Title

The multi-dimensional generalized Langevin equation for conformational motion of proteins

Permalink

https://escholarship.org/uc/item/56m8v955

Journal

The Journal of Chemical Physics, 150(17)

ISSN

0021-9606

Authors

Lee, Hee Sun
Ahn, Surl-Hee
Darve, Eric F

Publication Date

2019-05-07

DOI

10.1063/1.5055573

The Journal of Chemical Physics

The multi-dimensional generalized Langevin equation for conformational motion of proteins

Cite as: J. Chem. Phys. **150**, 174113 (2019); https://doi.org/10.1063/1.5055573
Submitted: 07 September 2018 . Accepted: 14 April 2019 . Published Online: 06 May 2019

Hee Sun Lee, Surl-Hee Ahn, and Eric F. Darve

AIP Publishing

The multi-dimensional generalized Langevin equation for conformational motion of proteins

Cite as: J. Chem. Phys. 150, 174113 (2019); doi: 10.1063/1.5055573
Submitted: 7 September 2018 • Accepted: 14 April 2019 •
Published Online: 6 May 2019

Hee Sun Lee,[1,a] Surl-Hee Ahn,[2,b] and Eric F. Darve[1,c]

AFFILIATIONS

[1] Mechanical Engineering Department, Stanford University, Stanford, California 94305, USA
[2] Chemistry Department, Stanford University, Stanford, California 94305, USA

[a] Electronic mail: hslee88@stanford.edu
[b] Electronic mail: sahn1@stanford.edu
[c] Electronic mail: darve@stanford.edu

ABSTRACT

Using the generalized Langevin equation (GLE) is a promising approach to build coarse-grained (CG) models of molecular systems since the GLE model often leads to more accurate thermodynamic and kinetic predictions than Brownian dynamics or Langevin models by including a more sophisticated friction with memory. The GLE approach has been used for CG coordinates such as the center of mass of a group of atoms with pairwise decomposition and for a single CG coordinate. We present a GLE approach when CG coordinates are multiple generalized coordinates, defined, in general, as nonlinear functions of microscopic atomic coordinates. The CG model for multiple generalized coordinates is described by the multidimensional GLE from the Mori-Zwanzig formalism, which includes an exact memory matrix. We first present a method to compute the memory matrix in a multidimensional GLE using trajectories of a full system. Then, in order to reduce the computational cost of computing the multidimensional friction with memory, we introduce a method that maps the GLE to an extended Markovian system. In addition, we study the effect of using a nonconstant mass matrix in the CG model. In particular, we include mass-dependent terms in the mean force. We used the proposed CG model to describe the conformational motion of a solvated alanine dipeptide system, with two dihedral angles as the CG coordinates. We showed that the CG model can accurately reproduce two important kinetic quantities: the velocity autocorrelation function and the distribution of first passage times.

I. INTRODUCTION

Many scientific problems deal with time-evolution equations that are computationally too costly to solve in full resolution of a system because of the system size and the time scale involved in the problems. For example, atomistic descriptions of biomolecular systems of practical interest often become infeasible to simulate for the time period needed to investigate the phenomena of interest with current computational power. A strategy to simulate those systems is to describe the system in reduced dimensions, which can be divided into two tasks: (1) identification of the coordinates that effectively describe the characteristics that we are interested in and (2) finding a governing equation for those selected coordinates, which should account for the effect of ignored degrees of freedom. Additionally, we need an efficient method to solve this governing equation numerically. Following the common naming convention in the context of biomolecular systems, the description in reduced dimensions will be referred to as a coarse-grained (CG) model; and the selected coordinates will be referred to as CG coordinates and the governing equation for CG coordinates as a CG equation.

For a chosen set of CG coordinates, a CG equation can be obtained using the Mori-Zwanzig (MZ) formalism.[1–3] This CG equation consists of three terms: the mean term, the memory term, and the fluctuation term. The exact memory term given by the MZ formalism is infeasible to compute for practical applications. To use this CG equation for applications, often approximations are made on the memory term and the fluctuation term is modeled. With a certain type of approximation on the memory term, the CG equation given by the MZ formalism reduces to the generalized Langevin equation (GLE). Since the memory term in the GLE can be calculated from trajectories of the full dynamics, the GLE has been used as a CG equation for many applications.[4–8] In those previous studies, the CG coordinates were center of mass (COM) of a group of atoms[5–8] or a single coordinate.[4,9,10] Multiple CG coordinates were only used with the Markovian Langevin equation (LE) in which the memory kernel is approximated by a delta function.

When COMs are chosen as CG coordinates, we can approximate the CG forces as pairwise forces, which lead to dissipative particle dynamics (DPD) models.[5–8] In this case, the memory kernel for each pairwise interaction in each direction is a scalar function. In Ref. 4, a single curved coordinate was used for the GLE to describe the conformational motion of the hexapeptide neurotensin. This coordinate was found by first using principal component analysis (PCA), and the first three principal components were reduced to a single curved coordinate by inspection. In Ref. 11, multiple CG coordinates were used for the Markovian LE to describe the conformational dynamics of heptaalanine. These CG coordinates were carefully chosen so that they represent slow and large-amplitude motions. This choice of CG coordinates is required to use the Markovian LE since the equation is only valid when the time scale of the motion of the CG coordinates is separable from that of the unresolved variables. To find those CG coordinates, the authors used dihedral angle based PCA[12] and did careful inspection of values of those principal components.

However, for some applications, those previously used CG coordinates are hard to identify. For example, choosing COM as a CG coordinate is only intuitive when the group of atoms are connected by bonds; finding a single coordinate that effectively describes a system is a complicated task. To use the Markovian LE, determining CG coordinates that are time scale separable from other coordinates is not always possible. So, it is important to be able to use the GLE model with multiple generalized coarse variables.

In this paper, we present the GLE when the CG coordinates are multiple generalized coordinates, which we will call the "multidimensional GLE." Since the forces on CG coordinates have nonzero correlation, the full memory matrix with nonzero off-diagonal entries should be considered. Also the mass matrix in the CG equation is a full matrix which is a function of CG coordinates. We also include the more accurate mean force term that has additional terms besides the gradient of the free energy; these additional terms are nonzero for CG coordinates that are nonlinear functions of atomistic Cartesian coordinates. Then to efficiently solve the multidimensional GLE, we map the multidimensional GLE to a Markovian system in higher dimensions. Although the mapping of the GLE to an extended Markovian system has been discussed and used in previous studies,[8–10,13] those studies did not discuss the case of the multidimensional GLE for which inclusion of the full memory matrix makes the mapping procedure more involved. We present a novel procedure for the mapping of the multidimensional GLE to an extended Markovian system.

We applied our approach to describe the conformational motion of the solvated alanine dipeptide system starting from an atomistic description. The two dihedral angles of the molecule were used as CG coordinates. We show that our CG model accurately describes the conformational dynamics of the peptide in terms of two key kinetic properties: the velocity autocorrelation (VAC) function and the distribution of first passage times (FPTs).

The rest of the paper is organized as follows. In Sec. II, we briefly review the MZ formalism and describe our new approach to calculate the memory kernel in the multidimensional GLE. Then, we present a procedure to map the multidimensional GLE to an extended Markovian system. In Sec. III, the CG model of the alanine dipeptide system is constructed with two dihedral angles as CG coordinates. We first compute the terms in the GLE from trajectories of the atomistic MD except the fluctuation term. Then, from the computed terms in the GLE, we find the coefficients in the mapped Markovian system in higher dimensions. In Sec. IV, we evaluate the CG model using the two kinetic properties and show agreement of the quantities obtained from the CG model and those from the reference model of atomistic MD. We also discuss the effect of using a nonconstant mass matrix in the CG model.

II. THE MULTIDIMENSIONAL GENERALIZED LANGEVIN EQUATION

A. The generalized Langevin equation from the Mori-Zwanzig formalism

We consider a dynamical system that is described by the following nonlinear ordinary differential equation (ODE):

$$\frac{d\boldsymbol{\phi}}{dt} = \mathbf{R}(\boldsymbol{\phi}(t)), \qquad \boldsymbol{\phi}(0) = \mathbf{x}, \tag{1}$$

where $\boldsymbol{\phi}$, $\mathbf{x} \in \boldsymbol{\Gamma} = \mathbb{R}^n$. Instead of following the entire variables $\boldsymbol{\phi}(t)$, we want to coarse-grain the problem, that is, to follow only CG variables $\mathbf{A}(\boldsymbol{\phi}(t)) \in \mathbb{R}^m\ (m < n)$. CG variables $\mathbf{A}(\boldsymbol{\phi})$ is an m–dimensional vector valued function defined on $\boldsymbol{\Gamma}$ such as a subset of the entire variables $\boldsymbol{\phi}$ or phase variables of $\boldsymbol{\phi}$. We want to build an approximate dynamics that only involves CG variables $\mathbf{A}(t)$ so that by solving this approximate dynamics we can approximate trajectories of CG variables, $\mathbf{A}(\boldsymbol{\phi}(t))$, from the original dynamics, Eq. (1).

The Mori-Zwanzig (MZ) formalism provides a good starting point to build such an approximate dynamics using a projection operator P. A projection operator P maps a function of fine-grained variables $\boldsymbol{\phi}$ to a function of CG variables $\mathbf{A}$. Using the MZ formalism, we can write the time evolution of $\mathbf{A}(\boldsymbol{\phi}(t))$ as follows:[1–3]

$$\begin{aligned}\frac{\partial}{\partial t}\mathbf{A}(\boldsymbol{\phi}(\mathbf{x},t)) = e^{tL}PL\mathbf{A}(\mathbf{x}) + \int_0^t e^{(t-s)L}PLe^{sQL}QL\mathbf{A}(\mathbf{x})ds \\ + e^{tQL}QL\mathbf{A}(\mathbf{x}).\end{aligned} \tag{2}$$

Here, L is a differential operator $L = \sum_{i=1}^n R_i(\mathbf{x})\frac{\partial}{\partial x_i}$, called the Liouville operator. And e^{tL} is an evolution operator associated with the operator L, and $Q = I - P$. The notation $\boldsymbol{\phi}(\mathbf{x}, t)$ is to emphasize an initial condition $\mathbf{x}$ of an arbitrary point in $\boldsymbol{\Gamma}$. For detailed discussion of the MZ formalism, we refer to the existing literature.[3,14,15]

The projection operator P in Eq. (2) can be defined in a few different ways.[3,16] We define P using the conditional expectation

$$(Pf)(\mathbf{A}) \equiv \frac{\int_{\mathbf{x}^*}\delta(\mathbf{A}(\mathbf{x}^*) - \mathbf{A})f(\mathbf{x}^*)\rho(\mathbf{x}^*)d\mathbf{x}^*}{\int_{\mathbf{x}^*}\delta(\mathbf{A}(\mathbf{x}^*) - \mathbf{A})\rho(\mathbf{x}^*)d\mathbf{x}^*} = \langle f\rangle_{\mathbf{A}}^{cond}, \tag{3}$$

where $\mathbf{A}$ indicates $\mathbf{A}(\mathbf{x})$ and $\rho(\mathbf{x})$ is an equilibrium probability density function (pdf). Note that $\mathbf{x}^*$ indicates integration axes, whereas $\mathbf{x}$ indicates a state of the system on the axes.

If the original dynamics (1) is volume conserving, $\nabla \cdot \mathbf{R} = 0$, the second term in Eq. (2) can be written in a more explicit form using the projection operator in Eq. (3),[5,16,17]

$$\frac{\partial}{\partial t}\mathbf{A}(\boldsymbol{\phi}(\mathbf{x},t)) = e^{tL}PL\mathbf{A}(\mathbf{x}) + \int_0^t e^{(t-s)L} \times \left[(\nabla_{\mathbf{A}} - \nabla_{\mathbf{A}}\mathcal{H}) \cdot P_{\mathbf{A}}\left[\mathbf{F}(0)\mathbf{F}^T(s)\right]\right]^T \mathrm{d}s + \mathbf{F}(t). \quad (4)$$

Here, $\mathcal{H}$ is defined as

$$\mathcal{H}(\mathbf{A}) \equiv -\ln \int_{\mathbf{x}^*} \delta(\mathbf{A}(\mathbf{x}^*) - \mathbf{A})\,\rho(\mathbf{x}^*)\,d\mathbf{x}^*, \quad (5)$$

and $\mathbf{F}(\mathbf{x}, t) \equiv e^{tQL}QL\mathbf{A}(\mathbf{x})$; in Eq. (4), we omit the argument $\mathbf{x}$ to simplify the notation. The derivation of the second term in Eq. (4) is given in the Appendix.

When the CG variables $\mathbf{A}$ are a set of generalized coordinates $\boldsymbol{\xi}$ and the corresponding momentums $\mathbf{p}_\xi$, that is $\mathbf{A} \equiv (\boldsymbol{\xi}, \mathbf{p}_\xi)$, Eq. (4) becomes

$$\frac{\mathrm{d}\mathbf{p}_\xi}{\mathrm{d}t} = -\beta^{-1}\frac{\partial \mathcal{H}}{\partial \boldsymbol{\xi}} + \int_0^t e^{(t-s)L} \times \left[\left(\frac{\partial}{\partial \mathbf{p}_\xi} - \beta \mathbf{M}_\xi^{-1}\mathbf{p}_\xi\right) \cdot P_{\mathbf{A}}[\mathbf{F}(0)\,\mathbf{F}^T(s)]\right]^T \mathrm{d}s + \mathbf{F}(t), \quad (6)$$

$$\frac{\mathrm{d}\boldsymbol{\xi}}{\mathrm{d}t} = \mathbf{M}_\xi(\mathbf{q})^{-1}\,\mathbf{p}_\xi. \quad (7)$$

The second equation (7), the equation for the position, includes the generalized mass $\mathbf{M}_\xi(\mathbf{q})$ and is not decomposed with the MZ terms as in Eq. (2). The generalized mass $\mathbf{M}_\xi(\mathbf{q})$ is defined by[18]

$$\mathbf{M}_\xi(\mathbf{q})^{-1} = \sum_k \frac{1}{m_k}\left(\frac{\partial \boldsymbol{\xi}}{\partial q_k}\right)\left(\frac{\partial \boldsymbol{\xi}}{\partial q_k}\right)^T, \quad (8)$$

where m_k is the mass of a microscopic particle with a coordinate q_k. In general, the generalized mass $\mathbf{M}_\xi(\mathbf{q})$ is a function of microscopic coordinates $\mathbf{q}$; the generalized mass should be approximated to a function of CG variables to use Eq. (7) as a CG equation. This approximation will be detailed in Sec. III A 1

Although theoretically appealing, the MZ formalism in the form presented above, Eq. (6), is computationally very expensive since the evaluation of the memory term requires to compute $P_{\mathbf{A}}[\mathrm{F}(0)\ \mathrm{F}^T(s)]$. Instead, we will follow below a different approach. The final form for the equation we are going to use instead of Eq. (6) is as follows:

$$\frac{\mathrm{d}\mathbf{p}_\xi}{\mathrm{d}t} = -\beta^{-1}\frac{\partial \mathcal{H}}{\partial \boldsymbol{\xi}} - \int_0^t \mathbf{K}_L(s)\,\mathbf{M}_\xi^{-1}\mathbf{p}_\xi(t-s)\,\mathrm{d}s + \mathbf{F}'(t). \quad (9)$$

The kernel $\mathbf{K}_L(t)$ is uniquely defined if we impose a certain condition on $\mathbf{F}'(t)$, e.g., $\langle \mathbf{F}'(t)\mathbf{p}_\xi^T(0)\rangle = \mathbf{0}$.

Once we have $\mathbf{K}_L(s)$, the fluctuation term $\mathbf{F}'(t)$ is simply defined as the "remainder" in the equation

$$\mathbf{F}'(t) = \frac{\mathrm{d}\mathbf{p}_\xi}{\mathrm{d}t} + \beta^{-1}\frac{\partial \mathcal{H}}{\partial \boldsymbol{\xi}} + \int_0^t \mathbf{K}_L(s)\,\mathbf{M}_\xi^{-1}\mathbf{p}_\xi(t-s)\,\mathrm{d}s. \quad (10)$$

This approach is exact in the sense that it does not require any assumption aside from the existence of the kernel $\mathbf{K}_L(s)$.

A key aspect of this decomposition is that the effect of the unresolved variables is captured entirely by $\mathbf{F}'(t)$ and this term will have to be modeled in some fashion. For the CG equation to be accurate, the term $\mathbf{F}'(t)$ should contain as "little" information as possible. Later on, we will approximate it using a multivariate Gaussian noise. A poor choice for the form of the decomposition in Eq. (9) leads ultimately to an inaccurate model for $\mathbf{F}'(t)$.

To motivate Eq. (9), we now consider what approximations are required in the MZ formalism to reach an equation of the form Eq. (9). However, it is important to realize that these steps merely serve as a motivation for Eq. (9). With our approach, we do not actually follow the MZ formalism but rather propose a procedure to calculate the memory kernel $\mathbf{K}_L(s)$ and the fluctuating force $\mathbf{F}'(t)$ such that Eq. (9) is satisfied exactly. These two approaches are therefore complementary, but ultimately they are different. The memory kernel $\mathbf{K}_L(t)$ that we will compute in Eq. (9) is distinct from the one obtained through the MZ derivation.

We can get an equation of the form Eq. (9) from the MZ equation (6) if we assume that the autocorrelation $P_{\mathbf{A}}[\mathbf{F}(s)\,\mathbf{F}^T(0)] = \langle \mathbf{F}(s)\,\mathbf{F}^T(0)\rangle_{\mathbf{A}}^{cond}$ does not depend on the CG variables $\mathbf{A}$, which results in $\langle \mathbf{F}(s)\,\mathbf{F}^T(0)\rangle_{\mathbf{A}}^{cond}$ equal to $\langle \mathbf{F}(s)\,\mathbf{F}^T(0)\rangle$. With that, the MZ equation reduces to an equation of the form Eq. (9). The validity of this assumption depends on the system to be coarse grained and the choice of the CG variables $\mathbf{A}$.

As a result of the MZ decomposition in Eq. (2), $\mathbf{F}(t)$ in Eq. (6) satisfies $P\mathbf{F}(t) = \mathbf{0}$; consequently, $\langle \mathbf{F}(t)g(\mathbf{A})\rangle = 0$ for any function of $\mathbf{A}$, $g(\mathbf{A})$. This result now motivates the following approach. Let us require that $\langle \mathbf{F}'(t)\mathbf{p}_\xi^T(0)\rangle = \mathbf{0}$. We show that this equation allows us to uniquely define and compute $\mathbf{K}_L(s)$. Let us right-multiply Eq. (9) with $\mathbf{p}_\xi^T(0)$ and take the expectation. Then, using the property $\langle \mathbf{F}'(t)\mathbf{p}_\xi^T(0)\rangle = \mathbf{0}$, we get

$$\left\langle\left(\frac{\mathrm{d}\mathbf{p}_\xi}{\mathrm{d}t} + \beta^{-1}\frac{\partial \mathcal{H}}{\partial \boldsymbol{\xi}}\right)\mathbf{p}_\xi^T(0)\right\rangle = -\int_0^t \mathbf{K}_L(s)\,\langle \mathbf{M}_\xi^{-1}\mathbf{p}_\xi(t-s)\mathbf{p}_\xi^T(0)\rangle\,\mathrm{d}s. \quad (11)$$

This is a Volterra equation of the first kind for the memory kernel $\mathbf{K}_L(t)$. This equation has a unique solution assuming the "diagonal" term $\langle \mathbf{M}_\xi^{-1}\mathbf{p}_\xi(0)\mathbf{p}_\xi^T(0)\rangle$ is nonsingular. Equation (11) can be solved for $\mathbf{K}_L(t)$ using the recurrence formula, which will be detailed in Sec. III A 3. Note that computing the memory kernel $\mathbf{K}_L(t)$ from Eq. (11) only requires trajectories of the fine-scale dynamics (1).

In Ref. 19, the authors proposed an iterative approach to calculate the memory kernel. The memory kernel was constructed so that the resulting GLE reproduces the force autocorrelation function or velocity autocorrelation function of the reference dynamics.

The memory kernel and the generalized mass $\mathbf{M}_\xi(\mathbf{q})$ in Eq. (7) are both full matrices for generalized coordinates $\boldsymbol{\xi}$. In our model, we will take into account the off-diagonal entries of the memory kernel and the generalized mass. To emphasize this aspect, we call Eqs. (9) and (7) the multidimensional GLE.

The first term in Eq. (9), the mean term, can be calculated by sampling $\mathcal{H}$ from trajectories of the fine-scale dynamics (1). To enhance the sampling of $\mathcal{H}$, the trajectories can be obtained from simulations using biased force[18,20–23] or from other enhanced sampling methods.[24]

The third term in Eq. (9), the fluctuating force $\mathbf{F}'(t)$, depends on fine-grained variables $\mathbf{x}$. So, to use Eq. (9) as a CG equation, $\mathbf{F}'(t)$ should be modeled. In this paper, we want to model $\mathbf{F}'(t)$ as a function of only t. A model of $\mathbf{F}'(t)$ should satisfy the following relation with the memory kernel when the CG equation describes an equilibrium state $\mathbf{K}_L(t) = \beta \langle \mathbf{F}'(t)\, \mathbf{F}'^T(0)\rangle$, which is an instance of the fluctuation dissipation theorem (FDT).

To satisfy the FDT, $\mathbf{K}_L(t) = \beta \langle \mathbf{F}'(t)\, \mathbf{F}'^T(0)\rangle$, a model of $\mathbf{F}'(t)$ should have entries that are correlated with each other since $\mathbf{K}_L(t)$ is a full matrix. To find this model of $\mathbf{F}'(t)$ is nontrivial. So, instead of modeling $\mathbf{F}'(t)$ in Eq. (9) in its non-Markovian form, we first want to map the multidimensional GLE to an approximate Markovian form in Sec. II B. The Markovian form leads to simpler models for $\mathbf{F}'(t)$.

B. Mapping of the GLE to an extended Markovian system

We map Eq. (9) into the following Markovian equation in extended dimensions:[10,15]

$$\begin{pmatrix}\dot{\mathbf{p}}_\xi\\ \dot{\mathbf{s}}\end{pmatrix} = \begin{pmatrix}-\beta^{-1}\frac{\partial\mathcal{H}}{\partial\xi}\\ \mathbf{0}\end{pmatrix} - \underbrace{\begin{pmatrix}\mathbf{0} & \mathbf{A_{ps}}\\ \mathbf{A_{sp}} & \mathbf{A_{ss}}\end{pmatrix}}_{\equiv\mathbf{A}}\begin{pmatrix}\mathbf{M}_\xi^{-1}\mathbf{p}_\xi\\ \mathbf{s}\end{pmatrix} + \underbrace{\begin{pmatrix}\mathbf{0}&\mathbf{0}\\ \mathbf{0}&\mathbf{B_{ss}}\end{pmatrix}}_{\equiv\mathbf{B}}\begin{pmatrix}\mathbf{0}\\ \boldsymbol{\eta}\end{pmatrix}. \tag{12}$$

In this Markovian equation, coupling of $\mathbf{p}_\xi$ to auxiliary variables $\mathbf{s}$ produces non-Markovian dynamics for $\mathbf{p}_\xi$. Using this mapped Markovian system, we avoid calculations of the integration in Eq. (9) and accordingly avoid storage of history of $\mathbf{p}_\xi$. So, using the mapped Markovian equation greatly improves the computational efficiency of the GLE.

It also ease the generation of the fluctuating force $\mathbf{F}'(t)$ in Eq. (9). The third term of the right-hand side of Eq. (12) represents a simple model of $\mathbf{F}'(t)$ in extended dimensions: a linear combination of the white noise gaussian process $\boldsymbol{\eta}$ with $\langle\boldsymbol{\eta}(t)\rangle = \mathbf{0}$ and $\langle \eta_i(t)\, \eta_j(0)\rangle = \delta_{ij}\delta(t)$. The coefficient matrix $\mathbf{B_{ss}}$ can be calculated from a simple relation with $\mathbf{A}$ given by the FDT.

The Markovian equation (12) effectively represents the non-Markovian equation (9), and the relations between the terms in Eqs. (9) and (12) are as follows:[10]

$$\mathbf{K}_L(t) = -\,\mathbf{A_{ps}}e^{-t\mathbf{A_{ss}}}\mathbf{A_{sp}} \qquad (t \ge 0), \tag{13a}$$

$$\mathbf{F}'(t) = -\int_0^t \mathbf{A_{ps}}e^{-(t-t')\mathbf{A_{ss}}}\mathbf{B_{ss}}\boldsymbol{\eta}(t')dt'. \tag{13b}$$

We get (13b) by setting $\mathbf{s}(0) = \mathbf{0}$. Equation (13a) indicates that the Markovian equation (12) can only represent the non-Markovian equation (9) having a memory kernel $\mathbf{K}_L(t)$ that is a combination of exponentially decaying oscillatory functions and exponentially decaying functions. So, for the non-Markovian equation (9) having other forms of the memory kernel $\mathbf{K}_L(t)$, the mapping to the Markovian equation (12) is approximate.

To map Eq. (9) to the Markovian equation (12), we first approximate the computed memory kernel $\mathbf{K}_L(t)$ with a combination of exponentially decaying oscillatory functions and/or exponentially decaying functions. Then, we find matrix $\mathbf{A}$ in Eq. (12) according to Eq. (13a). Previously, this mapping was tried only for a single or uncoupled $\mathbf{p}_\xi$;[8,10] each auxiliary variable s was coupled to one entry of $\mathbf{p}_\xi$. But, for the multidimensional GLE where the memory matrix $\mathbf{K}_L(t)$ is a full matrix, some auxiliary variables should be coupled to multiple entries of $\mathbf{p}_\xi$. It makes the mapping procedure more complicated. Below, we propose a mapping procedure of the multidimensional GLE to the Markovian system (12).

The FDT, $\mathbf{K}_L(t) = \beta \langle \mathbf{F}'(t)\, \mathbf{F}'^T(0)\rangle$, should be satisfied by the mapped Markovian equation (12). It can be shown that the following relation is a sufficient condition for the FDT:[10] $\mathbf{A_{sp}} = -\mathbf{A}_{\mathbf{ps}}^T$ and $\mathbf{B_{ss}}\mathbf{B}_{\mathbf{ss}}^T = \beta^{-1}(\mathbf{A_{ss}} + \mathbf{A}_{\mathbf{ss}}^T)$. So, we need to find $\mathbf{A_{ps}}$ and $\mathbf{A_{ss}}$ then $\mathbf{B_{ss}}$ and $\mathbf{A_{sp}}$ are determined from the FDT.

We introduce another notation $\mathbf{K}_L^M(t)$ as an "effective" memory kernel that is represented by the Markovian equation (12). With the above restriction on $\mathbf{A_{sp}}$, $\mathbf{A_{sp}} = -\mathbf{A}_{\mathbf{ps}}^T$, and $\mathbf{K}_L^M(t)$ becomes

$$\mathbf{K}_L^M(t) = \mathbf{A_{ps}}e^{-t\mathbf{A_{ss}}}\mathbf{A}_{\mathbf{ps}}^T \qquad (t \ge 0). \tag{14}$$

For the mapping, our goal is to find $\mathbf{A_{ps}}$ and $\mathbf{A_{ss}}$ such that $\mathbf{K}_L^M(t)$ closely approximates a given $\mathbf{K}_L(t)$ in Eq. (9) computed in Sec. II A.

We first want to determine $\mathbf{A_{ss}}$. Without loss of generality, we can always assume that $\mathbf{A_{ss}}$ is a block diagonal matrix that consists of 2 × 2 blocks and/or 1 × 1 blocks of scalar entries. A 2 × 2 block results in an exponentially decaying cosine and/or sine component of $\mathbf{K}_L^M(t)$ and a 1 × 1 block of scalar entry results in a pure exponential component of $\mathbf{K}_L^M(t)$.

Note that $\mathbf{K}_L(t)$ is an autocorrelation of a real vector $\mathbf{F}'(t)$ $\mathbf{K}_L(t) = \beta \langle \mathbf{F}'(t+t_0)\, \mathbf{F}'^T(t_0)\rangle$. Let us define $\mathbf{K}_L(t)$ for $t < 0$ using this autocorrelation, which results in $(K_L)_{i,j}(-t) = (K_L)_{j,i}(t)$. This means that when $\mathbf{K}_L(t)$ is expanded with Fourier series, it has symmetric cosine components and skew-symmetric sine components. This property of $\mathbf{K}_L(t)$ motivates the following choice for the diagonal blocks of $\mathbf{A_{ss}}$:

$$\begin{pmatrix} a & b\\ -b & a\end{pmatrix}. \tag{15}$$

With the above choice of $\mathbf{A_{ss}}$, $\mathbf{K}_L^M(t)$ in (14) can be written as follows:

$$\begin{aligned}\mathbf{K}_L^M(t) = \sum_i \Big(& e^{-a_i t}\cos(b_i t)\left(\mathbf{A_{ps}}\right)_i\left(\mathbf{A_{ps}}\right)_i^T\\ &+ e^{-a_i t}\sin(b_i t)\left(\mathbf{A_{ps}}\right)_i \mathbf{S}\left(\mathbf{A_{ps}}\right)_i^T\Big)\\ &+ \sum_i e^{-c_i t}\left(\mathbf{A}'_{\mathbf{ps}}\right)_i\left(\mathbf{A}'_{\mathbf{ps}}\right)_i^T.\end{aligned} \tag{16}$$

Here, $\left(\mathbf{A_{ps}}\right)_i$ denote ith $m \times 2$ block of the $\mathbf{A_{ps}}$ matrix, where m is the size of $\mathbf{p}_\xi$. $\left(\mathbf{A}'_{\mathbf{ps}}\right)_i$ denote a $m \times 1$ vector block of $\mathbf{A_{ps}}$, which shares a column with a 1 × 1 diagonal block of scalar c_i in the $\mathbf{A_{ss}}$ matrix, and

$$\mathbf{S} = \begin{pmatrix} 0 & -1\\ 1 & 0\end{pmatrix}. \tag{17}$$

Note that the coefficient of $e^{-a_i t}\cos(b_i t)$, $\left(\mathbf{A_{ps}}\right)_i\left(\mathbf{A_{ps}}\right)_i^T$, is symmetric, and the coefficient of $e^{-a_i t}\sin(b_i t)$, $\left(\mathbf{A_{ps}}\right)_i \mathbf{S}\left(\mathbf{A_{ps}}\right)_i^T$, is skew-symmetric.

For a given matrix $\mathbf{K}_L(t)$, we expand entries of the matrix using terms as in (16), which are exponential cosines, exponential sines and exponentials. For a certain accuracy, we can aim to find an approximation that minimizes the number of additional degrees of freedom, but it is more complicated to find that approximation. So, we propose a procedure to find an approximation $\mathbf{K}_L^M(t)$ that is less complicated to find but does not minimize the additional degrees of freedom. See the supplementary material for the restrictions on the coefficients of the expansion terms of $\mathbf{K}_L^M(t)$ in (16).

To find an approximation of each entry of $\mathbf{K}_L(t)$, we first do the discrete Fourier transform of $\mathbf{K}_L(t)$; and using the results of it as an initial guess, we use an optimization routine to find expansion coefficients. Then, for the approximation, we find $(\mathbf{A_{ps}})_i$ for each expansion term as in (16). The specific procedure to find an approximation and accordingly the matrix $\mathbf{A_{ps}}$ and $\mathbf{A_{ss}}$ is presented in the supplementary material.

When finding expansion coefficients of symmetric part of $\mathbf{K}_L^M(t)$ using an optimization routine, we propose to use the integration value of $\mathbf{K}_L(t)$ as an additional constraint. By matching the integration value of $\mathbf{K}_L^M(t)$ with that of $\mathbf{K}_L(t)$, we can improve the mapping to the Markovian system. For most applications, the integration value of $\mathbf{K}_L(t)$, $\int_0^T \mathbf{K}_L(t)dt$, $T \gg 0$, is difficult to obtain since it is challenging to accurately compute $\mathbf{K}_L(t)$ for large t. In those cases, we can impose this additional constraint on the integration value of $\mathbf{K}_L^M(t)$ with a free parameter instead of the computed integration value of $\mathbf{K}_L(t)$ from fine-scale simulations. This free parameter for the additional constraint can be set so that the resulting Markovian CG model reproduce a certain quantity of fine-scale simulations.

Using the above procedure, we find $\mathbf{A_{ps}}$ and $\mathbf{A_{ss}}$. As we explained earlier, the FDT gives $\mathbf{A_{sp}}$ and $\mathbf{B_{ss}}$. This completes finding the coefficients of the Markovian system (12).

III. THE MULTIDIMENSIONAL GLE DESCRIPTION OF CONFORMATIONAL MOTION OF THE ALANINE DIPEPTIDE

In this section, we used the proposed CG model to describe the conformational motion of the solvated alanine dipeptide system. In Sec. III A, we present the multidimensional GLE with two dihedral angles of the molecule as CG coordinates. We compute the terms in the multidimensional GLE from trajectories of the atomistic MD simulation of the system. Then, in Sec. III B, we find the approximate Markovian system of the multidimensional GLE following the procedure presented in Sec. II B.

The alanine dipeptide system consists of one alanine dipeptide molecule and 252 water molecules in a 20 Å cubic box. Alanine dipeptide is a small protein with 22 atoms, which was described by the OPLS-AA (Optimized Potential for Liquid Simulations - All Atom) force field, and the water molecules were described by the SPC (Simple Point-Charge) flexible water model. The atomistic MD simulation was carried out using LAMMPS (Large-scale Atomic/Molecular Massively Parallel Simulator);[25] the periodic boundary condition was used with the time step size of 0.5 fs. We carried out *NVT* simulation at T = 300 K with the Langevin thermostat of friction coefficient 1 ps^{-1}. We followed the simulation details from Ref. 26 except we used the single time step of 0.5 fs instead of the multiple time step scheme that they used. Each trajectory was computed for 100 ns long, and the trajectory's initial condition was $(\Phi, \Psi) = (-87.8531°, 161.927°)$, which belongs to one of the two metastable states of the system.

A. The multidimensional GLE with dihedral angles

We chose two dihedral angles of the alanine dipeptide molecule as CG coordinates $\boldsymbol{\xi}$, $\boldsymbol{\xi} = (\Phi, \Psi)$. Then, from Eqs. (9) and (7), the multidimensional GLE model of the alanine dipeptide system is given as follows:

$$\begin{pmatrix} \dot{p}_\Phi \\ \dot{p}_\Psi \end{pmatrix} = \beta^{-1}\begin{pmatrix} -\frac{\partial \mathcal{H}}{\partial \Phi} \\ -\frac{\partial \mathcal{H}}{\partial \Psi} \end{pmatrix} - \int_0^t \mathbf{K}_L(t-s)\,\mathbf{M}_\xi^{-1}\begin{pmatrix} p_\Phi(s) \\ p_\Psi(s) \end{pmatrix} ds + \mathbf{F}'(t), \quad (18)$$

$$\begin{pmatrix} \dot{\Phi} \\ \dot{\Psi} \end{pmatrix} = \mathbf{M}_\xi^{-1}\begin{pmatrix} p_\Phi \\ p_\Psi \end{pmatrix}. \quad (19)$$

Here, p_Φ and p_Ψ are momentums corresponding to Φ and Ψ coordinates, respectively. And $\mathcal{H}$ is defined according to Eq. (5).

1. The coarse-grained mass: A full mass matrix with coordinate dependent mass values

The mass matrix $\mathbf{M}_\xi$ in Eq. (19) is given by Eq. (8). This $\mathbf{M}_\xi$ depends on coordinates of all atoms, $\mathbf{q}$, since the CG coordinates $\boldsymbol{\xi}$ are two dihedral angles, each of which is a nonlinear function of the Cartesian coordinates of the four atoms. We want to approximate the mass matrix $\mathbf{M}_\xi$ as a function of CG coordinates $\boldsymbol{\xi}$ so that Eq. (19) can be evolved only with the information of CG coordinates.

We propose to approximate $\mathbf{M}_\xi$ by the following CG mass:

$$\mathbf{M}_{\mathrm{CG}}(\boldsymbol{\xi})^{-1} \equiv \langle \mathbf{M}_\xi(\mathbf{q})^{-1}\rangle_\xi^{cond} = \left\langle \sum_k \frac{1}{m_k}\left(\frac{\partial \boldsymbol{\xi}}{\partial q_k}\right)\left(\frac{\partial \boldsymbol{\xi}}{\partial q_k}\right)^T \right\rangle_\xi^{cond}. \quad (20)$$

With the CG mass, we neglect the fluctuations in the generalized mass (8). The value $\partial\boldsymbol{\xi}/\partial q_k$ in the above expression (20) is difficult to compute from a standard MD simulation. We use the following equalities to make the CG mass easily calculable from the simulation

$$\mathbf{M}_{\mathrm{CG}}(\boldsymbol{\xi})^{-1} = \left\langle \frac{\partial \boldsymbol{\xi}}{\partial \mathbf{q}}\,\mathbf{M}^{-1}\,\frac{\partial \boldsymbol{\xi}}{\partial \mathbf{q}}^T \right\rangle_\xi^{cond} \quad (21)$$

$$= \left\langle \frac{\partial \boldsymbol{\xi}}{\partial \mathbf{q}}\,(k_B T)^{-1}\,\langle \dot{\mathbf{q}}\dot{\mathbf{q}}^T\rangle\,\frac{\partial \boldsymbol{\xi}}{\partial \mathbf{q}}^T \right\rangle_\xi^{cond} \quad (22)$$

$$= (k_B T)^{-1}\left\langle \frac{\partial \boldsymbol{\xi}}{\partial \mathbf{q}}\,\dot{\mathbf{q}}\dot{\mathbf{q}}^T\,\frac{\partial \boldsymbol{\xi}}{\partial \mathbf{q}}^T \right\rangle_\xi^{cond} \quad (23)$$

$$= (k_B T)^{-1}\langle \dot{\boldsymbol{\xi}}\dot{\boldsymbol{\xi}}^T\rangle_\xi^{cond}. \quad (24)$$

With Eq. (24), the CG mass is calculated from the conditional variance of the time derivative of CG coordinates. See the supplementary material for the derivation. Here, $\mathbf{M}$ is the diagonal matrix with m_k on the diagonal. We call this CG mass $\mathbf{M}_{\mathrm{CG}}(\boldsymbol{\xi})$ the varying mass model.

We evaluated the CG mass from expression (24) using trajectories of the atomistic MD simulation. More specifically, we discretized the domain with 24 × 24 uniform bins; size of each bin was [12° × 12°]. We used 500 ns of trajectories, which is about 10^9 samples with a 0.5 fs time step. To get smooth mass values over the domain, the mass values were calculated from expression (24) for the bins with more than 800 samples, and the mass values in other regions were interpolated using the radial basis function (RBF) interpolation. In particular, inverse multiquadric functions were used as basis functions of the interpolation. Note that the mass matrix is symmetric, and we, therefore, have three unique components of the mass matrix. The contour plots of these three components of the mass matrix are shown in Fig. 1. The variation of the mass over the configuration seems to be non-negligible. For calculation of the mass and other terms in the GLE, we used the following units: ps for time, K for temperature, 10^{-24} g for mass, and pm (10^{-12} m) for length.

We also used the approximate constant mass to compare the CG model with the varying mass with that of the constant mass. The constant CG mass is defined by averaging the mass (24) over all CG coordinates $\boldsymbol{\xi}$,

$$\mathbf{M}_{\text{CG, const}}^{-1} \approx (k_B T)^{-1} \langle \dot{\boldsymbol{\xi}} \dot{\boldsymbol{\xi}}^T \rangle. \tag{25}$$

The constant CG mass was computed using the same trajectories of the atomistic MD that were used for the varying mass calculation. The entries of the constant CG mass matrix were $M_{\Phi,\Phi} = 15.6717$, $M_{\Phi,\Psi} = M_{\Psi,\Phi} = 4.5543$, and $M_{\Psi,\Psi} = 18.8405$. In Ref. 4, the authors described a peptide with the GLE using one curved CG coordinate. In that work, they employed constant mass, which was calculated by averaging the generalized mass (8) over all configurations. Using a varying mass model with the GLE is a novel approach to the authors' knowledge.

2. The mean force term

The mean force term $-\beta^{-1}\,\partial\mathcal{H}/\partial\boldsymbol{\xi}$ in Eq. (18) can be expanded to[18]

$$\begin{aligned}-\beta^{-1}\,\frac{\partial\mathcal{H}}{\partial\boldsymbol{\xi}} &= \left\langle -\frac{\partial H}{\partial\boldsymbol{\xi}}\right\rangle_{\xi,\mathbf{p}_\xi}^{cond} \\ &= -\frac{\mathrm{d}\mathcal{V}}{\mathrm{d}\boldsymbol{\xi}} - \frac{1}{2}\mathbf{p}_\xi^T \frac{\partial\mathbf{M}_{\text{CG}}(\boldsymbol{\xi})^{-1}}{\partial\boldsymbol{\xi}}\mathbf{p}_\xi + \frac{k_B T}{2}\left\langle \frac{\partial \log|\mathbf{M}_\xi^{-1}|}{\partial\boldsymbol{\xi}}\right\rangle_\xi^{cond}.\end{aligned} \tag{26}$$

Here, H denote the Hamiltonian for the reference atomistic system, and $|\cdot|$ denote the determinant of a matrix. And $\mathcal{V}$ is free energy for CG coordinates $\boldsymbol{\xi}$,

$$\mathcal{V}(\boldsymbol{\xi}) = -\beta^{-1} \ln \int_{\mathbf{q}^*,\mathbf{p}^*} \delta(\boldsymbol{\xi}(\mathbf{q}^*) - \boldsymbol{\xi})\, Z^{-1} e^{-\beta H(\mathbf{q}^*,\mathbf{p}^*)}\, \mathrm{d}\mathbf{q}^* \mathrm{d}\mathbf{p}^*. \tag{27}$$

See the supplementary material for the derivation. The mean force term (26) consists of the usual free energy term, $-\mathrm{d}\mathcal{V}/\mathrm{d}\boldsymbol{\xi}$, and the contributions from the generalized mass varying over CG coordinates. When $\mathbf{M}_\xi$ is not constant, from CG coordinates being nonlinear combinations of atomic coordinates, the second and the third terms are nonzero.

We calculated the mean term (26) using the same trajectories of the atomistic MD that we used for calculating the mass. To calculate the first term, $-\mathrm{d}\mathcal{V}/\mathrm{d}\boldsymbol{\xi}$, we discretized the domain with 72 × 72 uniform bins; size of each bin was [5° × 5°]. $\mathcal{V}(\boldsymbol{\xi})$ for each bin was calculated from the histogram $\mathcal{V}(\boldsymbol{\xi}_i) = -\beta^{-1} \ln N_i/N_{\text{tot}}$, where N_i is the number of passes of trajectories for each bin and $N_{\text{tot}} = \sum_{\text{total number of bins}} N_i$. Then, $\mathcal{V}(\boldsymbol{\xi}_i)$ for each bin was interpolated with the RBF interpolation to get a smooth free energy $\mathcal{V}(\boldsymbol{\xi})$ over the whole domain. For the gradient of $\mathcal{V}(\boldsymbol{\xi})$, analytical derivatives of the interpolating basis function were used to obtain smooth $-\mathrm{d}\mathcal{V}/\mathrm{d}\boldsymbol{\xi}$ over the domain.

Figure 2 shows the calculated $\mathcal{V}(\boldsymbol{\xi})$, Ramachandran plot, of the system. Color close to red represent lower free energy region; the two metastable states are labeled with P_{II} and α_R. One metastable state P_{II} is defined as a rectangular region [−110° ≤ Φ ≤ −60°] × [110° ≤ Ψ ≤ 180°], and the other metastable state α_R is defined as a rectangular region [−110° ≤ Φ ≤ −60°] × [−40° ≤ Ψ ≤ 10°].

For the calculation, the third term in Eq. (26) was approximated to $\frac{k_B T}{2}\frac{\partial \log|\mathbf{M}_{\text{CG}}(\xi)^{-1}|}{\partial\xi}$. The second and the third terms were calculated with $\mathbf{M}_{\text{CG}}(\boldsymbol{\xi})$ that was computed in Sec. III A 1. The second and the third terms are nonzero only for the varying mass model; for the varying mass model, the first term was the dominant term, and the second and the third terms were about 10^{-2} or smaller of the magnitude of the first term.

3. The memory matrix

The memory kernel $\mathbf{K}_L(t)$ in Eq. (18) is given by Eq. (11). After discretizing Eq. (11), $\mathbf{K}_L(t)$ at discrete time points can be calculated

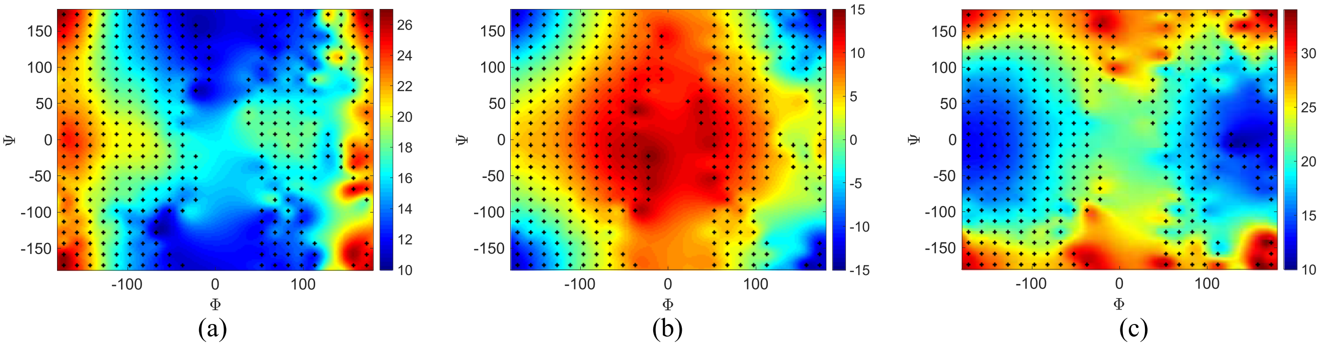

FIG. 1. The coarse-grained mass in the multidimensional GLE. Marked points indicate centers of bins where the mass values were calculated using expression (24); the values in other regions were from the interpolation. (a) $M_{\Phi,\Phi}$, (b) $M_{\Phi,\Psi} = M_{\Psi,\Phi}$, and (c) $M_{\Psi,\Psi}$.

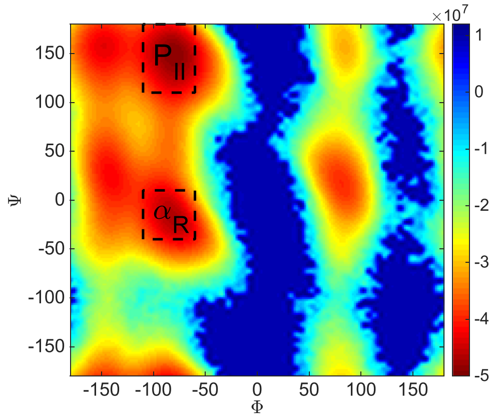

FIG. 2. Free energy $\mathcal{V}$ in Eq. (27) of the alanine dipeptide system. The two labeled regions α_R and P_{II} indicate the two metastable states of the system.

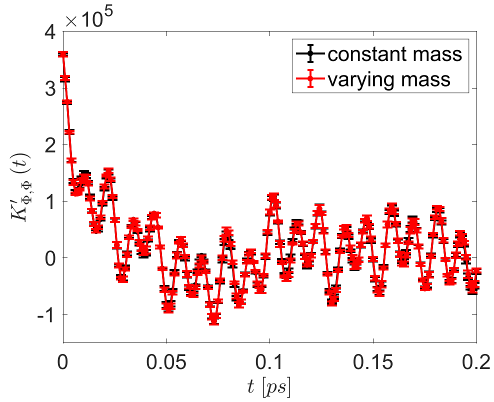

FIG. 3. The multidimensional memory kernel $\mathbf{K}_L(t)$ for two dihedral angles Φ and Ψ of the alanine dipeptide system. [Only $K_{\Phi,\Phi}(t)$ entry is shown.] The memory kernel was calculated both for the varying mass model and the constant mass model.

from the following recurrence formula:

$$\mathbf{K}_L^{(k-1)} = -\left[\langle \mathbf{f}^{(k)} \, (\mathbf{p}_\xi^{(0)})^T \rangle + \sum_{l=0}^{k-2} \mathbf{K}_L^{(l)} \, \mathbf{G}^{(k-l)} \, \Delta t \right] \left(\mathbf{G}^{(1)} \, \Delta t \right)^{-1}. \quad (28)$$

Here, superscripts denote that a function is evaluated at discrete time points; a superscript (k) denotes that a function is evaluated at $t = k\Delta t$. $\mathbf{f}(t) \equiv \mathrm{d}\mathbf{p}_\xi/\mathrm{d}t + \beta^{-1} \, \mathrm{d}\mathcal{H}/\mathrm{d}\boldsymbol{\xi}$ and $\mathbf{G}^{(k)} \equiv \langle \mathbf{M}_\xi{}^{-1} \mathbf{p}_\xi^{(k)} \, (\mathbf{p}_\xi^{(0)})^T \rangle$.

Calculating the memory kernel $\mathbf{K}_L^{(k)}$ from Eq. (28) only requires two ensemble averages $\langle \mathbf{f}^{(k)} \, (\mathbf{p}_\xi^{(0)})^T \rangle$ and $\langle \mathbf{M}_\xi{}^{-1} \mathbf{p}_\xi^{(k)} \, (\mathbf{p}_\xi^{(0)})^T \rangle$. These two ensemble averages were calculated for $0 \le k\Delta t \le 0.25$ ps with $\Delta t = 1.0$ fs. Since we found that those two ensembles obtained using 10 ns trajectories were sufficiently converged, two ensembles were calculated using a 10 ns long trajectory of the atomistic MD. More specifically, we cut a 10 ns trajectory into 40 000 of 0.25 ps segment to calculate those ensembles. We computed the two ensembles from 8 different 10 ns long trajectories. For each calculation of those ensembles, respectively, we calculated $\mathbf{K}_L^{(k)}$. The mean and the standard deviation of $\mathbf{K}_L^{(k)}$ from 8 independent calculations are shown in Fig. 3 for both the constant mass model and the varying mass model. Only one entry of $\mathbf{K}_L^{(k)}$, $K_{\Phi,\Phi}^{(k)}$, is shown in Fig. 3 [see the supplementary material for other entries of $\mathbf{K}_L^{(k)}$]. We observed long-lasting oscillations of a certain frequency (about 12 fs) in the memory kernels $\mathbf{K}_L(t)$. These seem to be due to the stiff bond potentials of the reference atomistic MD system.

As can be seen in Fig. 3, the memory kernel for the varying mass model and the constant mass model show little difference. This seems to be due to characteristics of the system and the way we set up the GLE. Regarding the former, even though the mass varies over the configuration, the last two terms in (26), which are the contribution from the gradient of the mass, were small compared to the first term. Regarding the latter, our memory kernel is already approximated to be independent of $\boldsymbol{\xi}$ in Eq. (9), and so we only used the two ensemble averages, $\langle \mathbf{f}^{(k)} \, (\mathbf{p}_\xi^{(0)})^T \rangle$ and $\langle \mathbf{M}_\xi{}^{-1} \mathbf{p}_\xi^{(k)} \, (\mathbf{p}_\xi^{(0)})^T \rangle$, to compute the memory. These two quantities differ little for the varying mass and the constant mass models since $\mathbf{M}_\xi{}^{-1}$ is only inside $\langle \cdot \rangle$.

In Ref. 27, the authors show that the friction experienced by a solute molecule from solvent depends on the external potentials; as the external potential increases the friction

TABLE I. Coefficients of the approximated memory kernel $\mathbf{K}_L^M(t)$.

		$k = 0$	$k = 1$	$k = 2$	$k = 3$	$k = 4$	Integration values
			The constant mass model				
$K_{\Phi,\Phi}$	p_k or q_k	$1.059 \cdot 10^4$	$1.307 \cdot 10^5$	$4.069 \cdot 10^4$	$1.947 \cdot 10^4$	$6.518 \cdot 10^4$	13, 200
	a_k or c_k	$9.408 \cdot 10^{-1}$	$1.877 \cdot 10^1$	3.294	$7.684 \cdot 10^1$	$3.991 \cdot 10^1$	
	b_k		$3.621 \cdot 10^1$	$3.093 \cdot 10^2$	$7.122 \cdot 10^2$	$1.423 \cdot 10^2$	
			The varying mass model				
$K_{\Phi,\Phi}$	p_k or q_k	$1.1822 \cdot 10^4$	$1.285 \cdot 10^5$	$3.849 \cdot 10^4$	$2.511 \cdot 10^4$	$6.074 \cdot 10^4$	13, 200
	a_k or c_k	1.069	$1.863 \cdot 10^1$	4.986	$9.681 \cdot 10^1$	$4.015 \cdot 10^1$	
	b_k		$3.632 \cdot 10^1$	$3.091 \cdot 10^2$	$6.652 \cdot 10^2$	$1.436 \cdot 10^2$	

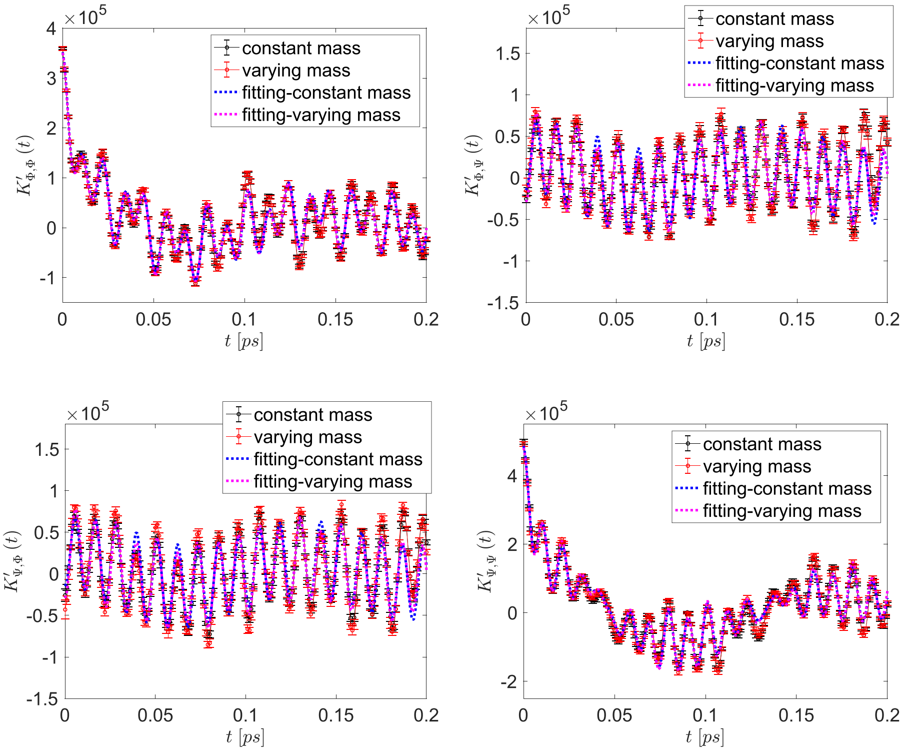

FIG. 4. The approximated memory kernel $\mathbf{K}_L^M(t)$, which is the "effective" memory kernel of the mapped Markovian equation, (denoted as "fitting" in legends) compared to the memory kernel $\mathbf{K}_L(t)$ in the original GLE.

experienced by a solute molecule increases. From this result, the authors speculate that for proteins, a local free-energy minimum would produce a local increase in the friction and conversely, a free-energy barrier would tend to reduce the local friction. This claim implies that the CG model with coordinate dependent friction would be useful. Although, in the scope of this paper, we employ coordinate independent friction model with the memory kernel, employing the coordinate dependent friction model could improve the CG model.

B. Mapping onto an extended Markovian system

We mapped Eq. (18) into an extended Markovian system following the procedure presented in Sec. II B. We ignored the skew-symmetric part of $\mathbf{K}_L(t)$ since the skew-symmetric part was much smaller than the symmetric part. Accordingly, the sine components were not used for the approximation of $\mathbf{K}_L(t)$. Specifically, the off-diagonal term was approximated with a combination of one exponential and 3 exponentially decaying cosines. Then, each modified diagonal term was approximated with a combination of one exponential and 4 exponentially decaying cosines. Using these approximated memory kernel $\mathbf{K}_L^M(t)$, the mapped Markovian equation uses 25 additional degrees of freedom to represent the original non-Markovian equation. When finding those approximated memory kernel $\mathbf{K}_L^M(t)$, we imposed an additional constraint on the integration values of it. Since it is difficult to compute an accurate integration value of the memory kernel $\mathbf{K}_L(t)$ from the atomistic MD simulations, we instead set the integration value of $\mathbf{K}_L^M(t)$ so that the resulting Markovian CG model matches the mean first passage time (MFPT) of the MD system.[28] We computed 50 ns long trajectories of the CG models to estimate the MFPTs of it.

The coefficients of the resulting approximated memory kernel $K_{\Phi,\Phi}(t)$ of $\mathbf{K}_L^M(t)$ are listed in Table I. See the supplementary material for the coefficients of other entries of $\mathbf{K}_L^M(t)$. Figure 4 shows the approximated memory kernel $\mathbf{K}_L^M(t)$ along with $\mathbf{K}_L(t)$ computed in Sec. III A 3. The figure shows that the $\mathbf{K}_L^M(t)$, which is the "effective" memory kernel represented by the mapped Markovian equation (12), is in a good agreement with the memory kernel $\mathbf{K}_L(t)$ in the GLE. The coefficients matrix $\mathbf{A}$ and $\mathbf{B}$ in the mapped Markovian equation (12) can be found from the coefficients of the approximated memory kernel $\mathbf{K}_L^M(t)$.

IV. EVALUATION OF THE CG MODEL AND DISCUSSION

In this section, we evaluate the CG model of the alanine dipeptide system using two important kinetic properties of the system: the velocity autocorrelation (VAC) and the first passage time (FPT) distributions. The VAC and the FPT are key characterizations of the kinetics of any process.[29] We evaluate the CG model with those kinetic properties since it is known that the kinetic properties can only be accurately predicted by including a proper model of "fluctuations" from the unresolved degrees of freedom, whereas the thermodynamic properties can be accurately predicted using the correct mean force term. In our CG model, by including a more sophisticated representation of the "fluctuations" using the memory than Langevin models, we aim to reproduce the kinetic properties of the system as well as thermodynamic properties.

To get the VAC and the FPTs of the CG model, we computed the trajectories of the CG model, Eqs. (12) and (19), using the velocity-Verlet integrator with the time step size $\Delta t = 1$ fs. Specifically, we computed 9 of 100 ns trajectories both for the varying mass model and the constant mass model using the same initial condition of (Φ, Ψ) used in the reference atomistic MD simulations and $\mathbf{s}(0) = \mathbf{0}$. The VAC and FPTs of the atomistic MD were obtained using the equivalent length of trajectories of the atomistic MD.

A. The velocity autocorrelation

Entries of the VAC matrix from the CG model of constant mass and varying mass were compared to those from the atomistic MD in Fig. 5. The VAC from the CG models matches well with that of the atomistic MD for a short time (up to about 0.1 ps). Off-diagonal entries of the VAC matrix were well reproduced with the CG models as well as the diagonal entries. The reproduction of the VAC by the CG models is directly related to the inclusion of the memory kernel in the CG model.[30] Our CG model includes the approximated memory kernel $\mathbf{K}_L^M(t)$, so the approximation of the memory kernel $\mathbf{K}_L(t)$ is directly related to the reproduction of the VAC. The VAC becomes less accurately reproduced for $t > 0.1$ ps since we considered a finite time period of the memory $\mathbf{K}_L(t)$ (up to 0.2 ps) to get the approximated memory kernel $\mathbf{K}_L^M(t)$. The off-diagonal entries of the VAC matrix were reproduced by the CG model since we included off-diagonal entries of the memory kernel. As can be seen in Fig. 5, a (nonoverdamped) Langevin model that does not include the memory kernel was not able to reproduce the VAC of the atomistic MD. The friction coefficients in the Langevin model were the half of the integration values of the memory kernel in Table I; the time step size to integrate the Langevin model was 1 fs. The VAC from the varying mass model and the constant mass model were comparable.

B. The first passage time distributions

We collected the FPTs of the two metastable states P_{II} and α_R from total of 900 ns long trajectories for the CG models and the atomistic MD. The FPT distributions were obtained for each 300 ns long trajectories; the mean and the standard deviation of the FPT distributions from 3 calculations are shown in Fig. 6. Note that we matched the MFPTs of the CG models with those of the atomistic MD by adjusting the integration values of $\mathbf{K}_L^M(t)$ in Sec. III B where we used preliminary 50 ns long trajectories of the CG models to find the integration values of $\mathbf{K}_L^M(t)$. The MFPTs and their ratio using the total of 900 ns long trajectories from the CG models and the atomistic MD are listed in Table II.

Overall, the FPT distributions from the CG models well reproduce those from the atomistic MD. There are relatively large discrepancies for the fast transitions [the first bin in Fig. 6(a) and the first two bins in Fig. 6(b)]. Fast transitions likely highly depend on the fluctuating force. But in our CG models, the fluctuating force is simply modeled as a Gaussian noise and merely matches the autocorrelation of it from the atomistic MD. This seems lead to fewer fast transitions in our CG models than those in the atomistic MD.

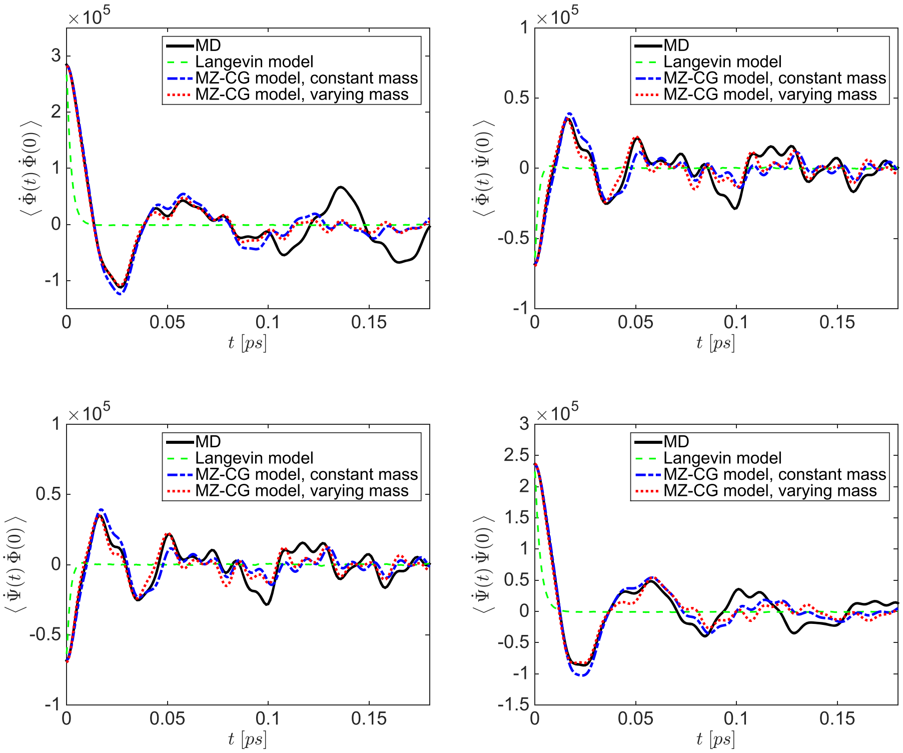

FIG. 5. The velocity autocorrelation (VAC) of two dihedral angles Φ and Ψ calculated from the atomistic MD simulation and from the CG models. Subplots on the diagonal show the autocorrelation for each dihedral angle, and subplots on the off-diagonal show the cross correlation for the two dihedral angles. For the CG models, an approximate Markovian system of the multidimensional GLE with the varying mass model and the constant mass model were shown compared to the Makrovian Langevin model.

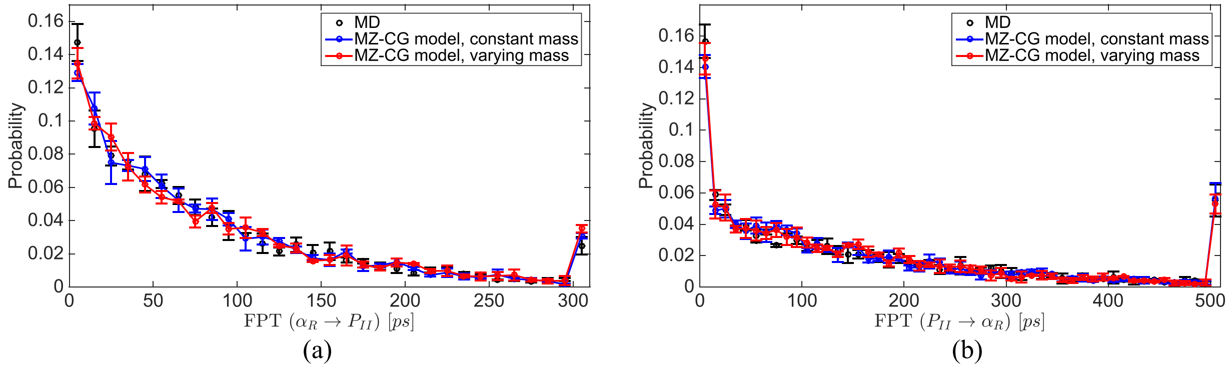

FIG. 6. The first passage time (FPT) distributions of $\alpha_R \rightleftharpoons P_{II}$ from the atomistic MD and the CG models with the varying mass model and the constant mass model. (a) $\alpha_R \to P_{II}$—Transition times greater than 300 ps are binned altogether. (b) $P_{II} \to \alpha_R$—Transition times greater than 500 ps are binned altogether.

In Ref. 31, the authors study the mean first passage times for the barrier crossing of a single massive particle by numerical simulations of the GLE in one dimension. They use single exponential memory kernel in the GLE. The authors show that barrier crossing is accelerated for intermediate memory time, and the barrier crossing is slowed down for the long memory time. We computed the MFPTs from the Makorivan LE with the friction coefficient being the integration values of the memory kernel in the GLE. The GLE model shows about 30%–40% shorter MFPTs compared to the LE model.

Overall, the results from the CG model of the constant mass and the varying mass model differ little for the present alanine dipeptide system. This may be because the chosen two properties for evaluation of the CG model depend on over all configurations. The varying mass model could be more advantageous for reproducing locally dependent kinetic properties. Or the varying mass model might be advantageous when the reference system for coarse-graining is a more complex system and the correction of the mean force term (26) given by the varying mass model is bigger.

We only tested our CG model with a small alanine dipeptide system and used two dihedral angles as CG coordinates. Using this small system allowed easier evaluation of the resulting CG model since it is easier to collect the data for evaluation of the CG models. Although we only used this small system, the application to a bigger peptide chain is a straightforward extension.

TABLE II. The mean first passage times of $\alpha_R \rightleftharpoons P_{II}$ and their ratio from the atomistic MD simulation and the CG models with the varying mass model and the constant mass model.

	$\alpha_R \to P_{II}$ (ps)	$P_{II} \to \alpha_R$ (ps)	Ratio
Atomistic MD	83.43	153.88	1.8445
CG, varying mass	87.43	156.46	1.7896
CG, constant mass	84.51	159.63	1.8888

V. CONCLUSION

In this paper, we modeled the conformational motion of proteins with the multidimensional GLE with the backbone dihedral angles as CG coordinates. Since the forces on a set of two dihedral angles are correlated, we employed the full memory matrix and the full mass matrix in the multidimensional GLE. Since the dihedral angles are nonlinear functions of atomistic Cartesian coordinates, we employed a position-dependent CG mass; for the mean force term, we considered additional terms resulting from the position-dependent mass. Then, we mapped the multidimensional GLE to an extended Markovian system for computational efficiency and to simplify the modeling of the fluctuating force term. Using the solvated alanine dipeptide system of atomistic MD, we showed that the proposed CG model could reproduce the two key kinetic characterizations of the reference atomistic MD system: the VAC and the FPT distributions.

Although we tested our CG model of the multidimensional GLE with the small alanine dipeptide system, our approach can be used for bigger peptides with multiple internal coordinates as CG coordinates. Besides the presented application of protein modeling, the proposed multidimensional GLE approach will be useful for many other applications with multiple CG coordinates. The off-diagonal entries of the memory matrix will be important each time we want to use several CG coordinates for which fluctuating forces from the unresolved degrees of freedom are strongly correlated.

SUPPLEMENTARY MATERIAL

See the supplementary material for all entries of the memory kernel in Fig. 3, additional details of the mapping procedure in Sec. II B, and entire coefficients of the fitted memory kernel in Table I. The supplementary material also contains the derivation of Eqs. (24) and (26).

ACKNOWLEDGMENTS

This work was supported by the Applied Mathematics Program within the Department of Energy (DOE) Office of Advanced

Scientific Computing Research (ASCR) as part of the Collaboratory on Mathematics for Mesoscopic Modeling of Materials (CM4).

APPENDIX: DERIVATION OF THE MEMORY TERM IN EQ. (4)

Let us consider the integrand of the memory term in Eq. (2) $e^{(t-s)L}PL\mathbf{F}(\mathbf{x}, s)$. Using the definition of the projection operator (3),

$$PL\mathbf{F}(\mathbf{x},s) = \frac{\int_{\mathbf{x}^*} L\mathbf{F}(\mathbf{x}^*,s)\delta\big(\mathbf{A}(\mathbf{x}^*)-\mathbf{A}\big)\rho(\mathbf{x}^*)d\mathbf{x}^*}{\int_{\mathbf{x}^*}\delta(\mathbf{A}(\mathbf{x}^*)-\mathbf{A})\rho(\mathbf{x}^*)d\mathbf{x}^*}. \tag{A1}$$

We will write the time argument of $\mathbf{F}(\mathbf{x}, s)$ as a subscript to simplify the notation. Let us focus on the numerator of (A1)

$$\int_{\mathbf{x}^*} L\mathbf{F}_s(\mathbf{x}^*)\delta\big(\mathbf{A}(\mathbf{x}^*)-\mathbf{A}\big)\rho(\mathbf{x}^*)d\mathbf{x}^*, \tag{A2}$$

$$= \int_{\mathbf{x}^*}\sum_i R_i(\mathbf{x}^*)\frac{\partial}{\partial x_i^*}\mathbf{F}_s(\mathbf{x}^*)\delta\big(\mathbf{A}(\mathbf{x}^*)-\mathbf{A}\big)\rho(\mathbf{x}^*)d\mathbf{x}^*, \tag{A3}$$

$$= \sum_i\int_{\mathbf{x}^*}\frac{\partial}{\partial x_i^*}\mathbf{F}_s(\mathbf{x}^*)R_i(\mathbf{x}^*)\delta\big(\mathbf{A}(\mathbf{x}^*)-\mathbf{A}\big)\rho(\mathbf{x}^*)d\mathbf{x}^*, \tag{A4}$$

$$= -\sum_i\int_{\mathbf{x}^*}\mathbf{F}_s(\mathbf{x}^*)\frac{\partial}{\partial x_i^*}\Big(R_i(\mathbf{x}^*)\delta\big(\mathbf{A}(\mathbf{x}^*)-\mathbf{A}\big)\rho(\mathbf{x}^*)\Big)d\mathbf{x}^*, \tag{A5}$$

$$= -\sum_i\int_{\mathbf{x}^*}\mathbf{F}_s(\mathbf{x}^*)\frac{\partial}{\partial x_i^*}\Big(R_i(\mathbf{x}^*)\delta\big(\mathbf{A}(\mathbf{x}^*)-\mathbf{A}\big)\Big)\rho(\mathbf{x}^*)d\mathbf{x}^*, \tag{A6}$$

$$= -\int_{\mathbf{x}^*}\mathbf{F}_s(\mathbf{x}^*)\sum_i\frac{\partial}{\partial x_i^*}\Big(R_i(\mathbf{x}^*)\delta\big(\mathbf{A}(\mathbf{x}^*)-\mathbf{A}\big)\Big)\rho(\mathbf{x}^*)d\mathbf{x}^*, \tag{A7}$$

$$\begin{aligned}= -\int_{\mathbf{x}^*}\mathbf{F}_s(\mathbf{x}^*)\Big(\sum_i\Big\{\frac{\partial}{\partial x_i^*}R_i(\mathbf{x}^*)\Big\}\delta\big(\mathbf{A}(\mathbf{x}^*)-\mathbf{A}\big)\\ + R_i(\mathbf{x}^*)\Big\{\frac{\partial}{\partial x_i^*}\delta\big(\mathbf{A}(\mathbf{x}^*)-\mathbf{A}\big)\Big\}\Big)\rho(\mathbf{x}^*)d\mathbf{x}^*.\end{aligned} \tag{A8}$$

In (A5), we use integration by parts. In (A6), we use $\partial\rho/\partial t = 0$; ρ is an invariant pdf. If $\{\frac{\partial}{\partial x_i^*}R_i(\mathbf{x}^*)\} = \nabla\cdot\mathbf{R} = 0$, which means the reference dynamics (1) is volume conserving, (A8) reduces to

$$= -\int_{\mathbf{x}^*}\mathbf{F}_s(\mathbf{x}^*)\Big(\sum_i R_i(\mathbf{x}^*)\Big\{\frac{\partial}{\partial x_i^*}\delta\big(\mathbf{A}(\mathbf{x}^*)-\mathbf{A}\big)\Big\}\Big)\rho(\mathbf{x}^*)d\mathbf{x}^*, \tag{A9}$$

$$\begin{aligned}= -\int_{\mathbf{x}^*}\mathbf{F}_s(\mathbf{x}^*)\Big(\sum_i R_i(\mathbf{x}^*)\{\nabla_{\mathbf{A}(\mathbf{x}^*)}\delta(\mathbf{A}(\mathbf{x}^*)-\mathbf{A})\}\cdot\frac{\partial\mathbf{A}(\mathbf{x}^*)}{\partial x_i^*}\Big)\\ \times\rho(\mathbf{x}^*)d\mathbf{x}^*,\end{aligned} \tag{A10}$$

$$= \int_{\mathbf{x}^*}\mathbf{F}_s(\mathbf{x}^*)\Big(\sum_i R_i(\mathbf{x}^*)\nabla_{\mathbf{A}}\delta(\mathbf{A}(\mathbf{x}^*)-\mathbf{A})\cdot\frac{\partial\mathbf{A}(\mathbf{x}^*)}{\partial x_i^*}\Big)\rho(\mathbf{x}^*)d\mathbf{x}^*, \tag{A11}$$

$$= \int_{\mathbf{x}^*}\big[\mathbf{F}_s(\mathbf{x}^*)\otimes L\mathbf{A}(\mathbf{x}^*)\big]\nabla_{\mathbf{A}}\delta(\mathbf{A}(\mathbf{x}^*)-\mathbf{A})\rho(\mathbf{x}^*)d\mathbf{x}^*. \tag{A12}$$

In (A11), we use the equality $\nabla_{\mathbf{A}(\mathbf{x}^*)}\delta(\mathbf{A}(\mathbf{x}^*) - \mathbf{A}) = -\nabla_{\mathbf{A}}\delta(\mathbf{A}(\mathbf{x}^*) - \mathbf{A})$. In (A12), we use $\sum_i R_i(\mathbf{x}^*)\frac{\partial}{\partial x_i^*}\mathbf{A}(\mathbf{x}^*) = L\mathbf{A}(\mathbf{x}^*)$, and $\otimes$ denotes an outer product of two vectors.

Then, (A12) becomes

$$\begin{aligned}PL\mathbf{F}_s(\mathbf{x}) = e^{\mathcal{H}}\Big[\nabla_{\mathbf{A}}\cdot\Big(\int_{\mathbf{x}^*}[L\mathbf{A}(\mathbf{x}^*)\otimes\mathbf{F}_s(\mathbf{x}^*)]\Big)\delta(\mathbf{A}(\mathbf{x}^*)-\mathbf{A})\\ \times\rho(\mathbf{x}^*)d\mathbf{x}^*\Big]^T,\end{aligned} \tag{A13}$$

$$= e^{\mathcal{H}}\Big[\nabla_{\mathbf{A}}\cdot(e^{-\mathcal{H}}P[L\mathbf{A}\otimes\mathbf{F}_s])\Big]^T, \tag{A14}$$

$$= \Big[(\nabla_{\mathbf{A}}-\nabla_{\mathbf{A}}\mathcal{H})\cdot P[(L\mathbf{A}-PL\mathbf{A})\otimes\mathbf{F}_s]\Big]^T, \tag{A15}$$

$$= \Big[(\nabla_{\mathbf{A}}-\nabla_{\mathbf{A}}\mathcal{H})\cdot P[\mathbf{F}_0\otimes\mathbf{F}_s]\Big]^T. \tag{A16}$$

In (A15), we use the equality $P\mathbf{F}_s = \mathbf{0}$.

REFERENCES

[1] R. Zwanzig, Phys. Rev. **124**, 983 (1961).

[2] H. Mori, Prog. Theor. Phys. **33**, 423 (1965).

[3] A. J. Chorin, O. H. Hald, and R. Kupferman, Phys. D: Nonlinear Phenom. **166**, 239 (2002).

[4] O. F. Lange and H. Grubmüller, J. Chem. Phys. **124**, 214903 (2006).

[5] C. Hijón, P. Español, E. Vanden-Eijnden, and R. Delgado-Buscalioni, Faraday Discuss. **144**, 301 (2010).

[6] Y. Yoshimoto, I. Kinefuchi, T. Mima, A. Fukushima, T. Tokumasu, and S. Takagi, Phys. Rev. E **88**, 043305 (2013).

[7] Z. Li, X. Bian, X. Li, and G. E. Karniadakis, J. Chem. Phys. **143**, 243128 (2015).

[8] Z. Li, H. S. Lee, E. Darve, and G. E. Karniadakis, J. Chem. Phys. **146**, 014104 (2017).

[9] L. Stella, C. D. Lorenz, and L. Kantorovich, Phys. Rev. B **89**, 134303 (2014).

[10] M. Ceriotti, G. Bussi, and M. Parrinello, J. Chem. Theory Comput. **6**, 1170 (2010).

[11] R. Hegger and G. Stock, J. Chem. Phys. **130**, 034106 (2009).

[12] Y. Mu, P. H. Nguyen, and G. Stock, Proteins: Struct. Funct. Bioinf. **58**, 45 (2004).

[13] A. Davtyan, J. F. Dama, G. A. Voth, and H. C. Andersen, J. Chem. Phys. **142**, 154104 (2015).

[14] S. Nordholm and R. Zwanzig, J. Stat. Phys. **13**, 347 (1975).

[15] R. Zwanzig, *Nonequilibrium Statistical Mechanics* (Oxford University Press, 2001).

[16] S. Izvekov, J. Chem. Phys. **138**, 134106 (2013).

[17] E. Darve, J. Solomon, and A. Kia, Proc. Natl. Acad. Sci. **106**, 10884 (2009).

[18] E. Darve and A. Pohorille, J. Chem. Phys. **115**, 9169 (2001).

[19] G. Jung, M. Hanke, and F. Schmid, J. Chem. Theory Comput. **13**, 2481 (2017); e-print arXiv:1709.07805.

[20] A. Barducci, M. Bonomi, and M. Parrinello, Wiley Interdiscip. Rev. Comput. Mol. Sci. **1**, 826 (2011).

[21] E. Darve, D. Rodríguez-Gómez, and A. Pohorille, J. Chem. Phys. **128**, 144120 (2008).

[22] J. Hénin and C. Chipot, J. Chem. Phys. **121**, 2904 (2004).

[23] J. Kästner, Wiley Interdiscip. Rev. Comput. Mol. Sci. **1**, 932 (2011).

[24] R. C. Bernardi, M. C. Melo, and K. Schulten, Biochim. Biophys. Acta, Gen. Subj. **1850**, 872 (2015).

[25] S. Plimpton, J. Comput. Phys. **117**, 1 (1995).

[26] J. A. Morrone, T. E. Markland, M. Ceriotti, and B. J. Berne, J. Chem. Phys. **134**, 014103 (2011).

[27] J. O. Daldrop, B. G. Kowalik, and R. R. Netz, Phys. Rev. X **7**, 041065 (2017).

[28] For this application, using the MFPTs to determine the integration values of $\mathbf{K}_L^M(t)$ is a natural choice. This is because the integration values of the memory kernel represent the magnitude of friction, which is inversely proportional to diffusion, and the diffusion is inversely related to the MFPTs.

[29] E. Suárez, A. J. Pratt, L. T. Chong, and D. M. Zuckerman, Protein Sci. **25**, 67 (2015).

[30] J. Porrà, K.-G. Wang, and J. Masoliver, Phys. Rev. E **53**, 5872 (1996).

[31] J. Kappler, J. O. Daldrop, F. N. Brünig, M. D. Boehle, and R. R. Netz, J. Chem. Phys. **148**, 014903 (2018).