Order parameters for macromolecules: Application to multiscale simulation

A. Singharoy, S. Cheluvaraja, and P. Ortoleva[a]
Center for Cell and Virus Theory, Indiana University, Bloomington, Indiana 47405, USA

(Received 10 May 2010; accepted 13 November 2010; published online 24 January 2011)

Order parameters (OPs) characterizing the nanoscale features of macromolecules are presented. They are generated in a general fashion so that they do not need to be redesigned with each new application. They evolve on time scales much longer than 10^{-14} s typical for individual atomic collisions/vibrations. The list of OPs can be automatically increased, and completeness can be determined via a correlation analysis. They serve as the basis of a multiscale analysis that starts with the N-atom Liouville equation and yields rigorous Smoluchowski/Langevin equations of stochastic OP dynamics. Such OPs and the multiscale analysis imply computational algorithms that we demonstrate in an application to ribonucleic acid structural dynamics for 50 ns. © *2011 American Institute of Physics*. [doi:10.1063/1.3524532]

I. INTRODUCTION

A multiscale framework for simulating the dynamics of macromolecules is developed. Their dynamics is divided into high frequency atomic vibrations and slow (coherent) large-spatial-scale conformational changes. A set of order parameters (OPs) is introduced to describe the coherent, overall structural changes while the small amplitude and high frequency atomic fluctuations are described by an equilibrium distribution following from entropy maximization constrained to instantaneous values of the OPs. In effect, OPs as conceived here filter out the high frequency atomistic fluctuations. These concepts are put on a firm mathematical basis via a multiscale analysis of the Liouville equation. The result of the latter analysis is a set of Langevin equations, all factors within which are related to an interatomic force field. This yields a force-field based multiscale algorithm that allows for all-atom simulation of macromolecular structural transitions with high CPU efficiency. We believe this computational approach will be of great value in fundamental and applied studies of the dynamics of macromolecules and their interactions with their microenvironment.

OPs characterize the state of organization of a system. As used here they describe the overall structure of a macromolecule. The objective of introducing OPs has been dimensionality reduction, i.e., to arrive at a practical computational algorithm for large systems and to understand the salient features of the structure/dynamics of nanoscale assemblies.[1] They are related to the underlying all-atom description, enabling a unified treatment based on the Newtonian Liouville equation.[2,3] The present objective is to analyze a set of OPs, showing that they facilitate a complete analysis of macromolecular dynamics.

Given that the OPs evolve much slower than the 10^{-14} s time scale of atomistic collisions/vibrations, the latter have sufficient time to arrive at a quasiequilibrium consistent with the instantaneous state of the OPs. The OPs modify the probability distribution of atomistic configurations, which, in turn, determines the diffusions and thermal-average forces mediating OP dynamics. In this view, macromolecular structural dynamics follows from the coupling of processes across multiple scales in space and time.[2–11] The result of the multiscale analysis of the Newtonian Liouville equation is a set of Langevin equations of stochastic OP dynamics. If the set of OPs is incomplete (i.e., their dynamics is coupled to that of other slowly changing variables), then they satisfy equations involving time delays (i.e., memory effects) as resulting from traditional projection operator analysis.[12] In contrast, the formal multiscale analysis presented here does not involve these memory effects because of the time scale separation enabled by the OPs.

A practical property of macromolecular OPs is that their construction be automatable. This has two implications: (1) the description can readily be enriched if it is found to be incomplete and (2) the tedious process of inventing new OPs for each macromolecule is avoided. The OPs must capture key features of the free-energy landscape in order to be complete dynamically. In this way, they capture key pathways for structural transitions and associated energy barriers. Earlier choices for OP-like variables include principal component analysis (PCA) modes to identify collective behaviors in macromolecular systems,[13–15] dihedral angles,[16,17] curvilinear coordinates to characterize macromolecular folding and coiling,[18] bead models wherein a peptide or nucleotide is represented by a bead which interacts with others via a phenomenological force, and spatial coarse-grained models.[19–21] These approaches suffer from one or more of the following difficulties: (1) characteristic variables are not slowly varying in time; (2) macromolecular twist is not readily accounted for; (3) their internal dynamics, and hence inelasticity of their collisions is neglected; and (4) the forces involved must be calibrated for most new applications. In contrast as discussed in Secs. II and III, these difficulties are overcome in the present approach.

However, PCA has been successfully used to study protein folding. The PCA modes (involving collective degrees of freedom) probe the global motions of proteins in an "essential" subspace.[22,23] Dihedral PCA (dPCA) has been

[a] Author to whom correspondence should be addressed. Electronic mail: ortoleva@indiana.edu.

suggested to naturally provide a correct separation of internal and overall motions.[16] Free-energy landscapes of several small molecules including protein and ribonucleic acid (RNA) hairpin have been analyzed using dPCA. However, it has been commented that dPCA produces distortions on the free-energy surface that can lead to artificial energy barriers and minima.[24] One of the simulation methods based on PCA is essential dynamics sampling (EDS). Here the PCA modes are used to guide the macromolecular dynamics. This method has been successfully used to study the folding path of cytochrome *c* over hundreds of ps. But the principal components vary significantly during large transformations and hence need to be carefully changed from those of the starting structure.[13,22] Another problem with EDS is the prerequisite of PCA modes that require up to several nanoseconds of molecular dynamics (MD) run. For big systems, this becomes computationally expensive.[21] Improvements to EDS have been made via a technique called amplified collective motion. However normal modes derived from this model via an anisotropic network[13] may not be consistent with all-atom models, especially if the structure undergoes severe deformation. Direct essential dynamics (DED) on the other hand uses the most active collective mode and a weak force to jump out of energy minima as discussed and reviewed in Ref. 13. This has been demonstrated on a 15 amino acid S-peptide where DED trajectories overcame local minima and energy barriers and folded the protein faster than conventional MD. Stock *et al.* developed a multidimensional Langevin model of biomolecules where dPCA modes are used to determine slow, large amplitude motions.[16] This has been used to model tri- and hepta-alanine structural transitions. The methodology stresses on the large dimensionality of the model essential for timescale separation.[17] Similar ideas arise from our multiscale analysis developed in Sec. III and other papers.[25,26] However, to accurately generate the stochastic driving field for the Langevin equation, long time MD data is required to calculate the PCA modes. For example, a 100 ns trajectory is required as an input for subsequent analysis of tri-alanine. "Scrambled" data from replica MD is suggested to suffice. These would again make calculations expensive for large systems that exhibit timescale separation. A variational coarse-graining approach was used to locate the coarse-grain (CG) sites on centers of masses of various collections of atoms identified via PCA or normal mode analysis based on the C_α atoms of each residue.[20] However, a simulation method has not yet been implemented using this CG representation. In order to account for considerable far-from equilibrium structures, a nonlinear dimensionality reduction free-energy profiling scheme based on the isometric mapping algorithm isomap has been developed and demonstrated on Src homology three protein using MD simulations.[27] Another coarse-graining approach is the rigid region decomposition model.[28] This has been implemented via algorithms like framework rigidity optimized dynamic algorithm and rigid cluster normal mode analysis to investigate protein mobility.[28]

In this study, we have several objectives to be achieved by introducing OPs: (1) provide a set of OPs for macromolecules that capture the essence of macromolecular structural dynamics (Sec. II), (2) provide an efficient computational algorithm to simulate structural dynamics (Sec. III), and (3) demonstrate the applicability of our OPs to the multiscale algorithm of Sec. III via viral RNA simulations (Sec. IV).

II. MACROMOLECULAR OPs

A key element of the multiscale analysis of a macromolecule is the identification of OPs that describe its nanoscale features. A central property of an OP is that it evolves slowly. Slow OP dynamics emerges in several ways including

- inertia associated with the coherent dynamics of many atoms evolving simultaneously;
- migration over long distances;
- stochastic forces that tend to cancel; and
- species population levels as in chemical kinetics or self-assembly, which involve many units, only a few of which change on the atomic timescale.

OPs considered here relate macromolecular features to a reference structure (e.g., from x-ray crystallography). They are introduced via (1) a transformation warping space[6] and (2) maximizing their information content to relate them to the atomistic configurations.[7] Consider OPs constructed by embedding the system in a volume V_S. Basis functions $U_k(\vec{r})$ for a triplet of labeling indices k are introduced. If computations are carried out using periodic boundary conditions to simulate a large system (e.g., to minimize boundary effects and to handle Coulomb forces), periodic basis functions (Fourier modes) can be used. Other possible basis functions would be spherical harmonics when the system is embedded in a spherical volume. More generally, as is familiar in quantum theory or hydrodynamics, the basis functions used are chosen for convenience to reflect the overall geometry of the system and the conditions imposed at the boundary. Furthermore the basis functions should be free of unphysical features (e.g., poles). In the present study, we found Legendre polynomials to be convenient for simulating systems with closed boundaries of rectangular geometry.[4,6,8,9]

In our approach, points $\vec{r}$ within the system are considered to be a displacement of original points $\vec{r}^{\,0}$.[6] A set of vector OPs $\vec{\Phi}_k$ are constructed as follows. The macromolecule deforms in 3D space such that a point $\vec{r}$ is displaced from an original point $\vec{r}^{\,0}$. Deformation of space taking any $\vec{r}^{\,0}$ to $\vec{r}$ is continuous and is used to introduce OPs $\vec{\Phi}_k$ via

$$\vec{r} = \sum_k U_k(\vec{r}^{\,0})\vec{\Phi}_k. \tag{2.1}$$

As the $\vec{\Phi}_k$ change, space is deformed, and so does the macromolecule embedded in it. The objective is to ensure that the dynamics of the $\vec{\Phi}_k$ reflects the physics of the macromolecule and that the deformation captures key aspects of the atomic-scale details of the structure. In this way, the $\vec{\Phi}_k$ constitutes a set of vector OPs if they are slowly varying in time (see below).

Let the i^{th} atom in the macromolecule ($i = 1, \ldots, N$) be moved from its original position $\vec{r}_i^{\,0}$ via the above deformation by evolving the $\vec{\Phi}_k$ and correcting for atomic-scale details as follows. Given a finite truncation of the k sum in Eq. (2.1), for e.g., choosing N_{OP} number of OPs, there will be some residual displacement (denoted $\vec{\sigma}_i$) for each atom in addition to the coherent deformation generated by the k sum:

$$\vec{r}_i = \sum_k \vec{\Phi}_k U_k\left(\vec{r}_i^{\,0}\right) + \vec{\sigma}_i. \tag{2.2}$$

To maximize the information content of the OPs, the magnitude of the $\vec{\sigma}_i$ can be minimized by the choice of basis functions and the number of terms in the k sum. Conversely, imposing a permissible size threshold for the residuals allows one to determine the number of terms to include in the k sum.

To start the multiscale analysis, the $\vec{\Phi}_k$ must be expressed in terms of the fundamental variables $\vec{r}_i$. To arrive at this relationship, we minimize the mass-weighted square residual $(m_1\sigma_1^2 + \cdots m_N\sigma_N^2)$ with respect to the $\vec{\Phi}_k$, where m_i is the mass of atom i. This implies

$$\begin{aligned} \sum_{k'} B_{kk'}\vec{\Phi}_{k'} &= \sum_{i=1}^{N} m_i U_k\left(\vec{r}_i^{\,0}\right)\vec{r}_i, \\ B_{kk'} &= \sum_{i=1}^{N} m_i U_k\left(\vec{r}_i^{\,0}\right) U_{k'}\left(\vec{r}_i^{\,0}\right). \end{aligned} \tag{2.3}$$

Thus, the OPs can be computed in terms of the atomic positions by solving Eq. (2.3). For convenience, we choose the basis functions U_k to be mass weighted orthogonal. In that case, the B-matrix equation (II.3) is diagonal. We view $U_k(\vec{r}_i^{\,0})$ as the ith component of an N-dimensional vector for an N-atom macromolecule. There are N_{OP} N-dimensional vectors, each labeled by its k value. Orthogonalization of these vectors is carried out using a Gram–Schmidt procedure. In the above notation, k is a set of integers (each of which can be 0,1...); with this, Φ_{100X} is the X component of the OP $\vec{\Phi}_{100}$ that weights the U_{100} contribution in Eq. (2.2) after the latter has been subjected to Gram–Schmidt orthogonalization. In our implementation, before orthogonalization $U_{k_1k_2k_3}$ is a product of Legendre polynomials of order k_1, k_2, k_3 for the X, Y, Z components of $\vec{r}_i^{\,0}$ respectively. The orthogonalization scheme preserves the physical nature of the three fundamental OPs (100X, 010Y, 001Z) because they are always chosen to be the first three members of the basis.[4] For other OPs, the k labeling corresponds to the original $U_k(\vec{r}_i^{\,0})$ from which each orthogonal vector was constructed via the Gram–Schmidt procedure.

Mass-weighted orthonormality of the basis functions implies that $B_{kk'}$ is 0 for $k \neq k'$. With this

$$\vec{\Phi}_k = \frac{\sum_{i=1}^{N} m_i U_k\left(\vec{r}_i^{\,0}\right)\vec{r}_i}{\tilde{\mu}_k}, \quad \tilde{\mu}_k = \sum_{i=1}^{N} m_i\left\{U_k\left(\vec{r}_i^{\,0}\right)\right\}^2. \tag{2.4}$$

Thus for a given set of atomic positions the corresponding OPs are uniquely defined.

Next consider the timescale of OP dynamics. The Liouville operator is defined $\mathcal{L} = -\sum_{i=1}^{N}\vec{p}_i/m_i \cdot \partial/\partial\vec{r}_i + \vec{F}_i \cdot \partial/\partial\vec{p}_i$, where $\vec{p}_i$ and $\vec{F}_i$ are the momentum of, and net force on atom i. Given Eq. (2.4), one may compute $d\Phi/dt$ as $-\mathcal{L}\Phi$, where $\Phi(\Gamma)$ is a set of OPs $\vec{\Phi}_k$. With this

$$\begin{aligned} \frac{d\vec{\Phi}_k}{dt} &= \frac{\vec{\tilde{\Pi}}_k}{\tilde{\mu}_k}, \\ \vec{\tilde{\Pi}}_k &= \sum_{i=1}^{N} U_k\left(\vec{r}_i^{\,0}\right)\vec{p}_i. \end{aligned} \tag{2.5}$$

Inclusion of m_i in developing Eq. (2.3) gives $\vec{\Phi}_k$ the character of a generalized center-of-mass-like (CM) variable. In fact, if U_k is a constant then $\vec{\Phi}_k$ is proportional to the CM. While the $\vec{\Phi}_k$ are given in terms of a sum of N-atomic displacements, many terms of which have similar directions due to the smooth variation of U_k with respect to $\vec{r}_i^{\,0}$, the momenta $\vec{\tilde{\Pi}}_k$ are given by a sum of atomic momenta, which tend to cancel near equilibrium. Hence the thermal average of $\vec{\tilde{\Pi}}_k$ is small and thus the $\vec{\Phi}_k$ tends to evolve slowly.

First consider the dynamics of the CM, i.e., $\vec{\Phi}_{000}$. From Eq. (2.5), $\vec{\Phi}_{000}$ satisfies $d\vec{\Phi}_{000}/dt = \vec{\Pi}_{000}/M$, where $\tilde{\mu}_{000} = M$ is the total mass of the macromolecule. Since M is large, $\vec{\Phi}_{000}$ evolves slowly relative to the time scale of atomic collision/vibration. This suggests that $\vec{\Phi}_{000}$ satisfies a key criterion to be an OP and serves as the starting point of a multiscale analysis.

To reveal the time scale on which the OPs evolve, it is convenient to define the smallness parameter $\varepsilon = m/M$, where m is a typical atomic mass. For any of the $\vec{\Phi}_k$, letting $\vec{v}_i$ be the velocity of particle i, the definition of ε and Eq. (2.5) yields

$$\begin{aligned} \frac{d\vec{\Phi}_k}{dt} &= \frac{\sum_{i=1}^{N} U_k\left(\vec{r}_i^{\,0}\right)\vec{p}_i}{\tilde{\mu}_k} = \frac{\sum_{i=1}^{N} U_k\left(\vec{r}_i^{\,0}\right)m_i\vec{v}_i}{\tilde{\mu}_k} \\ &= \frac{\sum_{i=1}^{N} U_k\left(\vec{r}_i^{\,0}\right)m\hat{m}_i\vec{v}_i}{M\mu_k} = \varepsilon\frac{\sum_{i=1}^{N} U_k\left(\vec{r}_i^{\,0}\right)\hat{m}_i\vec{v}_i}{\mu_k} = \varepsilon\frac{\vec{\Pi}_k}{\mu_k}, \end{aligned} \tag{2.6}$$

where $\mu_k = \tilde{\mu}_k/M$ and $\hat{m}_i = m_i/m$.

Thus $\vec{\Phi}_k$ changes at a rate O(ε) under the assumption that the atomic momenta tend to cancel as is consistent with the quasiequilibrium probability distribution $\hat{\rho}$ below. Special initial conditions could make the rate of OP change scale differently; examples of such conditions include an initial density discontinuity (leading to a shockwave), injection of the macromolecule at a high velocity or a sudden jump in temperature. Under these conditions, the slowness of motion within our reduced dimensionality framework (OPs) and all the resulting advantages (e.g., calculating thermal-average forces) is lost. Therefore, for any class of initial conditions, the slow rate of OP dynamics must be confirmed before applying the multiscale ideas developed in Sec. III. In this study we demonstrate the applicability of the $\varepsilon-$scaling for typical conditions underlying macromolecular behavior (Sec. IV).

A simple case of the $\vec{r}_i, \vec{\Phi}_k$ relationship suggests how it captures rigid rotation. Take $U_k, k = 100, 010, 001$ to be x^0, y^0, and z^0, respectively. Then neglecting the residuals, Eq. (2.2) becomes $x_i = \Phi_{100x} x_i^0 + \Phi_{010x} y_i^0 + \Phi_{001x} z_i^0$, and similar for y_i and z_i (where x_i, y_i, z_i are the three Cartesian components of $\vec{r}_i$ vector). The relationship can be written in the tensorial form $\vec{r}_i = \vec{\vec{\Phi}}\,\vec{r}^{\,0}_i$. It is seen that for a special case (i.e., where the tensor $\vec{\vec{\Phi}}$ is a rotation matrix), $\vec{\Phi}_k$ constitute a length preserving rotation about the CM if $\vec{r}_i$ is relative to the CM. More generally, for the above three basis functions, the $\vec{r}_i - \vec{\Phi}_k$ relationship corresponds to a mixed rotation, extension/compression. In fact the OPs defined here constitute a strain tensor thereby accounting for elastic deformations. In addition, our multiscale formulation (Sec. III) is all-atom and hence captures internal friction effects via the force field. Therefore, the theory accounts for both elastic and viscous effects. As discussed in the original paper, where the present OPs (denoted "global co-ordinates" there) were introduced,[6] the higher order OPs specifically second and third order capture twisting, bending, and more complex deformations. Such OPs were shown to capture polyalanine folding from a linear to a globular state. In another work, the OPs were shown to capture nucleation and front propagation in a virus capsid.[8] While it is not trivial to interpret all the deformations associated with the higher order polynomial-defined OPs, it is the generality of our multiscale approach (Sec. III) that accounts for all their dynamics. Ultimately the interpretation of the OPs is embodied in the description of the phenomenon itself, e.g., macromolecular structural transition. A commonplace example is hydrodynamics wherein one does not always interpret the meaning of each Fourier density mode, but rather speaks in terms of phenomena such as propagating waves or viscous boundary layers. Additional properties of the OPs are discussed in Appendix A.

The set Φ of OPs have technical advantages that greatly facilitate theoretical analyses. Consider an extended set Φ_{ex} of OP and OP-like variables, notably the $\vec{\Phi}_k$ for k in the list of OPs plus similarly defined variables $\vec{\Phi}_{k\text{res}}$ for k not in the OP list. Thus we write

$$\vec{r}_i = \sum_k^{OP} \vec{\Phi}_k U_k\left(\vec{r}_i^{\,0}\right) + \sum_k^{res} \vec{\Phi}_{kres} U_k\left(\vec{r}_i^{\,0}\right). \qquad (2.7)$$

This relation maps the $3N$ configuration variables r onto Φ_{ex}, also a $3N$-dimensional space. Expression (2.7) for $\vec{r}_i$ in terms of Φ and Φ_{res} provides a way to generate ensembles of $\vec{\Phi}_k$-constrained configurations by randomly varying the $\vec{\Phi}_{k\text{res}}$. An expression for $\vec{\sigma}_i$ in terms of Φ_{res} is obtained by comparing Eqs. (2.2) and (2.7). However, generating ensembles by randomly varying the $\vec{\sigma}_i$ typically leads to very high-energy configurations. This difficulty is readily avoided as long as $\vec{\sigma}_i$ is chosen by constraining $\vec{\Phi}_{k\text{res}}$ for higher order k to small values.[4] The lower $k - \vec{\Phi}_{k\text{res}}$ do provide major structural variations by moving atoms in the ensemble with a measure of coherence, avoiding near-atom overlap. Thus $\vec{\Phi}_{k\text{res}}$ provides a way to generate rich ensembles at fixed Φ and with modest energies (and hence Boltzmann relevance). In practice, a "hybrid" sampling method, wherein short MD runs are performed starting with configurations from the $\vec{\Phi}_{k\text{res}}$-generated sample is used to enrich fluctuations about the constant set of OPs Φ.[4] All these properties are critical for the practical implementation of a multiscale molecular dynamics/order parameter extrapolation (MD/OPX) approach[7] and more recently a fully self-consistent multiscale approach and software **SIMNANOWORLD**$^{\text{TM}}$.[4] Implementation of the former is based on the philosophy of equation-free multiscale approach.[29] In this, the absence of macroscopic equations of motion is overcome by extrapolating OPs over large time intervals using short bursts of MD simulation. In contrast, **SIMNANOWORLD** computations are guided by the Langevin equation for OPs (III.9). Its development is closer to the theme of the heterogeneous multiscale modeling,[30] in the sense, maximum knowledge (or ignorance) on the various scales in the system is utilized in deriving the quasi-equilibium probability density.

Note, even though the OPs are defined in terms of atoms in the macromolecule, multiscale analysis developed in Sec. III accounts for both the macromolecule and the medium, allowing simulation of the entire system. Since the OPs evolve on a long timescale, their dynamics filters out the high frequency atomistic fluctuations (residuals). Thus the slowly evolving OPs can be projected over large intervals in time. These timesteps are appreciably larger than simple MD timesteps and, therefore, efficiently probe the long time behavior of a macromolecule. As the above OPs are generated in an automated fashion, the set may be expanded by increasing the range of the k sum. As discussed in Sec. III, this addresses the difficulty that arises when a limited set of OPs couples to other slow variables.

III. MULTISCALE THEORY AND COMPUTATION

In this section, we use the OPs considered above and the Liouville equation to derive equations for the stochastic dynamics of a macromolecule. The analysis starts by writing the Liouville equation for the N-atom probability density Υ, i.e., $\partial\Upsilon/\partial t = \mathcal{L}\Upsilon$ for Liouville operator $\mathcal{L}$. Υ depends on the set of $6N$ positions and momenta Γ and time t.

Multiscale analysis starts with a transformation of the N-atom probability density from the $\Upsilon\,(\Gamma, t)$ formulation to one that makes the multiple ways on which Υ depends on Γ, t more explicit. This involves introduction of a set of OPs $\Phi(\Gamma)$ (i.e., $\vec{\Phi}_k$ of Sec. II for all k on the list of OPs) that depends on Γ and which are shown to evolve on a time scale much greater than that of individual atomic collisions/vibrations.

First we write Υ in a form that makes the dependence on Γ and t of various types explicit:

$$\Upsilon\,(\Gamma, t) = \rho\,\{\Gamma_0(\Gamma), \Phi(\Upsilon), t_0(t), \underline{t}(t); \varepsilon\}\,. \qquad (3.1)$$

Thus we make an *ansatz* that reformulated probability density ρ depends on the N-atom state Γ both directly (i.e., via $\Gamma_0(\Gamma) = \Gamma$) and, via a set of OPs $\Phi(\Gamma)$, indirectly. Similarly,ρ depends on the sequence of times $t_0(t), t_1(t), t_2(t), \ldots = t_0(t), \underline{t}(t)$, where $t_n(t) = \varepsilon^n t$. The times t_n for $n > 0$ are introduced to account for the slower behaviors in ρ; while t_0

accounts for processes on the fast timescale (i.e., t_0 changes by one unit when 10^{-14} s elapse). ε is a small parameter as defined in Sec. II. The ε dependence of ρ and scaling of time are justified later in this section.

In adopting this perspective, Φ is not a set of additional independent dynamical variables; rather, its appearance in ρ is a place holder for a special dependence of ρ on Γ that underlies the slow temporal dynamics of ρ. A simple example that elucidates our *ansatz* is the function $f(x) = \exp^{-\varepsilon x} \sin(x)$. We restate $f(x)$ as $f(x_0, x_1)$ where $x_0 = x$ and $x_1 = \varepsilon x$. In making this transformation, we do not add any independent variable to the description, rather we make the discrete dependencies on x explicit. It is shown below that the dual dependence of ρ on Γ can be constructed if ε is sufficiently small. An equation of stochastic OP dynamics that preserves the feedback between the atomistic and nanoscale variables[4] is now obtained via a multiscale perturbation analysis for a classical N-atom system.

In the following, we use the above framework to derive an equation for the OP probability distribution. One finds that $\mathcal{L}\Phi$ naturally reveals a small parameter ε, i.e., $d\vec{\Phi}_k/dt = \varepsilon(\vec{\Pi}_k/\mu_k)$ (Sec. II). Starting with Eq. (3.1), the Liouville equation for Υ and the chain rule, one obtains the multiscale Liouville equation (Appendix B),

$$\sum_{n=0}^{\infty} \varepsilon^n \frac{\partial \rho}{\partial t_n} = (\mathcal{L}_0 + \varepsilon \mathcal{L}_1)\,\rho. \tag{3.2}$$

Many authors (see Refs. 9, 31, and 32 for reviews) have analyzed such equations in the small ε limit. Equation (III.2) is solved perturbatively via a Taylor expansion in ε. As shown in Appendix B and,[25] $\mathcal{L}_0$ involves partial derivatives with respect to Γ_0 at constant Φ [when operating on ρ in the multiscale form (3.1)], and conversely for $\mathcal{L}_1$. With this $\mathcal{L}_0$ *and* $\mathcal{L}_1$ take the forms

$$\mathcal{L}_0 = -\sum_{i=1}^{N} \frac{\vec{p}_i}{m_i} \cdot \frac{\partial}{\partial \vec{r}_i} + \vec{F}_i \cdot \frac{\partial}{\partial \vec{p}_i} \tag{3.3}$$

$$\mathcal{L}_1 = -\sum_k \frac{\Pi}{\mu_k} \cdot \frac{\partial}{\partial \Phi}. \tag{3.4}$$

Note that $\mathcal{L}_0$ and $\mathcal{L}_1$ operate in the space of functions that depend explicitly on variables Γ_0 and Φ; Π signifies a set of $\vec{\Pi}_k$ and subscripts 0 on $\vec{r}_i$ and $\vec{p}_i$ in Eq. (3.3) are henceforth dropped because of the simple $\Gamma_0(\Gamma) = \Gamma$ dependence of ρ. While the space of functions on which $\mathcal{L}_0$ and $\mathcal{L}_1$ operates is composed of $6N + N_{\mathrm{OP}}$ variables (the $6N$ atomic positions and momenta Γ_0 plus the N_{OP} OPs Φ), the formalism does not assume that the variables are dynamically independent. Rather, from Eq. (3.2), one determines the dependence of ρ on Γ_0 and Φ, but ultimately through Eq. (3.1) how Υ depends on Γ. Hence Eqs. (3.2), (3.3), and (3.4) do not imply that Γ_0 and Φ are independent dynamical variables but, in accordance with Eq. (3.1) the equations track the multiple space and time dependencies of Υ. Therefore, there are still $6N$ dynamical variables as the OPs do not evolve independently of the $6N$ atomic positions and momenta. Equations (2.4) and (2.6) show the explicit dependencies of atomic and coarse-grained quantities. In contrast, one could introduce collective modes as new dynamical variables in addition to the $6N$ atomic positions and momenta Γ. However, this approach carries the burden of eliminating selective position and momentum variables to keep the number ($6N$) of degrees of freedom fixed. In summary, to uncloak the explicit space-time dependencies of the N-particle density Υ, we make use of $6N + N_{\mathrm{OP}}$ variables of which N_{OP} are not independent of the remainder [with dependencies defined via Eqs. (2.4) and (2.6)]. As no additional independent variables are added to the description of the N-atom system, Υ still remains a function of the $6N$ dynamical variables. Furthermore, the $O(\varepsilon)$ scaling of the Liouville equation is a natural consequence of the slowness of OPs. This justifies a perturbative solution and hence the ε dependence of the N-atom probability density. With this, the N-atom probability density is constructed as $\rho = \sum_{n=0}^{\infty} \varepsilon^n \rho_n$. Putting the perturbation expansion for ρ into Eq. (3.2) and analyzing different orders in ε (discussed in details in Appendices SI, SII, and SIII (Ref. 33), the Smoluchowski equation for the coarse-grained probability distribution $\tilde{W}$ is obtained,

$$\frac{\partial \tilde{W}}{\partial \tau} = \sum_{kk'} \frac{\partial}{\partial \vec{\Phi}_k} \left[\vec{\vec{D}}_{kk'} \left[\frac{\partial}{\partial \vec{\Phi}_{k'}} - \beta \vec{f}_{k'} \right] \tilde{W} \right]. \tag{3.5}$$

The diffusivity factors $\vec{\vec{D}}_{kk}$ are related to the correlation function of time derivatives of OPs via

$$\vec{\vec{D}} = \frac{1}{\mu_k \mu_{k'}} \int_{-\infty}^{0} dt_0' \left\langle \vec{\Pi}_k e^{-\mathcal{L}_0 t_{0'}} \vec{\Pi}_{k'} \right\rangle, \tag{3.6}$$

where $\vec{\Pi}_k$ is defined in terms of the OP time derivatives via Eqs. (2.5) and (2.6). In constructing the correlation functions the initial data is at fixed Φ; since Φ does not change appreciably during the period in which the correlation function is appreciable, $\vec{\vec{D}}_{kk'}$ depends on Φ.

The thermal-average force $\vec{f}_k$ is given by

$$\vec{f}_k = -\frac{\partial F}{\partial \vec{\Phi}_k} = \langle \vec{f}_k^{\,\mathrm{m}*} \rangle, \tag{3.7}$$

for Φ constrained Helmholtz free-energy F, where

$$F = -\frac{1}{\beta} \ln Q(\Phi, \beta), \tag{3.8}$$

$Q(\Phi, \beta)$ is the partition function associated with $\hat{\rho}$ [Appendices SI, SII, and SIII (Ref. 33)], and $\vec{f}_k^{\,\mathrm{m}*} = \sum_{i=1}^{N} U_k(\vec{r}_i^{\,0}) \vec{F}_i^{\,*}$ is the "OP force." Details of Eq. (3.7) derivation are provided in Appendix C.

Equivalent to Eq. (3.5) is an ensemble of OP time courses generated by the Langevin equations

$$\frac{\partial \vec{\Phi}_k}{\partial \tau} = \beta \sum_{k'} [\vec{\vec{D}}_{kk'} \vec{f}_{k'}] + \vec{\xi}_k. \tag{3.9}$$

The coherent part of the evolution is determined by the product of the diffusion factors and the thermal-average forces; the stochastic evolution is determined by the random force $\vec{\xi}_k$. The latter is constrained by requiring the integral of its autocorrelation function to be proportional to the diffusion coefficient.

The expression for diffusion factors provided above involves an integration of the correlation function over all time. However, if the correlation function decays on a long time scale (i.e., on that comparable to OP evolution), the above Smoluchowski equation would be replaced by one that is nonlocal in time. This would suggest that the set of OPs couples to other slow variables. Since the OPs are generated automatically (as described in Sec. II), new slow variables can be added in a straightforward way to make the existing set Φ complete, e.g., eliminate the long-time tail. Completing the set of OPs modifies the operator $\mathcal{L}_0$ [and hence the velocity correlation of Eq. (3.6)] as the latter involves derivatives with respect to Γ_0 at constant Φ. This modifies the diffusion factor, affecting dynamics of the OPs on the free-energy surface they define. Such an operator is automatically accounted for via standard MD codes when the correlation time of OP velocities is short relative to the timescale of OP evolution. Thus the long-time behavior of correlation functions provides a completeness criterion for the set of OPs and, thereby, a self-consistency check for the theory and computations.

Another self-consistency check is related to refreshing of the reference structure r^0. Our simulations begin with the energy-minimized and thermally equilibrated x-ray crystallographic or other all-atom structure as the reference structure. As the system evolves in time, the resulting deformation may increase some of the residuals. This may reflect the need for a new reference structure. The reference structure transition point is chosen when the maximum residual for a structure in the constant OP ensemble becomes comparable with its root mean square deviation (RMSD) from the initial reference structure $[|\vec{r}_1 - \vec{r}_1^{\,0}|^2 + \cdots |\vec{r}_N - \vec{r}_N^{\,0}|^2/N]^{1/2}$, i.e., when some local change in a structure reaches the order of an overall deformation [Fig. S1 in Ref. 33]. The increase in residual may indicate the presence of coherent motions that are not accounted for by the set of OPs initially chosen. If the emerging modes couple to the previously defined set of OPs, long-time OP velocity autocorrelation tails are also expected. The remedy for this difficulty is to increase the number of OPs considered, a process greatly facilitated by the automated generation of OPs (Sec. II). Increase in the residuals may indicate the presence of an improbable fluctuation in the MD generated part of the finite ensemble used for a practical computation. This increase is generally minimized via the low $k - \vec{\Phi}_{k\text{res}}$ sampling and, for the thermal-average forces, via the Boltzmann factor (structures with larger values of $\vec{\Phi}_{k\text{res}}$ for higher order k tend to be high in energy as suggested in Fig. S2 in Ref. 33). However, for these cases, a simple reference structure renewal would suffice to account for the resulting motions, and additional OPs are not required.

As stressed above, new OPs should be added in order to redefine the constant OP ensemble (i.e., to eliminate the long-time tails of the correlation functions) and to account for the systematic growth of residuals that is not accounted for via re-referencing. In principle, a simple addition of higher order OPs to the set of existing ones, till the complete elimination of the long-time correlation tail would suffice. However, this might include some unnecessarily high frequency modes as OPs. As a result, the Langevin timestep must be reduced, affecting the efficiency of multiscale simulation. Efficiency can be restored by optimal choice of the OPs to be added. For example, one could carry out a ns conventional MD run and use the resulting configuration time courses to rank the omitted OPs according to their average rate of change. Variables that qualify as OPs are slow and coherent, whereas residuals (that are described by $\vec{\Phi}_{k\text{res}}$) are highly fluctuating with mean close to zero (Sec. IV). The former are added, one at a time, to the list of existing OPs and new ensembles are generated by sampling the remaining variables. This is repeated till the residuals become reasonably small and the long-time tail in the autocorrelation function is eliminated. This procedure allows for addition of only slow variables and hence consideration of high frequency modes as OPs is avoided. Such a procedure is also used in selecting OPs to start the SIMNANOWORLD simulations.

Multiscale analysis provides a numerical simulation approach implied by the feedback between nano and atomic scale variables.[4] As $\vec{f}_k$ and $\vec{\vec{D}}_{kk'}$ are OP-dependent, they must be computed at each Langevin timestep to account for the interscale feedback. A finite Langevin timestep Δt advancement takes the OPs from time t to a time $t + \Delta t$ via Eq. (3.9). Thermal forces $\vec{f}_k$ are efficiently computed via an ensemble/Monte Carlo integration method enabled by the nature of our OPs.[4] Atomic forces obtained from the residual generated OP constrained ensemble (Sec. II) are used to calculate the OP force $\vec{f}_k^{\,\mathrm{m}}$. Monte Carlo integration averaging of $\vec{f}_k^{\,\mathrm{m}}$ over the ensemble is carried out to obtain the thermal ($\hat{\rho}$) average force $\vec{f}_k$. Hence the free-energy driving force is obtained via the all-atom probability density $\hat{\rho}(\Gamma_0, \Phi)$, capturing the cross talk between the OPs and individual atomic degrees of freedom. Since $\hat{\rho}(\Gamma_0, \Phi)$ reflects the OP constrained ensemble, the $6N$ atomic degrees are consistent with the state of the OPs. Note, the definition of $\vec{f}_k$ as OP derivative of free-energy Eq. (3.7), and $\vec{\vec{D}}_{kk'}$ as time integral of OP velocity autocorrelation function Eq. (3.6) is independent of the linearity in the $r - \Phi$ relationship. This implies that the multiscale analysis developed can be applied to any complete set of slow variables provided the $\mathrm{O}(\varepsilon)$ time scaling, and Newton's laws of motion hold for their dynamics. As mentioned in Appendix A, Secs. II and IV, the present OPs form a set of slow variables that has suitable properties to serve as a basis of our approach.

Other sets of slow variables can also be used as long as all criteria of OPs dynamics are satisfied.

Adiabatic schemes have been successfully implemented to perform approximate quantum dynamics simulations[34,35] and Car-Parrinello type *ab initio* quantum dynamics.[36] The latter belongs to a family of extended Lagrangian approaches wherein the time scales of faster and slower degrees of freedom are adjusted to ensure the adiabatic propagation of the former in response to motions in the latter. This adjustment is achieved via attributing fictitious masses and kinetic energy to the faster modes.[36,37] For many atom systems, a free-energy profiling scheme, adiabatic free-energy dynamics (AFED) based on an adiabatic partitioning of the slow and fast variables have been developed.[38] The AFED allows for application of higher temperature for the slow variables facilitating rare events sampling and high mass leading to an adiabatic decoupling.[38] Therefore, the separation of timescale between those of OPs versus atomic dynamics Eq. (2.6), the relatively high OP masses $\tilde{\mu}_k$,[39] and the role of average forces [Eqs. 11, 12, and 13 in Ref. 36, Eq. (3.7) in the present paper and elsewhere[12,40]) for driving the slow variables across the free-energy landscape suggests implementation of an adiabatic dynamics algorithm for thermal-average force calculations. Unlike explicit variable transformation in AFED, extended phase space approaches like temperature accelerated molecular dynamics (TAMD)[41] are also used for simultaneous propagation of slow and fast degrees of freedom. Solving the resulting equations of motion in the extended phase space is equivalent to solving Eq. (3.9). This bypasses explicit averaging. Another related approach, driven adiabatic free-energy dynamics (d-AFED) employs explicit masses and dynamics closer to that of AFED in formulating TAMD.[42] This allows direct generation of multidimensional free-energy surfaces for complex systems from the probability distribution function of the extended phase-space variables. Even though an adiabatic decoupling appears naturally in our analysis by the O(ε) scaling of the Liouville operator, the relative efficiency of an adiabatic relaxation scheme versus the present residual-generated sampling scheme remains to be determined. In particular the present scheme requires the development of a rich ensemble of atomistic configurations at each Langevin timestep, while the adiabatic scheme requires coevolution of the slow and many [O(N)] fast variables. This issue is of critical importance for the efficiency of the simulation of systems involving 10^5 or more atoms. However, in analogy to AFED, using higher temperature for propagating the slower variables (OPs here) would yield an algorithm for the simulation of rare-events phenomenon. This is further discussed in Sec. V.

The diffusion factors are computed via the correlation functions of Eq. (3.6) using short MD runs. The latter is allowed because the correlation times in these functions are much shorter than the characteristic timescale of OP evolution (Sec. IV). All factors in the OP dynamics equation (III.9) are computed from the interatomic force field via Monte Carlo integration and MD. Thus the only element of the calibration in constructing the thermal-average forces and diffusions is through the existing force fields (e.g., CHARMM or AMBER). At each Langevin timestep, the updated OPs are used to generate the atomistic configurations of the macromolecule; then the host medium is introduced via a resolvation module[43] and the entire system is thermalized. An ensemble of such equilibrated atomistic configurations is used to generate the thermal-average forces and diffusions. The latter factors are used to update the OPs completing one cycle of the Langevin timesteping. A simple description of steps involved in the SIMNANOWORLD simulation workflow is included in the supplementary material as Appendix SIV in Ref. 33.

In the present approach the OPs are defined over the macromolecule. However, like for the atomic configurations of the macromolecule, water, and ion are accounted for via the quasiequilibrium ensemble (i.e., the configuration of the water and ions rapidly explores a quasiequilibrium ensemble at each stage of the OP dynamics). This assumption holds only when water/ion equilibrate on a timescale much smaller than that of OPs. Similar resolvation scheme has been used with OPs in simulating virus capsid expansion in Na^+ and Ca^{2+} solutions.[43] Therefore, fluctuations from water and ions modulate the residuals generated within the MD part of the constant OP sampling and hence affect the thermal-average force. This is the rationale behind not including the water and ions in the definition of the OPs. If slow hydrodynamic modes are found to be of interest, these atoms can be included in the definition of the OPs. Ions when tightly bound to the macromolecule are considered a part of OPs. After every Langevin timestep, an ion accessible surface is constructed and ions close to the surface are tracked during the MD ensemble enrichment calculation. Those with appreciable residence time within the surface are included in the definition of the OPs henceforth.

The notion of water and ion dynamics being much faster than a set of slow modes is similar to that used in normal mode Langevin (NML).[44] If these fast modes correspond to the coordinated motion of just a few atoms, then there is no clear separation in their time scale from that of single atom vibrations. Thus such fast modes cannot be described by Langevin dynamics unless memory kernels are used.[12] However, NML identifies the modes by diagonalizing a Hessian matrix and fully relaxes (overdamps) the high frequency modes near their energy minimum (respecting the subspace of low frequency normal modes). In the limit of large damping coefficient, this is equivalent to using Brownian dynamics to propagate high frequency modes. In contrast, for our approach rapidly fluctuating variables are allowed to explore a representative ensemble of configurations. For our purposes, using the overdamped value of the fast variables to construct an all-atom configuration and using that configuration to compute the OP force neglects the difference between the OP forces as a function of the average configuration versus the average of the OP force over an ensemble of atomic configurations. Hence, overdamping of fast modes is avoided. Furthermore, the NML approach would become computationally expensive for the N of O (10^6) atom systems of interest here. In contrast, our approach avoids the need to diagonalize a large Hessian matrix, a feature that follows directly from the explicit formulation of our order parameters via Eqs. (2.2) and (2.4). Finally our OPs are essentially normal modes in that they are linear combinations of atomic positions and furthermore are slowly

TABLE I. Input parameters for the NAMD and SIMNANOWORLD simulations.

Parameter	Values
Temperature	300 K
Langevin damping	5
Timestep	1 fs
Full-Elect Frequency	2 fs
Non-bonded Frequency	1 fs
Box size	145 Å × 145 Å × 145 Å[a] or 162 Å × 162 Å × 162 Å[b]
Force-field parameter	par_all27_prot_na.prm
1—4 scaling	1.0
Switch dist	10.0 Å
Cutoff	12.0 Å
Pair list dist	20.0 Å
Steps per cycle	2
Rigid bond	Water

[a] Box for free RNA simulation for 3 ns.

[b] Box for free RNA simulation from 3–50 ns and protein-bound RNA.

varying. However, our equations of motion are highly nonlinear in the OPs and are dissipative. Thus, unlike for normal mode analysis, our formulation can account for states that are far from a free-energy minimizing structure.

For some choice of initial data, the O (ε) contribution to $\tilde{W}(\Phi, t)$ can have short timescale dependence (e.g., due to a shock wave). Under this condition our basic assumption of the lowest order quasiequilibrium behavior is violated as the $O(\varepsilon)$ scaling of the OP motion is disturbed (Sec. II). Thus the theory breaks down. In such a case one expects Fokker–Planck behavior. The present formulation can be generalized to accommodate such inertial effects.[45] However, for the macromolecular phenomena considered here this class of initial data is ignored.

IV. NUMERICAL SIMULATIONS

A study was undertaken to assess the viability of the OPs of Sec. II as a basis of a multiscale algorithm for simulating macromolecules. An evaluation of other variables in this context is also obtained. Comparisons with traditional MD were made to determine the accuracy and efficiency of the method. All multiscale simulations were done using the **SIMNANOWORLD** software[4,5] that is based on the OPs of Sec. II and the multiscale analysis of Sec. III. The CHARMM22/27 force-field and NAMD are incorporated in **SIMNANOWORLD** as part of the computation of the thermal-average forces and diffusion factors for our OPs.

The demonstration system was the RNA of *Satellite Tobacco Mosaic Virus* (STMV).[46] This molecule contains 949 nucleotides. The initial state of the system was that believed to be at equilibrium when the RNA resided with the associated proteins within the STMV capsid. Evolution followed after the capsid was removed instantaneously. The host medium was 0.3M NaCl solution and the temperature was 300 K. The system was placed within a cube and NVT conditions were applied. Details of the NAMD settings are given in Table I. **SIMNANOWORLD** simulations were done in these conditions (and not those used in other studies[46]) as more dramatic structural changes occur because the RNA is more stabilized by Mg^{2+} than by Na^+.[47] This is because Na^+ is expected to be diffusively bound to RNA[47] where as Mg^{2+} remains tightly bound. The structure used to initialize the simulation was generously provided by Prof. K. Schulten; the same structure was used in Ref. 46.

As mentioned in Sec. II, in the following we examine the slowness in rate of change of a typical OP (001Z) to ensure its applicability within our multiscale framework. The one considered in particular exhibits properties of dilation/extension about the *Z*-axis. The time evolution of this OP is compared to other variables only to validate that some of the latter are not suitable as slow variables for the purpose of our multiscale analysis. If the fluctuations in these variables probe short space-time events then they are expected to be accounted for by the quasiequilibrium ensemble, otherwise larger space-time events are accounted for by one or more of our OPs. Appropriateness of these other variables for specific problems have been discussed in the literature cited below and is only included here to distinguish them from our OPs.

Consider some other variables commonly used to characterize macromolecules, denoted "structural parameters"(SPs) here and compare their dynamics with that of the OPs of Sec. II. The SPs analyzed in this paper are different types of dihedrals and their cosines,[48,49] radius of gyration, end-to-end distances,[49] and typical components of the unit vectors along the bonds connecting monomers.[50] We designate these SPs: SP_1, SP_2, $SP_{3,}$ and $SP_{4,}$ respectively. Each of them was calculated over a 1 ns MD trajectory for the RNA under conditions mentioned above. Their time evolutions are plotted in Fig. S3 in Ref. 33. We calculated the moving averages over a window of 50 ps to filter out the coherent part of the variations from the fluctuations (λ). Details on λ computation are provided in Appendix D. The coherent and fluctuating parts of SP variations are plotted in Fig. S4 in Ref. 33. The dihedrals γ, δ, and ε defined according to Fig. 1(a) in Ref. 17 were chosen for this comparison. The time evolution of these depends on their location in the back bone. δ fluctuates the least due to geometric constraints imposed by a five membered ring.[17] In contrast, those not associated with the backbone fluctuate even more than γ or ε. Fluctuations of variables characterizing the overall size of the macromolecule, like SP_2 or SP_3, are much smaller. For $SP_{4,}$ fluctuations are maximum and are also sensitive to bond location.

We compared the time evolution of these SPs to that of the OPs from Sec. II. In Fig. S5 in Ref. 33 the coherent and fluctuating parts are shown for the same 50 ps window. Fluctuations in SP_1 and SP_4 are several orders of magnitude higher in amplitude than those for OPs, and/or their characteristic time scale is much shorter. Hence they do not evolve slowly and cannot serve as a basis of our multiscale analysis. Finally SP_2 and SP_3 are suitable as OPs but do not readily enable the generation of SP ensembles as they are a subset of a more general set of OPs.[6] This presents difficulties in computing thermal-average forces and diffusion factors needed for a multiscale analysis. In contrast, our OPs are automatically generated and, therefore, the set Φ can be augmented for sys-

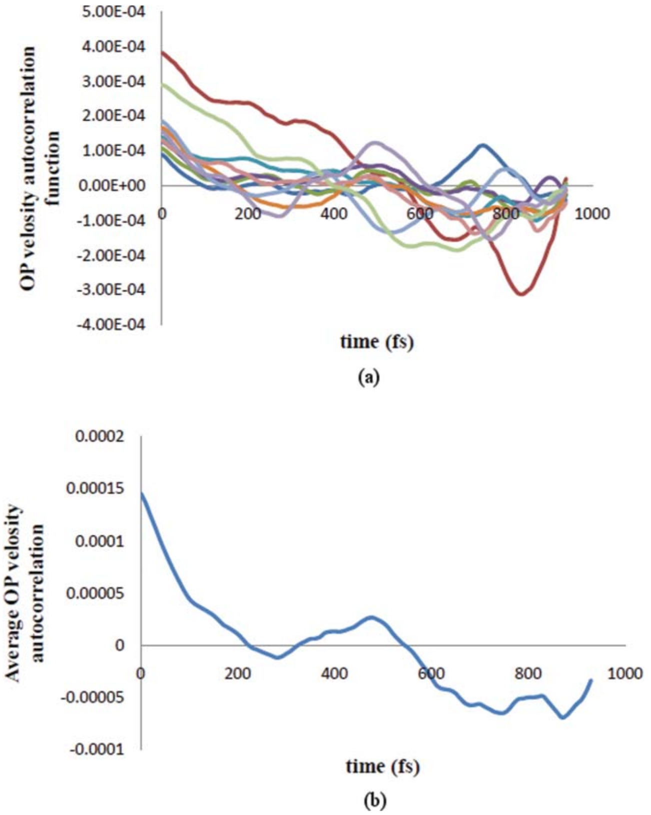

FIG. 1. (a) Ten OP velocity autocorrelation function time-courses for 001Z starting from same initial conditions but different initial velocity random seeds. (b) A Boltzmann average OP velocity autocorrelation function time-course showing the absence of a long-time tail, and hence the lack of coupling to other slow variables not included in the set of OPs. Decay trend similar to a single MD derived autocorrelation function of Fig. S8 (enhanced online) [URL: http://dx.doi.org/10.1063/1.35234532.1].

tems of higher complexity. Furthermore, the end-to-end distance and radius of gyration are accounted for via our more general OPs (shown below). Another OP like variable often found in literature[27] is the number of macromolecular hydrogen bonds. Later in this section we show that our OPs also account for these slow variables.

As system size increases the OPs become better filters of the high frequency fluctuations. To validate this, we simulated hepta-alanine with 8000 water molecules in a cubic box of side 44 Å under settings mentioned in Table I. Fluctuations in OPs for larger systems are found to be much smaller than those of smaller ones [i.e., the same OP shows amplified fluctuations as the system is changed from RNA to hepta-alanine (Fig. S6 in Ref. 33)]. Figure S7 in Ref. 33 compares the moving averages in lower and higher order OPs for RNA and hepta-alanine. Distinct differences in length scales allow better separation of the lower order OPs from those of the higher ones in the RNA. This facilitates filtering of the low and high frequency modes. For smaller systems like hepta-alanine, the length scale separation diminishes. Thus OPs cannot facilitate filtering of high frequency fluctuations, and consequently the implementation of multiscale analysis becomes inefficient.

The choice of OPs depends on the characteristics of the system of interest. This can be understood *a priori* via analyzing OP evolution for short MD trajectories and observing the decay in the OP velocity autocorrelation functions. It is found most efficient for the present problem to only use four OPs (i.e., the center of mass and three corresponding to overall extension–dilatation). To verify completeness of this set of OPs for the present problem, we plot the OP velocity autocorrelation functions for a window of 1 ps in Fig. S8 in Ref. 33. The correlation decays sharply on a time scale much shorter than that of OP evolution (i.e., the OPs were constant over the time of autocorrelation decay). The decay zone is followed by a fluctuating phase that reflects insufficient statistics for constructing long-time correlation function behavior. To illustrate this, we plot several autocorrelation functions for 1 ps trajectories with identical starting structure, initial conditions but different random seeds for generating initial velocities [Fig. 1(a)]. In principle, an average of such single MD simulation derived correlation functions is required to compute the diffusion factors. However, using only the early part of a single MD correlation function (wherein the most statics are accumulated) was found to suffice [Fig. 1(b)]. Furthermore, the correlation analysis validates the completeness of the set of OPs as there is no long-time tail behavior in the correlation functions.

Omission of a slow variable that couples with the existing set can lead to a long-time correlation tail. To validate this, we redid the correlation calculation without the 100X OP. This leads to a long-time tail in the velocity autocorrelation function for the 001Z and 010Y OPs (Fig. S9 in Ref. 33). In general terms, the deformational behavior in a given Cartesian direction is driven by forces that depend on the OPs in all directions. Therefore, a missing OP will create an ensemble of atomic configurations that reflect its absence, which in turn is expressed in slower behavior of the retained OP velocities, and hence the associated autocorrelation functions. Simultaneously, the ensemble had a major population of structures with very high residuals (∼10 Å) also signaling omission of a slow mode. Therefore, the diffusion calculation indicates the absence of a key slow variable that can be optimally added to the existing set via the procedure of Sec. III. Due to orthogonality of our basis functions, the cross-correlation functions between different OPs are several orders of magnitude smaller than the autocorrelation functions; this implies that the OPs are not coupled through the diffusion factors but only through the OP dependence of the thermal-average forces. This greatly simplifies the construction of the random forces as they are related to the diffusion matrix.

Constructing higher order OPs from an MD run via Eq. (2.4) shows that they are highly fluctuating and, therefore, not appropriate as OPs in the sense used here (Fig. S10 in Ref. 33). Rather they are accounted for via the quasiequilibrium probability density (i.e., in the ensemble used to calculate averages). As presented earlier,[4] the ensemble of atomistic configurations is generated via Eq. (2.2). The residuals ($\vec{\sigma}_i$) are generated by a formula similar in structure to that used to obtain the atomic positions but the sum over basis functions does not include those associated with the OPs (Sec. II).[4]

Addition of OPs to a pre-existing set is needed in various cases. If the system is changed, for e.g., if it is composed of multiple macromolecules then more OPs are required to

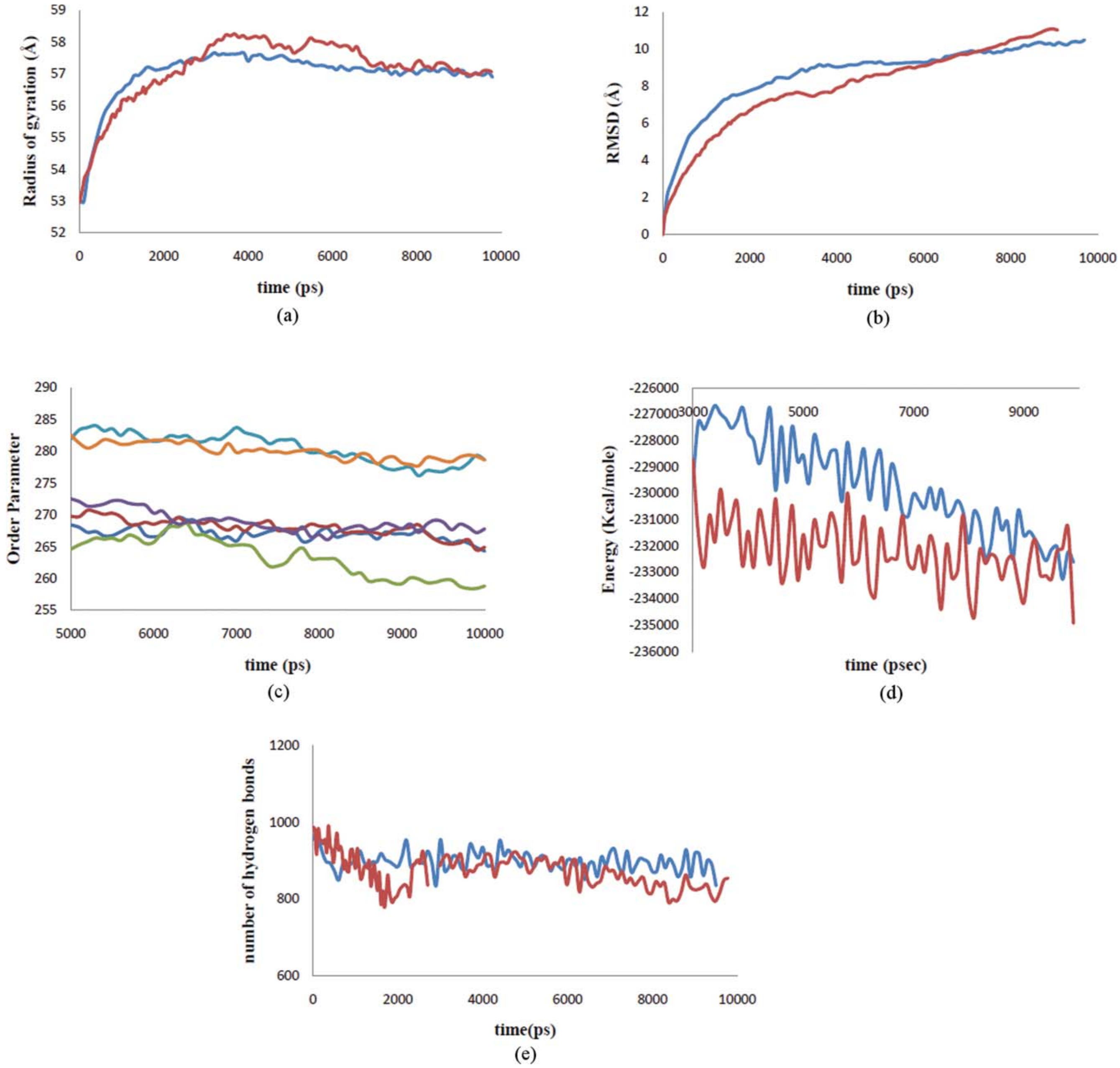

FIG. 2. Time evolution of the (a) radius of gyration via conventional MD (blue) and SIMNANOWORLD (red), (b) RMSD from the 0 ns structure to that after 10 ns for conventional MD (blue) and SIMNANOWORLD (red) simulations, (c) OPs 010Y (MD [blue]; SIMNANOWORLD[orange]), 100X (MD[mauve]; SIMNANOWORLD[red]) and 001Z (MD[deep-blue]; SIMNANOWORLD[green]) showing the OP-equivalence of SIMNANOWORLD and conventional MD, (d) RNA potential energy via MD (red) and SIMNANOWORLD (blue), (e) number of nucleic acid–nucleic acid hydrogen bonds via MD (blue) and SIMNANOWORLD (red) (enhanced online) [URL: http://dx.doi.org/10.1063/1.35234532.2].

form a complete set. The added OPs probe complex intermacromolecular motions. New OPs could also be added in a dynamic fashion in the course of a simulation to account for types of motions absent initially. As mentioned above, the appearance of long-time tails in the correlation functions later in a simulation is a key indicator of the need to augment the set of OPs.

To assess the accuracy of the multiscale OP dynamics comparisons were carried out with conventional MD simulations for trajectories of 10 ns. On removal of the viral capsid, the RNA is no longer constrained and tends to expand. Following initial expansion the RNA shrinks, and finally fluctuates among a range of distinct atomistic states of similar energy. Figure 2(a) shows the radius of gyration obtained with MD and **SIMNANOWORLD**, while Fig. 2(b) shows the progress of the RMSD from the initial structure as a function of time; agreement of the radius, and the RMSD between MD and the multiscale simulation is excellent. Figure S11 in Ref. 33 shows the overall and pentameric structural alignment[51] of the MD and the **SIMNANOWORLD** generated RNA structures at the end of 10 ns. We further plot OP time courses from the final 5 ns of conventional MD and

SIMNANOWORLD [Fig. 2(c)]. These results show that both the structures in Fig. S11 in Ref. 33 are essentially a part of similar OP ensembles having similar overall characteristics and confirm that multiscale simulation is generating configurations consistent with the same value of the OPs that arise in MD. However, it is inappropriate to compare our predictions with that of a single MD since the former corresponds to an ensemble of MD simulations (see below). Significantly, the multiscale simulations capture the overall structural dynamics, which is often the main interest. However, the atomistic configurations are also accounted for via the quasiequilibrium distribution. This illustrates that radius of gyration and end-to-end distance (as mentioned above) is accounted for by our OPs (Fig. S12 in Ref. 33). Figure S13 in Ref. 33 shows the potential energy for the multiscale simulations. It fluctuates about a constant value. Energies show identical trend and are within a percent of those from the MD run. Figure 2(d) is the same as Fig. S13 in Ref. 33 but without the water–water and water–ion interactions. This shows that the RNA gains stability as the potential energy gradually decreases. Energies from the MD and SIMNANOWORLD generated trajectories show excellent agreement in trend as well as in magnitudes. The observed difference is within limits of those from multiple MD or SIMNANOWORLD runs starting from the same initial structure with different initial velocities. As another basis of comparison, time evolutions of the number of intramacromolecular hydrogen bonds for both methods are shown [Fig. 2(e)]. Hydrogen bonds are defined solely on the basis of geometric parameters (bond angle: 20°; bond-length: 3.8 Å) between donors and acceptors. Initial expansion reduced the number of these bonds (primarily the ones involved in the RNA tertiary structure). The number of bonds decreased less rapidly in the later part of the trajectories when expansion ceased. A detailed account of the various types of hydrogen bonds will be given below.

An obvious advantage of multiscaling is the potential to use timesteps of tens or hundreds of ps or greater (in contrast to the 10^{-15} s of conventional MD or 0.2–1 ps for the Langevin PCA approach[16,17]). For relatively slow processes in large systems, the speed-up over conventional MD is significant. To assess the efficiency of our approach, 128 processors were used. During the initial transient, 40 ps timesteps were used. To accommodate the initial expansion and account for the structural anisotropy (Fig. S14 in Ref. 33) the RNA was resolvated in a bigger box (Table I) after the first 3 ns. Post initial transient, Langevin evolution was executed using 150 ps timesteps, reflecting the longer characteristic time for this phase. In this slower evolution regime (probed till 50 ns for the study) efficiency becomes 11 fold. However comparison with a single MD run is not appropriate since SIMNANOWORLD correspond to ensemble MD. In this study a single SIMNANOWORLD simulation corresponds to an ensemble of 168 traditional MD runs. While for each MD run the OP time course is essentially the same as that predicted by SIMNANOWORLD, the detailed atomistic configuration varies dramatically among members of the ensemble. This factor of 168 comes from the sample size used in the Monte Carlo integration to compute the thermal-average forces. Finally, a single MD run may not be representative of an ensemble of possible time courses, which, in contrast is automatically overcome in our approach. If the finer short timescale structural transition is of interest they can be pursued by either shorter timescale traditional MD runs or by including more OPs (although this will decrease the minimum characteristic time of OP dynamics (Fig. S10 in Ref. 33) and, therefore, reduce the efficiency of multiscale simulations). Unlike the Langevin PCA model[16] where single or multiple ns MDs are used to generate input, here only short ps MDs are required to generate a constant OP ensemble and thereby equivalent trajectory ensemble. Selecting OPs for running the multiscale simulation requires an initial analysis of their time trends over 1 ps—1 ns timescale. However, this analysis need not be repeated in the course of simulations until the emergence of new OPs, thereby restoring efficiency of the multiscale simulation.

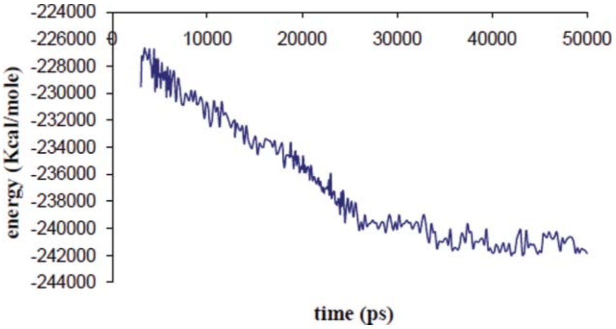

FIG. 3. Time evolution of the RNA potential energy via 50 ns SIMNANOWORLD simulation.

Making use of the above efficiency, we probed the long-time behavior of the RNA with SIMNANOWORLD. In Fig. 3 we plot potential energy versus time for the RNA. The energy decreases and fluctuates about a minimum. The radius of gyration is plotted for the 50 ns trajectory in Fig. 4. Following rapid initial expansion, RNA gradually shrinks for 30 ns before reaching a dynamic equilibrium wherein it fluctuates about 50 Å. These overall changes in shape and size are tracked by our OPs (Fig. S15 in Ref. 33). Their values also gradually decrease before flattening out. However, the magnitude of the three OPs is different, probing different extents of deformation along the three Cartesian axes. This is consistent with the fact that even though overall shape and size follow simple trends (reflected in Fig. 4), the anisotropy in the system leads to a symmetry breaking which is tracked by our OPs

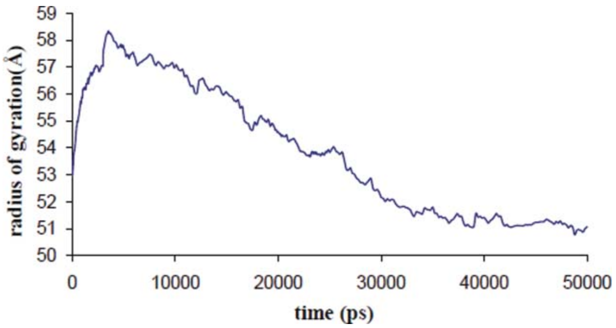

FIG. 4. Time evolution of the RNA radius of gyration via 50 ns SIMNANOWORLD simulation.

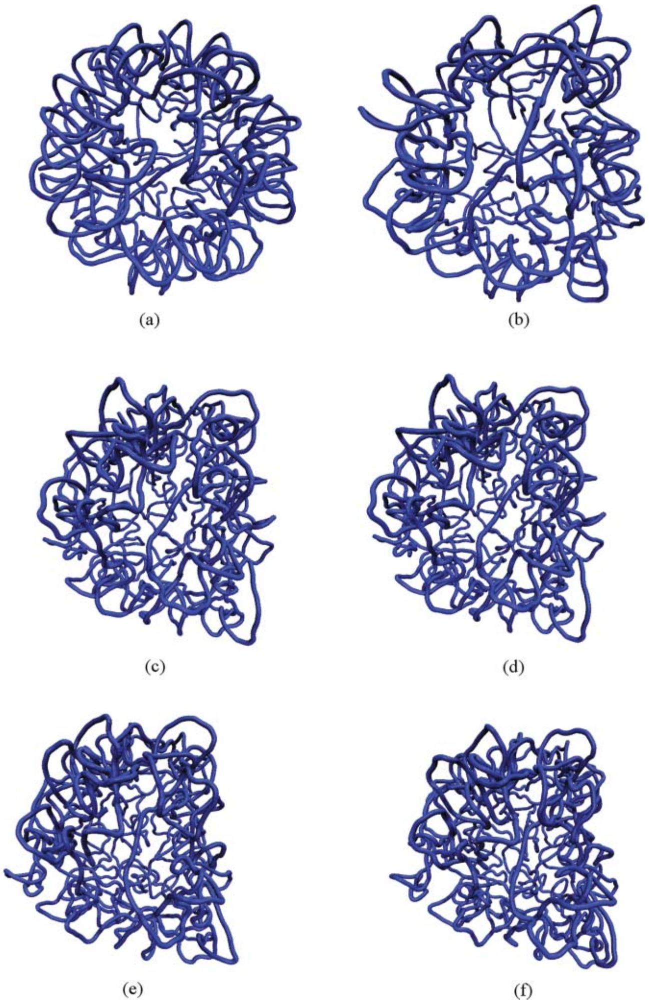

FIG. 5. RNA structure snapshots at (a) 0 ns, (b) 10 ns, (c) 20 ns, (d) 30 ns, (e) 40 ns, and (f) 50 ns.

and the constant OP ensemble. Figure 5 validates that the initial symmetry is completely lost in the course of the simulation. In the final structure (after 50 ns) the tertiary structure of the RNA is highly disrupted, though secondary structure still remained. The latter is in agreement with experiments that suggest free RNA can possess some secondary structure.[51] Figure 6 shows the RMSD over the entire 50 ns trajectory. RMSD shows a rapid increase followed by a gradual one. In Fig. S16 in Ref. 33 we plot RMSD versus energy for the final 20 ns of our trajectory. The increase in RMSD signifies that even though the energy, overall shape and size (OPs) change negligibly, there are local (noncoherent) changes that are accounted for by the constant OP ensemble (i.e., fluctuations in the higher order OPs). Thus SIMNANOWORLD captures the exploration of multiple isoenergetic configurations by the RNA.

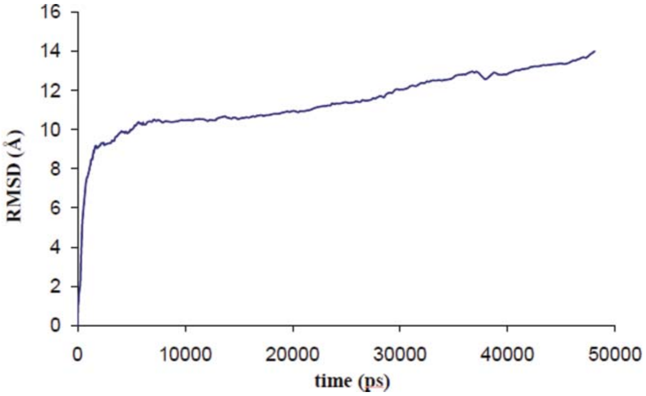

FIG. 6. As in Fig. 2(b) but for a total simulation time of 50 ns.

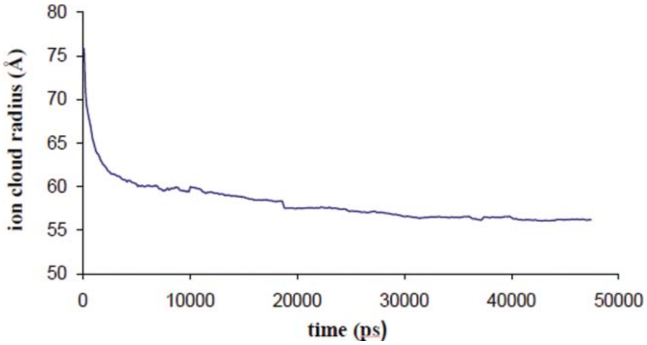

FIG. 7. Time evolution of mobile ion (Na^+) cloud radius over 50 ns.

The gradual shrinkage of RNA is explained on the basis of ion shielding effects. Figure 7 shows the radius of the ion cloud decreases with time. Thus the ion cloud concentrates and distributes about the RNA (Fig. 7 and movie 1 in the supplementary material), shielding the electrostatic repulsion between similarly charged nucleic acid residues in the RNA, causing them to come closer. When the cloud is removed similarly charged groups mutually repel and the RNA expands instead of shrinking. To confirm the above physical picture, the structure at the end of 20 ns was deionized and a further 7.5 ns simulation was carried out in aqueous solvent. The expansion due to electrostatic repulsion is reflected in the radius of gyration and OP changes over this simulation (Fig. S17 in Ref. 33 and movie 2 in supplementary material). Note, since we use 1:1 electrolyte the ions are diffusively bound. Therefore, they exchange positions unlike for tightly bound ions such as Mg^{2+}, justifying their inclusion as a part of the ensemble and not as an OP. Even though the ions are not included in an OP calculation, their rapid quasiequilibrium redistribution accompanying structural changes in the OP defined macromolecule at each Langevin timestep correctly accounts for the ion cloud around the RNA. Closely related to the distribution of ions is the distribution of water and hydrogen bonds. The total number of hydrogen bonds remains constant throughout the 50 ns simulation. However, the number of nucleic acid–water hydrogen bond decreases, while those for water–water hydrogen bonds increases, conserving the total number of bonds (Fig. 8). This phenomenon is consistent with mobile ion screening induced RNA shrinkage. As the RNA shrinks, water from the inner RNA cavity is expelled, thereby increasing the number of bulk water–water interactions. These shifts in sodium ion population, coordinated with hydrogen bond rearrangement, guide the system to the final structure. There is also a redistribution of internucleotide hydrogen bonds as the RNA samples isoenergetic configurations in the final 20 ns. However, the total

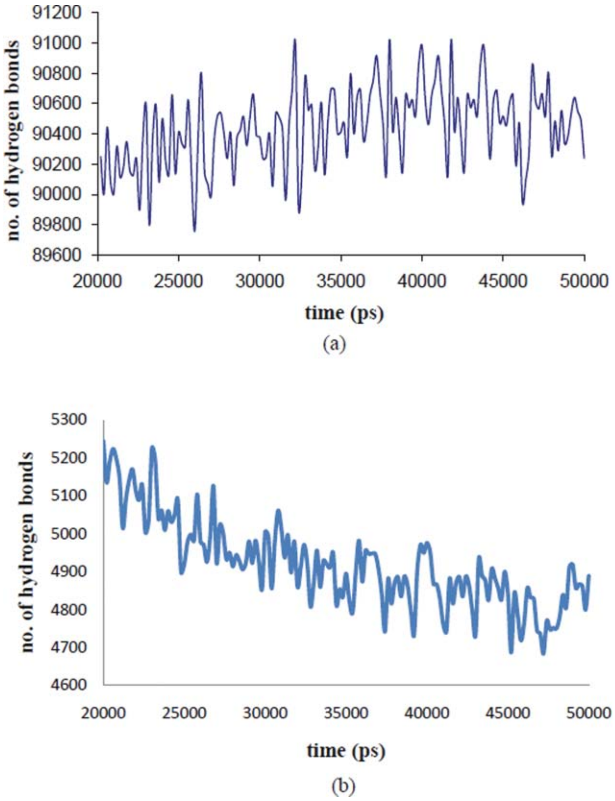

FIG. 8. Time evolution of the water–water (a) and nucleic acid–water (b) hydrogen bonds for simulated RNA dynamics during the final 30 ns of the Fig. 5 simulation.

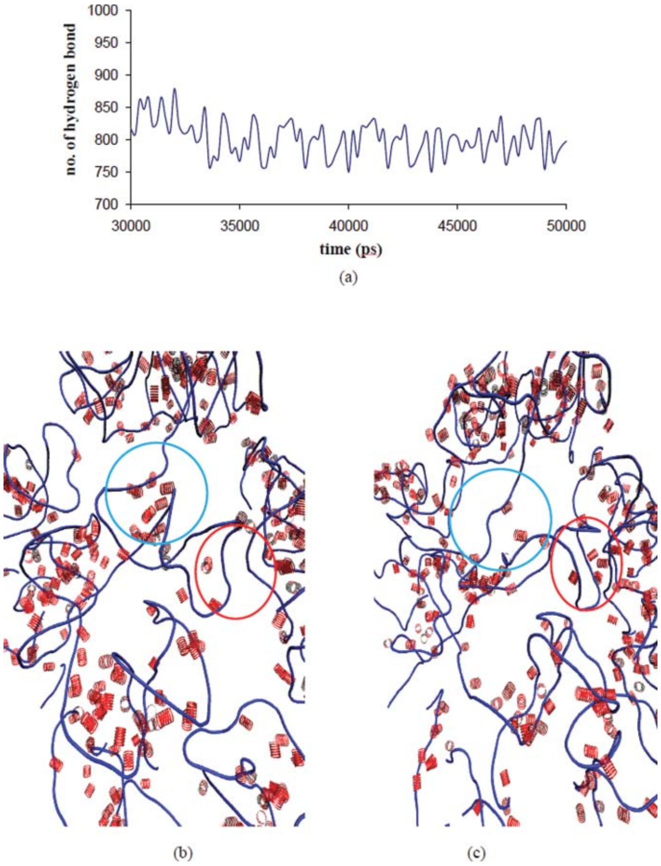

FIG. 9. (a) Time evolution of the number of nucleic acid–nucleic acid hydrogen bonds for the final 20 ns of the Fig. 5 simulation. [(b), (c)] Shift in hydrogen bonds from blue encircled to the red encircled region.

number of nucleic acid–nucleic acid hydrogen bonds is conserved (Fig. 9).

While coherent structural dynamics are tracked by changes in the OPs, additional high frequency macromolecular motions are captured by the residual modified/MD generated ensemble. These high frequency modes capture small-timescale local alterations, over and above the OP mediated deformations in the RNA, and consequent effects on atom scale features like the hydrogen bond distribution. However, other complex and/or slow modes like bending or twisting can arise in the course of the RNA dynamics and affect the hydrogen bonds. As stressed above, emergence of new modes can be captured by our OPs and are signaled by our self-consistency checks. Within the simulated period of time (50 ns), we have not come across any long-time velocity autocorrelation tail or a significant population of high residual structures signifying absence of additional slow modes. A plausible explanation is that in viral RNAs it is possible that the bending modes leading to an unfolding transition appear much later in the time course (>50 ns) of structural evolution or secondary structure disruption occurs at a temperature much higher than 300K.[52,53] Another possibility is the sampling limitation of our implementation. RNA unfolding has often been modeled using higher temperature sampling techniques like replica exchange MD.[54,55] Such sampling techniques can take into account contributions from rare events and prevent entrapment of a structure in deep potential wells. Therefore, they efficiently probe unfolding conformations. Our approach has not yet been modified to take into account the above and, therefore, can suffer from problems of rare event sampling as does conventional MD. However most of the simulation results in this paper are independent of such events as they deal with reaching the energy minima rather than being entrapped in one. Starting from the last few ns of the reported simulation during which the system tends to equilibrate, application of rare event sampling techniques becomes useful in taking the system away from the obtained minima by sampling other free-energy basins. Comments on other efficient sampling techniques are included in Sec. V.

We did a control experiment to compare the deformation in the free RNA versus that in protein encapsulated RNA. We redid the simulation using identical physical conditions (temperature, salinity) and software settings for the RNA core of STMV. The RNA core is composed of capsid protein strands (residue 2–27) complexed with the RNA.[56] This complex is found to be stable with a radius distribution of ∼50 Å. The added protein segments complex with the RNA, reducing the degrees of freedom. The structure was energy minimized and thermally equilibrated before starting the multiscale simulation. Time evolution of the OPs, RMSD and structures for this simulation are shown in Fig. 10. Figure 10(b) also shows the RMSD for free RNA. RMSD of RNA in the bound state is much less than in the free state. Unlike the previous case where the difference in the OPs was large, here the difference is small and their change is slower; this suggests the preservation of the symmetry originally imposed by the

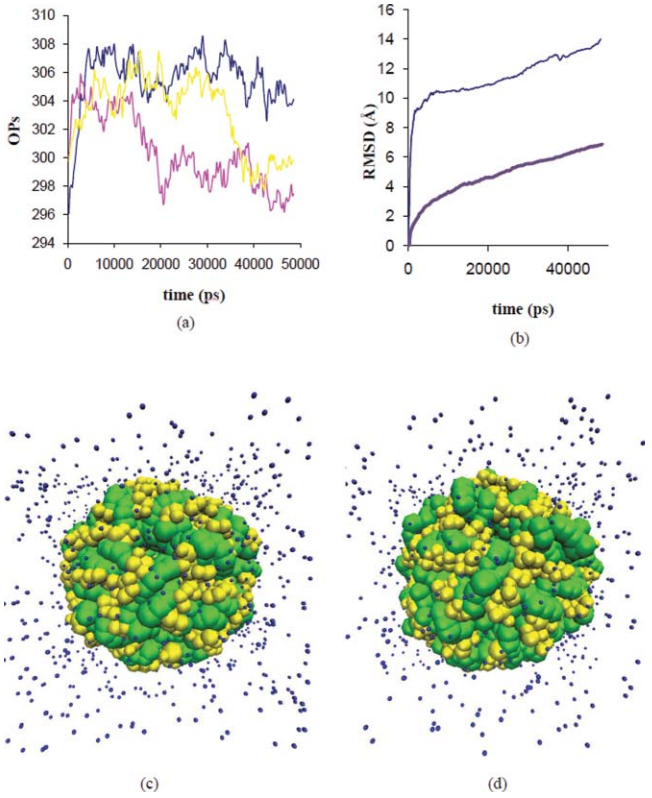

FIG. 10. Time evolution of the Ops (a) RMSD (b) and structure [(c),(d)] for the protein bound RNA in 0.3 M NaCl solution at 300K showing the restriction on RNA motion imposed by the proteins.

capsid. Thus changes in the protein RNA complex are much less than those of free RNA. This longer characteristic time allows more efficient application of OP dynamics as now timesteps of the order of 250 ps are possible. This implies that **SIMNANOWORLD** is 16 times faster than a single conventional MD for the present problem.

V. CONCLUSIONS

An efficient multiscale algorithm that probes the dynamics of macromolecules using OPs is demonstrated and validated. The OPs are shown to be slowly varying and hence are well suited as a basis of our multiscale algorithm. Completeness of a set of OPs is determined by the shortness of correlation times relative to the characteristic time of OP evolution. Automated construction of order parameters enables efficient augmentation of the set of OPs to address incompleteness. **SIMNANOWORLD** results show excellent agreement with those from a conventional MD run. **SIMNANOWORLD** efficiency increases for larger systems undergoing slow transformations, as these allow significant length and timescale separation for the OPs to filter out the high frequency fluctuations from the coherent dynamics. Thus OP mediated coarse graining of the free-energy landscape allows for a Langevin timestep of a few hundred ps (10^5 times greater than conventional MD). **SIMNANOWORLD** predictions correspond to an ensemble of MD trajectories and hence are more statistically significant than results from a single MD run. Multiscale simulation via the OP description is found to capture significant structural details like ion screening, hydrogen bond rearrangement, and symmetry breaking transitions.

The approach presented has been successful for phenomenon that does not involve high barrier-crossing events. However, this does not imply that our approach is lacking in the basic physical framework; rather it implies that simulating very high barrier-crossing events is computationally challenging.[54] Note, as OPs coarse grain the free-energy landscape, some of the low energy barriers are naturally addressed. A number of authors have presented approaches for simulating rare barrier crossing events and corresponding applications to free-energy profiling. An incomplete list includes umbrella sampling,[57] thermodynamic integration,[58–60] metadynamics,[61,62] and adiabatic molecular dynamics.[38] More recently, a multiscale conformation space exploration scheme is developed and applied to proteins.[63] These methods can be used to complement the present approach to address rare events and hence explicitly construct the free-energy landscape. In this direction, potential modifications of our approach are as follows: (i) In analogy to metadynamics, sequential free-energy basin discovery can be implemented via a memory dependent thermal force averaging that minimizes the forces which would have driven the systems to basins already discovered in an earlier step. With this, each sequential elimination simulation is guaranteed to find a new basin. (ii) In analogy with, Ref. 64 our procedure can be run at a set of high temperatures, for each of which the first passage time between two basins of interest is computed. This high temperature data can be fit to an Arrhenius-type law and the results are extrapolated down to the temperature of interest. This yields an approximation to the rate of crossing high free-energy barriers. Such modifications were not required in the current simulations as in these we aimed to probe the most likely pathways along the free-energy surface and not rare events.

Recently, schemes like reconstruction algorithm for coarse-grained structures[65] have been developed for constructing low-energy all-atom protein structures form configurations with only C_α atoms. In the present context, this could facilitate **SIMNANOWORLD** simulations starting from low resolution structures (e.g., electron cryomicroscopic models[66]) via reconstructing all-atom input structures. A higher order Langevin solver shall be incorporated into the **SIMNANOWORLD** implementation. This will utilize data from multiple time intervals thereby increasing the accuracy of the coherent part of the Langevin dynamics and over-all CPU efficiency. In summary, we show that our OP-multiscale computational approach is ideally suited for studying structural dynamics in large macromolecules and macromolecular complexes.

ACKNOWLEDGMENTS

We appreciate efforts of the JCP editor and reviewers whose suggestions greatly improved the manuscript. We thank Professor Srini Iyengar for fruitful discussions on extended Lagrangian approaches. This project was supported in part by the National Science Foundation (Collaborative Research in Chemistry program), National Institutes of Health (NIBIB), Department of Energy (Office of Basic Science), METAcyt through the Center of Cell and Virus Theory, and Indiana University College of Arts and Sciences.

APPENDIX A: LINEAR RELATIONSHIP BETWEEN OPS AND POSITIONS

The relationship between $\vec{\Phi}_k$ and $\vec{r}_i$ is taken here to be linear. This ensures that $\vec{r}_i$ has a unique value for a given set of $\vec{\Phi}_k$ and residuals Eq. (2.2). Should this not be the case then, as Newton's equations evolve r, the system could spontaneously transition to another state of order without a change of microstate. By similar arguments application of $\sum_{k'} B_{kk'}\vec{\Phi}_{k'} = \sum_{i=1}^{N} m_i U_k(\vec{r}_i^{\,0})\vec{g}(\vec{r}_i)$, where $\vec{g}$ is a function of $\vec{r}_i$, may not always be suitable. This would have been implied by replacing Eq. (2.2) with $\vec{g}(\vec{r}_i) = \sum_k \vec{\Phi}_k U_k(\vec{r}_i^{\,0}) + \vec{\sigma}_i$. If $\vec{g}$ is linear in $\vec{r}_i$, then the unique relation between $\vec{\Phi}_k$, residuals and $\vec{r}_i$ holds, allowing for the multiscale analysis of Sec. III. However, if $\vec{g}$ is a nonlinear function of $\vec{r}_i$ then this uniqueness could be lost; multiple solutions for $\vec{r}_i$ could exist for a given set of $\vec{\Phi}_k$ and residuals. This could create situations wherein an initial r state evolves to multiple states allowed by the nonlinearity of $\vec{g}$. Newtonian mechanics prohibits this dynamical bifurcation of states (i.e., simultaneous evolution of one state into multiple ones); hence, in the present OP construction formalism, inclusion of a nonlinear function $\vec{g}$ could lead to unphysical results if uniqueness of the $r - \Phi$ relationship is violated. Other transformations can be designed that allow for nonlinear combinations of r without violating this uniqueness. However, this was not pursued here. The suggested bifurcation of states should not be confused with the multiplicity of atomic configurations that arise due to $\vec{\sigma}_i$ sampling (Sec. II). The present formulation does not imply that our OP dynamics is linear, i.e., the thermal-average forces driving OP dynamics in general are related to the OPs in a highly nonlinear fashion. This nonlinearity is critical in simulating far-from equilibrium structures.[27]

In the above context, relationship (2.4) is the origin of the unique value of $\vec{\Phi}_k$ for a given set of $\vec{r}_i$. While r implies $\vec{\Phi}_k$ uniquely, the converse is not true, i.e., there is an ensemble of r for given Φ. This stems from the fact that a theory with $N_{\mathrm{OP}}(\ll N)$ OPs cannot predict $3N$ atomic co-ordinates uniquely; this is the motivation for adding the residuals to Eq. (2.1) and generating an ensemble of atomic configurations consistent with the OPs in Eq. (2.2). In particular, N $\vec{r}_i$ cannot be uniquely expressed in terms of $N_{OP}\vec{\Phi}_k$ from Eqs. (2.3) or (2.4). Therefore, the $r - \Phi$ relationship is not one-to-one, as it should not be.

Generation of atomic ensemble consistent with coarse-grained variables has been discussed in other multiscale approaches.[67] However, in these approaches the dynamics of all atoms was not accounted for, leading to issues in treating diffusion and long range electrostatics.[67] One suggested way of overcoming this is to couple the atomistic and coarse-grained representations via the boundary conditions.[67] This has been implemented by parameterization of a coarse-grained model with MD simulation data, and demonstrated on transmembrane proteins.[67] In contrast, the present approach accounts for the all-atom configurations via a quasiequilibrium probability distribution, which evolves adiabatically with the OPs. Viscoelastic effects in the dynamics of the macromolecule are accounted via thermal-average forces and diffusion coefficients (Secs. II and III).

APPENDIX B: DERIVATION OF THE MULTISCALE LIOUVILLE OPERATOR

Here we derive the multiscale Liouville operator of Eq. (3.2) using the *ansatz* Eq. (3.1) and the chain rule, starting from the classical Liouville operator $\mathcal{L}$. With this, the contribution to $\mathcal{L}\rho$ from particle i is given by

$$-\frac{\vec{p}_i}{m_i}\cdot\frac{\partial\rho}{\partial\vec{r}_i} - \vec{F}_i\cdot\frac{\partial\rho}{\partial\vec{p}_i} - \sum_{\alpha,\alpha',k}\frac{p_{i\alpha}}{m_i}\left(\frac{\partial\Phi_{\alpha'}}{\partial r_{i\alpha}}\right)\left(\frac{\partial\rho}{\partial\Phi_{\alpha'}}\right)_{\Gamma=\Gamma_0,} \tag{B1}$$

where α and α' signifies Cartesian components of the position/momentum and OPs respectively.

The first two terms in Eq. (B1) are the ith contribution to $\mathcal{L}_0$. Extending the above to N particles,

$$\mathcal{L}_0 = -\sum_{i=1}^{N}\left\{\frac{\vec{p}_i}{m_i}\cdot\frac{\partial}{\partial\vec{r}_i} + \vec{F}_i\cdot\frac{\partial}{\partial\vec{p}_i}\right\}. \tag{B2}$$

Furthermore considering the third term of Eq. (B1), the $\vec{\Phi}_k - \vec{r}$ relationship (II.4), and the definition of μ_k (II.6) yields

$$\mathcal{L}_1 = \sum_k \frac{\vec{\Pi}_k}{\mu_k}\left(\frac{\partial\rho}{\partial\vec{\Phi}_k}\right)_{\Gamma=\Gamma_0}. \tag{B3}$$

This justifies the form of Liouville operator in Eq. (3.2) through Eq. (3.4).

APPENDIX C: THERMAL-AVERAGE FORCES AS FREE-ENERGY GRADIENTS

By definition the αth Cartesian component of the thermal-average force is given by

$$f_{k\alpha} = -\frac{\partial F}{\partial\Phi_{k\alpha}} = \frac{1}{\beta Q(\beta,\Phi)}\frac{\partial}{\partial\Phi_{k\alpha}}\int \omega d\Gamma^*\Delta(\Phi-\Phi^*)e^{-\beta H^*}, \tag{C1}$$

where Φ^* is the set of the first N_{OP} components of Φ^*_{ex} (i.e.,Φ_{ex} evaluated at Γ^*); ω and Δ are defined in Appendix SII and references therein. The contribution $\tilde{Q}$ from the r^* integration of Eq. (C1) is given by

$$\tilde{Q} = \int \omega d^3r^*\Delta(\Phi-\Phi^*)e^{-\beta H^*}. \tag{C2}$$

As Φ does not affect the momentum contribution to the Q-integral in deriving the thermal-average forces, we proceed with the analysis of $\partial\tilde{Q}/\partial\Phi_{k\alpha}$.

Note, the thermal averaging involves integration over all positions and momenta Γ^* which are consistent with a given value of Φ as imposed through the $\triangle$ factor. In taking the derivative of the integral with respect to $\Phi_{k\alpha}$, in concept, one calculates the integral at two closely lying $\Phi_{k\alpha}$ values and divides the difference by the increment in $\Phi_{k\alpha}$; at all stages of this conceptual calculation, H is taken to be a function of Γ^* and not Φ^* directly. This implies the $\mathcal{L}_1 H = 0$ condition encountered in the course of constructing the multiple dependencies of ρ_0 is not violated.

Taking the derivative inside the integral in Eq. (C1), and using the fact that $\triangle$ only depends on the difference ($\Phi - \Phi^*$), one obtains

$$\frac{\partial \tilde{Q}}{\partial \Phi_{k\alpha}} = -\int \omega d^3 r^* \frac{\partial \Delta(\Phi - \Phi^*)}{\partial \Phi^*_{k\alpha}} e^{-\beta H^*}. \tag{C3}$$

Using Eq. (2.7) and the chain-rule yields

$$\begin{aligned}\frac{\partial \tilde{Q}}{\partial \Phi_{k\alpha}} &= -\int \omega d^3 r^* \sum_{\alpha'=1}^{3} \sum_{i=1}^{N} \frac{\partial \Delta(\Phi - \Phi^*)}{\partial r^*_{i\alpha'}} \frac{\partial r^*_{i\alpha'}}{\partial \Phi^*_{k\alpha}} e^{-\beta H^*} \\ &= -\int \omega d^3 r^* \sum_{\alpha'=1}^{3} \sum_{i=1}^{N} \delta_{\alpha\alpha'} U_k\left(\vec{r}_i^{\,0}\right) \frac{\partial \Delta(\Phi - \Phi^*)}{\partial r^*_{i\alpha'}} e^{-\beta H^*},\end{aligned} \tag{C4}$$

where Kronecker delta $\delta_{\alpha\alpha\prime}$ arises since $\partial r^*_{i\alpha\prime}/\partial \Phi^*_{k\alpha}$ is zero if $\alpha \neq \alpha\prime$. Integration by parts of Eq. (C4), and simplification of the expression yield

$$\frac{1}{\beta} \frac{\partial \tilde{Q}}{\partial \Phi_{k\alpha}} = \int \omega d^3 r^* \Delta(\Phi - \Phi^*) f^{m*}_{k\alpha} e^{-\beta H^*}, \tag{C5}$$

where $f^{m*}_{k\alpha} = \sum_i U_k(\vec{r}_i^{\,0}) F^*_{i\alpha}$, for $F^*_{i\alpha} = -\partial V^* / \partial r^*_{i\alpha}$. Here $V^* = V(\vec{r}^{\,*}_1, \vec{r}^{\,*}_2 \cdots \vec{r}^{\,*}_N)$ is the N-atom potential evaluated at r^*.

Reinstating vector notation, integrating over the momenta, and dividing both sides by Q yields

$$\vec{f}_k = \frac{1}{Q} \int \omega d\Gamma^* \Delta(\Phi - \Phi^*) \vec{f}^{\,\mathrm{m}*}_k e^{-\beta H^*} = \left\langle \vec{f}^{\,\mathrm{m}*}_k \right\rangle. \tag{C6}$$

This validates Eq. (3.7) and also shows how the interatomic forces $\vec{F}_i$ influence $\vec{f}_k$ through the N-atom potential in H. On arriving at Eq. (C6), all the position dependence of H is relegated to r^*. Therefore, the condition $\mathcal{L}_1 H = 0$ is sustained.

APPENDIX D: λ COMPUTATION

Fluctuations λ in SPs is calculated using $\lambda(SP_j, t_i) = (\sqrt{\sum_{i=0}^{N_f - 1} [SP_j(t_i) - \langle SP_j \rangle]^2 / N_f}) / SP_{j,\max}(t_i)$; $j = 1, \cdots 4$, where t_i is the ith time, N_f is the total number of MD time frames used for the moving averages, $\langle SP_j \rangle$ is the moving average (over 50 ps), and $SP_{j,\max}$ is the maximum absolute value of the SP within the range of SPs sampled for the moving average calculation. Fluctuations are defined about a moving absolute maxima rather than a moving average to avoid singularities for zero moving averages. The normalization makes λ dimensionless. Thus λ can be compared between different SPs and OPs.

[1] J. R. Williamson, Nat. Chem. Biol. **4**, 458 (2008).
[2] P. Ortoleva, J. Phys. Chem. B **109**, 21258 (2005).
[3] P. Ortoleva, P. Adhangale, S. Cheluvaraja, M. W. A. Fontus, and Z. Shreif, IEEE Eng. Med. Biol. Mag. **28**, 70 (2009).
[4] S. Cheluvaraja and P. Ortoleva, J. Chem. Phys. **132**, 075102 (2010).
[5] S. Cheluvaraja, A. Roy, and P. Ortoleva, "Roadmap for SimNanoworld$^{\mathrm{TM}}$ an all-atom nanosystem simulator. In preparation, 2011 (unpublished).
[6] K. Jaqaman and P. J. Ortoleva, J. Comput. Chem. **23**, 484 (2002).
[7] Y. Miao and P. Ortoleva, J. Comput. Chem. **30**, 423 (2009).
[8] Y. Miao and P. Ortoleva, Biopolymers **93**, 61 (2010).
[9] S. Pankavich, Y. Miao, J. Ortoleva, Z. Shreif, and P. Ortoleva, J. Chem. Phys. **128**, 234908 (2008).
[10] S. Pankavich, Z. Shreif, Y. Miao, and P. Ortoleva, J. Chem. Phys. **130**, 194115 (2009).
[11] Z. Shreif, P. Adhangale, S. Cheluvaraja, R. Perera, R. J. Kuhn, and P. Ortoleva, Sci. Model. Simul. **15**, 363 (2008).
[12] R. Zwanzig, *Nonequilibrium Statistical Mechanics* (Oxford University Press, New York, 2001).
[13] C. Chen and Y. Xiao, Biophys. J. **88**, 3276 (2005).
[14] S. Hayward, A. Kitao, and H. J. C. Berendsen, Proteins: Struct., Funct., Bioinf. **27**, 425 (1997).
[15] S. Hayward, A. Kitao, and N. Go, Proteins: Struct., Funct., Genet. **23**, 177 (1995).
[16] A. Altis, P. H. Nguyen, R. Hegger, and G. Stock, J. Chem. Phys. **126**, 2444111 (2007).
[17] L. Riccardi, P. H. Nguyen, and G. Stock, J. Phys. Chem. B **113**, 16660 (2009).
[18] Z. Shreif and P. Ortoleva, J. Stat. Phys. **130**, 669 (2008).
[19] A. Arkhipov, P. L. Freddolino, and K. Schulten, Structure **14**, 1767 (2006).
[20] Z. Zhang, J. Pfaendtner, A. Grafmüller, and G. A. Voth, Biophys. J. **97**, 2327 (2009).
[21] Z. Zhang, L. Lu, W. G. Noid, V. Krishna, J. Pfaendtner, and G. A. Voth, Biophys. J. **95**, 5073 (2008).
[22] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen, Proteins: Struct., Funct., Bioinf. **17**, 412 (1993).
[23] G. G. Maisuradze and D. M. Leitner, Proteins: Struct., Funct., Bioinf. **67**, 569 (2007).
[24] K. Hinsen, Proteins: Struct., Funct., Bioinf. **64**, 795 (2006).
[25] Y. Miao and P. Ortoleva, J. Chem. Phys. **125**, 44901 (2006).
[26] Y. Miao and P. J. Ortoleva, J. Chem. Phys. **125**, 214901 (2006).
[27] P. Das, M. Moll, H. Stamati, L. E. Kavraki, and C. Clementi, Proc. Natl. Acad. Sci. U.S.A. **103**, 9885 (2006).
[28] H. Gohlke and M. F. Thorpe, Biophys. J. **91**, 2115 (2006).
[29] Y. G. Kevrekidis, C. W. Gear, and G. Hummer, AIChE J. **50**, 1346 (2004).
[30] E. Weinan, B. Engquist, X. Li, W. Ren, and E. Vanden-Eijnden, Comm. Comp. Phys. **2**, 367 (2007).
[31] S. Izvekov and G. A. Voth, J. Phys. Chem. B **109**, 2469 (2005).
[32] S. Izvekov and G. A. Voth, J. Chem. Phys. **123**, 134105 (2005).
[33] See Supplementary Material at http://dx.doi.org/10.1063/1.3524532 for appendices SI, SII, SIII, SIV, and Figs. S1–S17.
[34] J. Cao and G. J. Martyna, J. Chem. Phys. **104**, 2028 (1996).
[35] D. Marx, M. E. Tuckerman, and G. J. Martyna, Comput. Phys. Commun. **118**, 166 (1999).
[36] R. Car and M. Parrinello, Phys. Rev. Lett. **55**, 2471 (1985).
[37] S. S. Iyengar, H. B. Schlegel, J. M. Millam, G. A. Voth, G. E. Scuseria, and M. J. Frisch, J. Chem. Phys. **115**, 10291 (2001).
[38] L. Rosso, P. Minary, Z. Zhu, and M. E. Tuckerman, J. Chem. Phys. **116**, 4389 (2002).
[39] S. Pankavich and P. Ortoleva, J. Math. Phys. **51**, 063303 (2010).
[40] J. M. Deutch and I. Oppenheim, Faraday Discuss. Chem. Soc. **83**, 1 (1987).
[41] L. Maragliano and E. Vanden-Eijnden, Chem. Phys. Lett. **426**, 168 (2006).
[42] J. B. Abrams and M. E. Tuckerman, J. Phys. Chem. B **112**, 15742 (2008).
[43] Y. Miao and P. Ortoleva, J. Phys. Chem. B **114**, 11181 (2010).
[44] C. R. Sweet, P. Petrone, V. S. Pande, and J. A. Izaguirre, J. Chem. Phys. **128**, 145101 (2008).
[45] S. Pankavich, Z. Shreif, and P. Ortoleva, Phys. A **387**, 4053 (2008).
[46] P. L. Freddolino, A. S. Arkhipov, S. B. Larson, A. McPherson, and K. Schulten, Structure **14**, 437 (2006).
[47] A. Singharoy, A. Yesnik, and P. Ortoleva, J. Chem. Phys. **132**, 174112 (2010).
[48] J. M. Ginder and A. J. Epstein, Phys. Rev. B **41**, 10674 (1990).
[49] M. Severin and O. Inganas, Europhys. Lett. **25**, 347 (1994).
[50] V. A. Ivanov, M. R. Stukan, M. Müller, W. Paul, and K. Binder, J. Chem. Phys. **118**, 10333 (2003).
[51] A. Schneemann, Annu. Rev. Microbiol. **60**, 51 (2006).
[52] M. M. Lin, L. Meinhold, D. Shorokhov, and A. H. Zewail, Phys. Chem. Chem. Phys. **10**, 4227 (2008).
[53] E. J. Sorin, M. A. Engelhardt, D. Herschlag, and V. S. Pande, J. Mol. Biol. **317**, 493 (2002).
[54] Y. Sugita and Y. Okamoto, Chem. Phys. Lett. **314**, 141 (1999).

[55] A. Villa, E. Widjajakusuma, and G. Stock, J. Phys. Chem. B **112**, 134 (2008).
[56] J. Day, Y. G. Kuznetsov, S. B. Larson, A. Greenwood, and A. McPherson, Biophys. J. **80**, 2364 (2001).
[57] G. M. Torrie and J. P. Valleau, J. Comput. Phys. **23**, 187 (1977).
[58] J. G. Kirkwood, J. Chem. Phys. **3**, 300 (1935).
[59] E. A. Carter, G. Ciccotti, J. T. Hynes, and R. Kapral, Chem. Phys. Lett. **156**, 472 (1989).
[60] P. Fleurat-Lessard and T. Zeigler, J. Chem. Phys. **123**, 084101 (2005).
[61] A. Laio and M. Parinello, Proc. Natl. Acad. Sci. U.S.A. **99**, 12562 (2002).
[62] M. Ianuzzi, A. Laio, and M. Parinello, Phys. Rev. Lett. **90**, 238302 (2003).
[63] A. Shehu, L. E. Kavraki, and C. Clementi, Proteins: Struct., Funct., Genet. **76**, 837 (2009).
[64] V. S. Pande and D. S. Rokhsar, Proc. Natl. Acad. Sci. U.S.A. **96**, 9062 (1999).
[65] A. P. Heath, L. E. Kavraki, and C. Clementi, Proteins: Struct., Funct., Genet. **68**, 646 (2007).
[66] Y. Modis, B. L. Trus, and S. C. Harrison, EMBO J. **21**, 4754 (2002).
[67] G. S. Ayton and G. A. Voth, J. Struct. Biol. **157**, 570 (2007).