1}^{N} |\boldsymbol{\Delta}_I(\mathbf{M}_\mathbf{R}^N(\mathbf{r}^n))|^2 \right\rangle \geqslant \chi^2[\mathbf{F}]. \end{aligned} \tag{34}$$

The third term on the right hand side of the first equality vanishes as a result of the lemmas proved in the Appendix. The $\geqslant$ in Eq. (34) holds as an equality if and only if $\mathbf{G}_I(\mathbf{R}^N)=\mathbf{F}_I(\mathbf{R}^N)$ for all I and $\mathbf{R}^N$.

This analysis provides the theoretical basis for the following variational principle: The global minimum of the functional $\chi^2[\mathbf{G}]$ for $\mathbf{G}$ in the vector space of CG force fields is achieved when $\mathbf{G}$ is $\mathbf{F}$, i.e., the appropriate CG force field for a consistent model, and the global minimum is unique.

Equation (34) suggests a physically relevant definition of a norm in the vector space of CG force fields. For any vector $\mathbf{G}$ in the space,

$$\begin{aligned}\|\mathbf{G}\| &\equiv \left\langle \sum_{I=1}^{N} |\mathbf{G}_I(\mathbf{M}_\mathbf{R}^N(\mathbf{r}^n))|^2 \right\rangle^{1/2} \\ &= \left(\int d\mathbf{R}^N \sum_{I=1}^{N} |\mathbf{G}_I(\mathbf{R}^N)|^2 p_R(\mathbf{R}^N) \right)^{1/2}, \end{aligned} \tag{35}$$

where $p_R(\mathbf{R}^N)$, given in Eq. (16), is the equilibrium probability density of CG coordinates that is implied by the equilibrium properties of the atomistic model. The norm $\|\mathbf{G}\|$ can be regarded as the length of the vector $\mathbf{G}$. With this definition, Eq. (34) becomes

$$\chi^2[\mathbf{G}] = \chi^2[\mathbf{F}] + \frac{1}{3N}\|\mathbf{G}-\mathbf{F}\|^2. \tag{36}$$

Suppose that we identify a finite set of N_D linearly independent vectors in the vector space of CG force fields, $\boldsymbol{\mathcal{G}}_D$ for $D=1,\ldots,N_D$. This set of vectors forms a finite, and hence incomplete, basis set $\{\boldsymbol{\mathcal{G}}_D\}$ for the vector space. We consider the subspace spanned by this basis, and we find the vector in that subspace that minimizes $\chi^2[\mathbf{G}]$. Using the same reasoning that led to Eq. (34), it is straightforward to show that the minimum exists and is unique. Moreover, from Eq. (36) it is clear that the force field $\mathbf{G}$ in the given vector subspace that minimizes χ^2 is also the force field in the subspace that minimizes $\|\mathbf{G}-\mathbf{F}\|$; i.e., it is the member of the subspace that minimizes the distance [as defined by Eq. (35)] from the CG force field $\mathbf{F}$ determined by the exact many-body PMF. Consequently, this variational principle can be used to construct the following practical algorithm for calculating approximations to the PMF, $U(\mathbf{R}^N)$, from simulations of an atomistic model:

(1) Choose a finite set of N_D linearly independent force field vectors $\boldsymbol{\mathcal{G}}_D$ for $D=1,\ldots,N_D$, defining a vector subspace of functions $\mathbf{G}_I(\mathbf{R}^N)$. Each force field vector $\boldsymbol{\mathcal{G}}_{I;D}(\mathbf{R}^N)$ should be of the form

$$\mathcal{G}_{I;D}(\mathbf{R}^N) = -\frac{\partial U_D(\mathbf{R}^N)}{\partial \mathbf{R}_I}, \tag{37}$$

where $U_D(\mathbf{R}^N)$ should be a scalar function that is differentiable with regard to all its arguments.

(2) Consider

$$\mathbf{G}_I(\mathbf{R}^N) = \sum_{D=1}^{N_D} \phi_D \mathcal{G}_{I;D}(\mathbf{R}^N). \tag{38}$$

Calculate $\chi^2[\mathbf{G}]$ by approximating the ensemble average in Eq. (33) by an appropriate average over computer simulation trajectories of the atomistic model designed to sample from a canonical ensemble.

(3) Find the set of coefficients $\{\phi_D^*\}$ that minimize χ^2. Then, Eq. (38) with those coefficients is an approximation to the CG force field $\mathbf{F}$ determined by the many-body PMF, and

$$U^*(\mathbf{R}^N) = \sum_{D=1}^{N_D} \phi_D^* U_D(\mathbf{R}^N) \tag{39}$$

is an approximation to the atomistically consistent many-body PMF, $U(\mathbf{R}^N)$.

In the limit that the set of N_D basis functions is sufficiently complete, an accurate numerical approximation to the many-body force field $\mathbf{F}$, which is determined by gradients of the many-body PMF, can be obtained using this variational principle, provided that the simulation data are sufficiently accurate to evaluate ensemble averages with sufficiently small statistical error. In practice, the basis set will not be complete and the data will have some statistical error. As discussed in the previous paragraph, within the subspace of CG force fields spanned by any incomplete but linearly independent set of force field basis vectors $\{\mathcal{G}_D\}$, there exists a unique CG force field $\mathbf{G}$ of the form given by Eq. (38) that minimizes $\chi^2[\mathbf{G}]$ in Eq. (33) among all force fields in that subspace. In practice, for any incomplete basis set, the $\mathbf{G}$ of the form of Eq. (38) that minimizes an approximate $\chi^2[\mathbf{G}]$ calculated by averaging over computer simulation trajectories will also be unique if there are enough simulation data, and sufficient data should be used so that this is the case for the basis set chosen. Then, the combination of basis set and simulation data set determines a unique approximation to the CG force field. Moreover, as a result of the way the basis set is constructed, the force field calculated from applying the variational principle will be the gradient of a potential energy function and hence suitable for use in computer simulations with the CG model. Finally, the force field determined by applying this variational principle within the subspace of CG force fields spanned by the given basis set will be the member of the subspace whose "distance" is the smallest from $\mathbf{F}$, subject, of course, to the effect of statistical error. In this sense, the result of the variational calculation represents an optimal approximation to the many-body CG force field derived from the PMF, given the basis set chosen and the simulation data set. A systematic improvement to this approximation can be made by improving the basis set and/or the data set. Paper II (Ref. 61) demonstrates the numerical implementation of the MS-CG variational algorithm and discusses the practical treatment of incomplete basis sets, imperfect sampling, and statistical noise within the method.

B. Previous development of the MS-CG method

The MS-CG method previously developed by Izvekov and Voth[28,29] determined a CG force field through an extension of a force-matching method originally introduced for constructing potential energy functions from *ab initio* MD data.[62] In early applications of the MS-CG method, the CG force field was determined by variationally minimizing the residual,[30]

$$\chi^2_{\mathrm{MS}}[\mathbf{F}^{\mathrm{MS}}] = \frac{1}{3n_t N}\sum_{t=1}^{n_t}\sum_{I=1}^{N} |\mathbf{f}_I(\mathbf{r}_t^n) - \mathbf{F}_I^{\mathrm{MS}}(\mathbf{M}_{\mathbf{R}}^N(\mathbf{r}_t^n))|^2 \tag{40}$$

$$= \frac{1}{3N}\left\langle \sum_{I=1}^{N} |\mathbf{f}_I(\mathbf{r}_t^n) - \mathbf{F}_I^{\mathrm{MS}}(\mathbf{M}_{\mathbf{R}}^N(\mathbf{r}_t^n))|^2 \right\rangle_t . \tag{41}$$

In Eqs. (40) and (41), the index t denotes one of n_t configurations sampled during the course of simulations of the atomistic model, $\mathbf{r}_t^n$ indicates the tth sampled atomistic configuration, and the angular brackets labeled by t represent an average over the sampled configurations. $\mathbf{F}_I^{\mathrm{MS}}(\mathbf{R}^N)$ is the MS-CG force on CG site I as a function of the CG site positions. The quantity $\mathbf{f}_I(\mathbf{r}_t^n)$ is the "atomistic force" on CG site I determined by the forces on the atoms used to define the CG site I according to the atomically detailed interaction potential, $u(\mathbf{r}^n)$. A comparison of Eqs. (33) and (41) indicates that $\chi^2_{\mathrm{MS}}=\chi^2$ and that the original MS-CG variational minimization method is a numerical implementation of the general MS-CG variational principle discussed in Sec. III A if $\mathcal{F}_I(\mathbf{r}^n)=\mathbf{f}_I(\mathbf{r}^n)$.

In previous applications of the MS-CG method,[28,29,45–51,63] CG sites have frequently been defined as the centers of mass for disjoint subsets of atoms, and the atomistic force on the CG site has been defined as

$$\mathbf{f}_I(\mathbf{r}^n) = \sum_{i\in\mathcal{I}_I} \mathbf{f}_i(\mathbf{r}^n). \tag{42}$$

Clearly, this definition is consistent with the mapping $c_{Ii}=d_{Ii}=m_i/M_I$ for $i\in\mathcal{I}_I=\mathcal{S}_I$ and $M_I=\Sigma_{i\in\mathcal{I}_I}m_i$, as long as no rigid intramolecular constraints connect atoms involved in distinct CG sites. In this case, $\mathcal{F}_I(\mathbf{r}^n)=\mathbf{f}_I(\mathbf{r}^n)$. Consequently, the variational calculation employed in these previous applications of the MS-CG method[29] is a specific case of the general MS-CG algorithm described in Sec. III A for determining an optimal approximation to the many-body PMF.

IV. DISCUSSION

CG models are becoming an increasingly important tool for studying complex long time- and long length-scale processes that cannot be adequately investigated with more atomistically detailed models. A fundamental assumption implicit in much CG modeling is that an investigation of the CG model would lead to the same conclusions as an investigation of a more detailed model if the latter investigation

were feasible. In other words, it is assumed that the CG model is consistent with a high resolution model. Previous studies have implicitly employed various notions of consistency of CG models. Quite commonly, the assessment of consistency has focused on reproducing either particular thermodynamic properties[24,44,64] or local structural information reported in pair distribution functions.[27,32,38,41] The present work proposes and investigates an explicit definition of consistency: a CG model is consistent with a particular atomistic model if the equilibrium distribution of CG structures (and, when appropriate, also momenta) generated by the CG model is equal to the distribution determined by the specified CG mapping and the equilibrium atomistic configuration space (or phase space) distribution function. This definition is highly restrictive but allows the present analysis and is sufficient to ensure that the ensemble of low resolution structures sampled with a consistent CG model is structurally equivalent to the ensemble of structures obtained by coarse-graining the ensemble of structures sampled with the high-resolution model. Moreover, Eq. (26) indicates that such consistent CG models certainly exist, regardless of the technical difficulty in developing such models.

The analysis of the present work suggests the following prescription for constructing a CG model that is consistent with a given atomistic model. This prescription is applicable if the CG model has no rigid intramolecular constraints. The first four steps are sufficient to guarantee consistency of the CG model with the atomistic model in configuration space. The last three steps should also be included if consistency in phase space is desired. They are as follows:

(1) For each molecule in the atomistic model, the number of CG sites describing the molecule and the atoms involved in the definition of each site must be chosen. An atom is specific to a site if it is involved in the definition of that site and is not involved in the definition of any other site. There must be at least one atom that is specific to each site.
(2) The c_{Ii} coefficients must be chosen subject to the simple normalization condition in Eq. (11). A coefficient c_{Ii} may be nonzero only if atom i is involved in the definition of site I.
(3) The set of d_{Ii} coefficients must be chosen subject to the normalization condition in Eq. (24). A coefficient d_{Ii} may be nonzero only if the atom i is specific to site I. The d_{Ii} coefficients for atoms i that are members of a constrained set must satisfy condition (6) in Sec. II H.
(4) The potential U (many-body PMF) of the CG model can be calculated from simulation data for the atomistic system using the variational principle in Sec. III A. (Note that the calculation of the CG potential and forces requires only configuration data for equilibrium systems, not momentum data, so it can be obtained from either MD simulations or Monte Carlo simulations.) The MS-CG method provides a practical numerical implementation of this variational calculation.
(5) A mass M_I must be assigned to each CG site according to Eq. (29).
(6) Each atom may be involved in the definition of at most one site.
(7) If one or more members of constrained sets are involved in the definition of sites, condition (7) in Sec. II H must be satisfied.

In principle, if these conditions are satisfied, a complete set of basis functions is used in the variational calculation, and the atomistic simulation data used in the calculations have sufficiently small statistical error, the resulting CG model will be consistent with the atomistic model, in the sense defined at the beginning of this paper.

From the above discussion, it is clear that consistent CG models may be developed for a remarkably diverse set of CG mappings. Consistent CG models may associate sites with the center of mass, center of geometry, or other normalized mappings for groups of atoms, as long as the $\{c_{Ii}\}$ and $\{d_{Ii}\}$ coefficients are appropriately chosen. Consistent CG models may be developed in which some atoms in the atomistic model are not involved in the definition of any CG sites. For example, consistent solvent-free CG models[38,39,65] may be systematically developed in which the solvent molecules have been integrated out of the model, but their effect is incorporated into the many-body PMF for the remaining CG sites. Similarly, consistent mixed resolution models[51,66] may be developed in which certain parts of the system are modeled in complete atomistic detail (i.e., in some parts of the system, every atom corresponds to a separate site), while the remainder of the system is modeled in reduced CG detail (i.e., in other parts of the system, every site corresponds to more than one atom). Furthermore, CG models generating a consistent configuration space distribution may be developed in which one or more atoms are involved in the definition of multiple sites. However, as mentioned above, no atom can be involved in the definition of more than one site for a model that is consistent in phase space.

The analysis of Sec. III B demonstrated that the variational principle discussed in Sec. III A forms the fundamental basis for the MS-CG method originally developed in previous work.[28,29] The MS-CG method employs force information determined from atomically detailed MD simulations to calculate an effective interaction potential for simulations of a CG model of the same system. If the CG mapping is in accord with the development of Sec. II, if the atomistic simulations have adequately sampled the canonical distribution function for the atomically detailed model, and if the variational calculation is performed using a basis set that spans the space of all possible CG force fields, then the MS-CG method determines the exact many-body PMF for the given atomistic model. Simulations of the MS-CG model employing this PMF as a potential energy function will be consistent in configuration space with the given atomistic model. Even more significantly, though, if the variational MS-CG method is employed within the vector subspace spanned by an incomplete set of force field basis vectors, then the MS-CG method will determine the CG force field within that subspace that is closest to the force field determined by the exact many-body PMF. In other words, if the CG force field is assumed to be of a certain form, then the

MS-CG method determines the CG force field of the assumed form that is the optimal approximation to the exact many-body force field, subject to any statistical noise in the data sampled in the atomistic simulations. Moreover, the MS-CG variational principle provides a systematic methodology for improving an approximation to the exact many-body PMF by expanding the space of trial force fields, e.g., by explicitly including basis functions that describe three-body interactions between CG sites.[34,36] The variational principle discussed in Sec. III ensures that as the space of trial CG force fields is expanded by introducing additional basis functions, the optimized MS-CG force field becomes an increasingly accurate approximation to the exact many-body PMF.

Previous applications of the MS-CG method have employed pairwise additive potentials for modeling nonbonded interactions between CG sites.[28,29,45–51,63] This can be regarded as a particular, physically motivated, choice of the type of basis function to be used for the variational calculation. Since basis sets of this type are not complete, the calculated MS-CG force field is not an exact representation of the many-body CG force field $\mathbf{F}$ that is given by the gradients of the many-body PMF. Instead, the MS-CG method determines the force field within the space of functions spanned by the basis set used that is closest to this many-body CG force field, in the sense defined in Sec. III A. However, despite the fact that basis sets that employ pairwise additive potentials are not complete, previous applications of the MS-CG method have been shown to determine accurate CG models for a number of complex systems when compared with the results of atomistic MD simulations.[28,29,45–51,63] This suggests that nonbonded interactions between CG sites are often well represented, at least to a good first approximation, as pairwise additive. A recent analysis[30] has also demonstrated that the MS-CG equations for these pair potentials are related to well-known exact kinetic equations for the liquid state[67] and systematically incorporate critical three-body correlations present in the underlying atomistic model, which may be a key feature in the previous successes of the MS-CG method.

Paper II (Ref. 61) discusses the numerical implementation of the MS-CG method and some examples.

V. CONCLUDING REMARKS

The present work introduces a procedure for developing a CG model that is consistent with a more detailed atomistic model of the same system. A precise definition of consistency is given, one that is based on a mapping operator that maps each point in the atomistic phase space onto a corresponding point in the CG phase space. The present analysis identifies conditions that are sufficient for proving that a CG model and an atomistic model are consistent. On this basis, it is clear how to construct a CG model that is consistent with a given atomistic model of interest. A crucial step in this construction is the calculation of the appropriate CG potential $U(\mathbf{R}^N)$, which is, in fact, a many-body PMF for the CG sites that incorporates both energetic and entropic effects from the atomistic model. The MS-CG variational principle provides the basis for calculating this PMF using equilibrium simulations of the atomistic system. Consequently, the MS-CG method may be employed, in principle, to develop a CG model that is consistent with a given atomistic model of the same system. Assuming adequate sampling of the atomistic configuration space, the MS-CG variational principle determines an optimal approximation to the exact many-body PMF when employed to parametrize a CG potential that is not completely flexible. The numerical aspects of implementing the MS-CG method are discussed in Paper II.

ACKNOWLEDGMENTS

This research was supported by a Collaborative Research in Chemistry grant from the National Science Foundation (Grant No. CHE-0628257). W.G.N. acknowledges funding from the National Institutes of Health through a Ruth L. Kirschstein National Research Service Award postdoctoral fellowship (Grant No. 5 F32 GM076839-02). W.G.N. also gratefully acknowledges many stimulating conversations with Dr. S. Iuchi.

APPENDIX: DERIVATION OF EQUATION (34)

Equation (28) defines a certain type of average $\langle g(\mathbf{r}^n)\rangle_{\mathbf{R}^N}$ of any function $g(\mathbf{r}^n)$. The following lemmas may be immediately derived from this definition and may be used to obtain Eq. (34).

Lemma 1. If $g(\mathbf{r}^n)$ and $G(\mathbf{R}^N)$ are functions of the atomistic and CG coordinates, respectively, and if

$$g(\mathbf{r}^n) = G(\mathbf{M}_{\mathbf{R}}^N(\mathbf{r}^n)), \tag{A1}$$

then

$$\langle g(\mathbf{r}^n)\rangle_{\mathbf{R}^N} = G(\mathbf{R}^N) \quad \text{for all } \mathbf{R}^N. \tag{A2}$$

Lemma 2. If $g(\mathbf{r}^n)$ is a function that satisfies $\langle g\rangle_{\mathbf{R}^N}=0$ for all $\mathbf{R}^N$, then $\langle g\rangle=0$.

Lemma 3. If $g(\mathbf{r}^n)$ is a function that satisfies $\langle g\rangle_{\mathbf{R}^N}=0$ for all $\mathbf{R}^N$ and $J(\mathbf{R}^N)$ is an arbitrary function of the site positions, then

$$\langle g(\mathbf{r}^n)J(\mathbf{M}_{\mathbf{R}}^N(\mathbf{r}^n))\rangle_{\mathbf{R}^N} = 0 \quad \text{for all } \mathbf{R}^N. \tag{A3}$$

Hence

$$\langle g(\mathbf{r}^n)J(\mathbf{M}_{\mathbf{R}}^N(\mathbf{r}^n))\rangle = 0. \tag{A4}$$

It is easily shown that the same lemmas hold for the case of an atomistic system with intramolecular constraints if the CG model does not have any constraints.

[1] M. P. Allen and D. P. Tildesley, *Computer Simulation of Liquids* (Oxford University, New York, 1987).
[2] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*, 2nd ed. (Academic, New York, 2002).
[3] M. Karplus and J. A. McCammon, Nat. Struct. Mol. Biol. **9**, 646 (2002).
[4] E. Lindahl, B. Hess, and D. van der Spoel, J. Mol. Model. **7**, 306 (2001).
[5] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodham, W. E. Reihar III, B. Raux, M. Schlenkrich, J. C. Smith, R. Store, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus, J. Phys. Chem. B **102**, 3586 (1998).
[6] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa,

C. Chipot, R. D. Skeel, L. Kale, and K. Schulten, J. Comput. Chem. **26**, 1781 (2005).
[7] J. M. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, J. Comput. Chem. **25**, 1157 (2004).
[8] M. Shirts and V. S. Pande, Science **290**, 1903 (2000).
[9] J. Mervis, Science **293**, 1235 (2001).
[10] D. A. Reed, Computer **36**, 62 (2003).
[11] J.-W. Chu and G. A. Voth, Proc. Natl. Acad. Sci. U.S.A. **102**, 13111 (2005).
[12] P. L. Freddolino, A. S. Arkhipov, S. B. Larson, A. McPherson, and K. Schulten, Structure (London) **14**, 437 (2006).
[13] K. Y. Sanbonmatsu, S. Joseph, and C. S. Tung, Proc. Natl. Acad. Sci. U.S.A. **102**, 15854 (2005).
[14] H. L. Scott, Curr. Opin. Struct. Biol. **12**, 495 (2002).
[15] P. D. Blood and G. A. Voth, Proc. Natl. Acad. Sci. U.S.A. **103**, 15068 (2006).
[16] B. J. Peter, H. M. Kent, I. G. Mills, Y. Vallis, P. J. G. Butler, P. R. Evans, and H. T. McMahon, Science **303**, 495 (2004).
[17] C. D. Snow, N. Nguyen, V. S. Pande, and M. Gruebele, Nature (London) **420**, 102 (2002).
[18] J. N. Onuchic and P. G. Wolynes, Curr. Opin. Struct. Biol. **14**, 70 (2004).
[19] K. Simons and D. Toomre, Nat. Rev. Mol. Cell Biol. **1**, 31 (2000).
[20] T. Pawson and P. Nash, Science **300**, 445 (2003).
[21] B. K. Ganser, S. Li, V. Y. Klishko, J. T. Finch, and W. I. Sundquist, Science **283**, 80 (1999).
[22] V. Tozzini, Curr. Opin. Struct. Biol. **15**, 144 (2005).
[23] G. S. Ayton, W. G. Noid, and G. A. Voth, Curr. Opin. Struct. Biol. **17**, 192 (2007).
[24] S. J. Marrink, A. H. de Vries, and A. E. Mark, J. Phys. Chem. B **108**, 750 (2004).
[25] M. Levitt, J. Mol. Biol. **104**, 59 (1976).
[26] S. Miyazawa and R. L. Jernigan, Macromolecules **18**, 534 (1985).
[27] A. P. Lyubartsev and A. Laaksonen, Phys. Rev. E **52**, 3730 (1995).
[28] S. Izvekov and G. A. Voth, J. Phys. Chem. B **109**, 2469 (2005).
[29] S. Izvekov and G. A. Voth, J. Chem. Phys. **123**, 134105 (2005).
[30] W. G. Noid, J.-W. Chu, G. S. Ayton, and G. A. Voth, J. Phys. Chem. B **111**, 4116 (2007).
[31] J.-W. Chu, G. S. Ayton, S. Izvekov, and G. A. Voth, Mol. Phys. **105**, 167 (2007).
[32] F. Muller-Plathe, ChemPhysChem **3**, 754 (2002).
[33] A. Liwo, S. Oldziej, C. Czaplewski, U. Kozlowska, and H. A. Scheraga, J. Phys. Chem. B **108**, 9421 (2004).
[34] A. Liwo, C. Czaplewski, J. Pillardy, and H. A. Scheraga, J. Chem. Phys. **115**, 2323 (2001).
[35] I. G. Kevrekidis, C. W. Gear, and G. Hummer, AIChE J. **50**, 1346 (2004).
[36] N. V. Buchete, J. E. Straub, and D. Thirumalai, Protein Sci. **13**, 862 (2004).
[37] E. Villa, A. Balaeff, L. Mahadevan, and K. Schulten, Multiscale Model. Simul. **2**, 527 (2004).
[38] A. P. Lyubartsev, Eur. Biophys. J. **35**, 53 (2005).
[39] S. Takada, Proteins: Struct., Funct., Genet. **42**, 85 (2001).
[40] J.-W. Chu, S. Izvekov, and G. A. Voth, Mol. Simul. **32**, 211 (2006).
[41] J. C. Shelley, M. Y. Shelley, R. C. Reeder, S. Bandyopadhyay, and M. L. Klein, J. Phys. Chem. B **105**, 4464 (2001).
[42] J. C. Shelley, M. Y. Shelley, R. C. Reeder, S. Bandyopadhyay, P. B. Moore, and M. L. Klein, J. Phys. Chem. B **105**, 9785 (2001).
[43] S. O. Nielsen, C. F. Lopez, I. Ivanov, P. B. Moore, J. C. Shelley, and M. L. Klein, Biophys. J. **87**, 2107 (2004).
[44] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries, J. Phys. Chem. B **111**, 7812 (2007).
[45] Y. T. Wang, S. Izvekov, T. Y. Yan, and G. A. Voth, J. Phys. Chem. B **110**, 3564 (2006).
[46] S. Iuchi, S. Izvekov, and G. A. Voth, J. Chem. Phys. **126**, 124505 (2007).
[47] S. Izvekov and G. A. Voth, J. Chem. Phys. **125**, 151101 (2006).
[48] S. Izvekov and G. A. Voth, J. Chem. Theory Comput. **2**, 637 (2006).
[49] J. Zhou, I. F. Thorpe, S. Izvekov, and G. A. Voth, Biophys. J. **92**, 4289 (2007).
[50] S. Izvekov, A. Violi, and G. A. Voth, J. Phys. Chem. B **109**, 17019 (2005).
[51] Q. Shi, S. Izvekov, and G. A. Voth, J. Phys. Chem. B **110**, 15045 (2006).
[52] B. M. Forrest and U. W. Suter, J. Chem. Phys. **102**, 7256 (1995).
[53] G. Ciccotti and J. P. Ryckaert, Comput. Phys. Rep. **4**, 345 (1986).
[54] M. Fixman, Proc. Natl. Acad. Sci. U.S.A. **71**, 3050 (1974).
[55] H. Goldstein, C. Poole, and J. Safko, *Classical Mechanics* (Addison-Wesley, Reading, MA, 2002).
[56] E. Helfand, J. Chem. Phys. **71**, 5000 (1979).
[57] N. Go and H. A. Scheraga, Macromolecules **9**, 535 (1976).
[58] D. Chandler and B. J. Berne, J. Chem. Phys. **71**, 5386 (1979).
[59] M. Sprik and G. Ciccotti, J. Chem. Phys. **109**, 7737 (1998).
[60] J. P. Ryckaert and G. Ciccotti, J. Chem. Phys. **78**, 7368 (1983).
[61] W. G. Noid, P. Liu, Y. Wang, J.-W. Chu, G. S. Ayton, S. Izvekov, H. C. Andersen, and G. A. Voth, J. Chem. Phys. **128**, 244115 (2008).
[62] S. Izvekov, M. Parrinello, C. J. Burnham, and G. A. Voth, J. Chem. Phys. **120**, 10896 (2004).
[63] P. Liu and G. A. Voth, J. Chem. Phys. **126**, 045106 (2007).
[64] A. A. Louis, J. Phys.: Condens. Matter **14**, 9187 (2002).
[65] G. Brannigan, L. C.-L. Lin, and F. L. H. Brown, Eur. Biophys. J. **35**, 104 (2006).
[66] M. Praprotnik, L. Delle Site, and K. Kremer, J. Chem. Phys. **126**, 134902 (2007).
[67] J.-P. Hansen and I. R. McDonald, *Theory of Simple Liquids*, 2nd ed. (Academic, New York, 1990).