

Data-driven molecular modeling with the generalized Langevin equation



Francesca Grogan^{a,*}, Huan Lei^{b,c}, Xiantao Li^d, Nathan A. Baker^{a,e}

^a Pacific Northwest National Laboratory, Richland, WA 99352, United States

^b Department of Computational Mathematics, Science and Engineering, Michigan State University, East Lansing, MI 48824, United States

^c Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, United States

^d Department of Mathematics, Pennsylvania State University, State College, PA 16801, United States

^e Division of Applied Mathematics, Brown University, Providence, RI 02912, United States

ARTICLE INFO

Article history:

Received 3 December 2019

Received in revised form 5 May 2020

Accepted 1 June 2020

Available online 3 June 2020

Keywords:

Molecular dynamics

Generalized Langevin equation

Coarse-grained models

Dimension reduction

Data-driven parametrization

ABSTRACT

The complexity of molecular dynamics simulations necessitates dimension reduction and coarse-graining techniques to enable tractable computation. The generalized Langevin equation (GLE) describes coarse-grained dynamics in reduced dimensions. In spite of playing a crucial role in non-equilibrium dynamics, the memory kernel of the GLE is often ignored because it is difficult to characterize and expensive to solve. To address these issues, we construct a data-driven rational approximation to the GLE. Building upon previous work leveraging the GLE to simulate simple systems, we extend these results to more complex molecules, whose many degrees of freedom and complicated dynamics require approximation methods. We demonstrate the effectiveness of our approximation by testing it against exact methods and comparing observables such as autocorrelation and transition rates.

© 2020 Elsevier Inc. All rights reserved.

1. Introduction

Molecular dynamics methods simulate atomic trajectories using Newton's second law of motion. Full atomic-detail molecular dynamics (MD) simulations are often prohibitively expensive due to the complexity and size of the systems under study. Model reduction based on surrogates [1–6] and projection operators [7,8] is a popular approach for reducing dimension and complexity in a wide range of computational science applications. One such model is the generalized Langevin equation (GLE) [7,8], which describes the system in terms of collective degrees of freedom and simulates dynamics in terms of coarse-grained collective variables (CVs). The GLE reduces the problem size by only explicitly representing the dynamics of these CVs; the remaining degrees of freedom are described implicitly. GLE-based approaches have been successfully used in a variety of application areas [9–14]. An important component of the GLE is a time-dependent memory kernel that accounts for the implicit degrees of freedom and their impact on the evolution of the explicitly resolved CVs. This memory term plays a crucial role in non-equilibrium dynamics but is often hard to characterize and evaluate, particularly for high-dimensional systems [15–17]. The kernel is sometimes simplified to reduce computational requirements; however, this often renders the model unable to accurately represent system dynamics [18–20].

* Corresponding author.

E-mail addresses: francesca.grogan@pnnl.gov (F. Grogan), leihuan@msu.edu (H. Lei), xli@math.psu.edu (X. Li), nathan.baker@pnnl.gov (N.A. Baker).

Ideally, construction of the memory kernel should balance computational cost and accuracy. For theoretical convergence analysis and error bounds for various memory kernel approximations, see [21]. Previous work by the authors [22] introduced a data-driven approach to parameterize the GLE memory kernel via a rational approximation in Laplace space. This modeling ansatz—along with the introduction of appropriate auxiliary variables—transforms the GLE into an extended system driven by white noise, where the second fluctuation-dissipation theorem (FDT) [23] can be satisfied by properly choosing the covariance matrix. Numerical studies on simple systems (a tagged particle in solvent) show that this approach can successfully characterize the non-equilibrium dynamics beyond Einstein's Brownian motion theory and accurately predicts observables such as transition rates between a double-well potential. In the data-driven algorithm, modeling accuracy relies on the approximation order of the memory kernel. Data-driven model reduction methods have also been developed by others using a variety of approaches [24–27]. Additionally, Zhu and Venturi have demonstrated the use of polynomial approximations of memory kernels for GLE-like problems [28] as well as a first-principles method for systems with local interactions [29]. In this work, we extend the data-driven parameterization approach [22] to construct a reduced model for the small molecule system of benzyl bromide (BnBr) in explicit water. We recently developed a data-driven approach [5] for uncertainty quantification of the equilibrium properties (e.g., solvation energy) with respect to the non-Gaussian conformation fluctuations using this solvated BnBr system. To quantify the non-equilibrium dynamics, the non-Markovian memory will need to be accurately constructed. In particular, this system is more complex than the benchmark problems considered previously by us [22]: its energy landscape has multiple energetic minima and both the intra- as well as inter-molecular interactions contribute to the energy-dissipation process. On the other hand, the small size of BnBr allows MD simulations to achieve near-ergodic sampling of its conformational space within a tractable amount of time. The transition rate between the two conformational states can be directly evaluated by MD simulation and compared with the predictions from the reduced model. Recently, similar work has been reported by Lee and co-workers [30], where a reduced model of the molecule alanine dipeptide is constructed by the GLE in terms of two dihedral angles and is then parameterized through time-series expansions. In the current study, we present an alternative approach that constructs the memory kernel in Laplace space based on a modification of our earlier approach. These modifications were made to accommodate the more complex gradient system we study in this work, see Section 2.2 for details. An advantage of our approach is that accuracy can be adaptively tuned by adjusting the order of the memory kernel approximation. We demonstrate the applicability of our GLE method on model reduction for molecules in aqueous environments.

The paper is organized as follows. Section 2 introduces the GLE and presents our methodology for constructing the data-driven reduced-order model. We discuss our simulation setup at the end of Section 2. In Section 3, we present results testing our approximation against exact methods and comparing observables such as autocorrelation and transition rates which show the exact memory term is well-modeled by its data-driven parametrization. We briefly conclude and discuss avenues for future work in Section 4.

2. Methods

We begin in Section 2.1 by introducing the GLE and the CVs used. Section 2.2 discusses two approaches to build a rational approximation to the GLE memory kernel. Using extended dynamics to represent the GLE is presented in Section 2.3, with initial and noise conditions detailed in Section 2.4. Finally, we provide simulation setup details in Section 2.5.

2.1. Preliminaries

Before introducing the GLE, we discuss the CVs to be calculated from our BnBr simulations. We perform principal component analysis (PCA) on the BnBr atom positions $\mathbf{x}(t) : [0, \infty) \mapsto \mathbb{R}^N$ and velocities $\dot{\mathbf{x}}(t)$ obtained from an MD trajectory, where N is the number of degrees of freedom in the system (usually $N = 3n - 6$ for n atoms). The covariance matrix $\mathbf{C} \in \mathbb{R}^{N \times N}$ is defined as

$$\mathbf{C} = \left\langle (\mathbf{x}(t) - \langle \mathbf{x}(t) \rangle) (\mathbf{x}(t) - \langle \mathbf{x}(t) \rangle)^T \right\rangle,$$

where $\langle \cdot \rangle$ denotes the ensemble average with respect to the equilibrium distribution of $\mathbf{x}$.

We project the BnBr trajectory onto the principal modes using the eigen-decomposition $\mathbf{C} = \mathbf{V}\mathbf{D}\mathbf{V}^T$ to obtain the principal components $\mathbf{q}(t) : [0, \infty) \mapsto \mathbb{R}^N$

$$\mathbf{q}(t) = \mathbf{V}^T (\mathbf{x}(t) - \langle \mathbf{x}(t) \rangle)$$

and associated velocities $\dot{\mathbf{q}}(t)$. These principal components provide an understanding of the dynamic behavior of BnBr by highlighting the dominant motions of the molecule. In our study, we use the first principal component as our CV; however, it is possible to generalize our method to multi-dimensional as well as nonlinear CVs.

We next introduce the CV mass matrix to be used in the GLE. Generally, for a nonlinear CV, $f(\mathbf{x})$, the mass matrix $\mathbf{M}$ is the diagonal matrix whose elements M_{ii} are given by

$$M_{ii} = \left(\int \sum_{i=1}^n \frac{1}{\mu_i} \left(\frac{\partial f}{\partial x_i} \right)^2 \rho(\mathbf{x}) d\mathbf{x} \right)^{-1},$$

where μ_i is the mass associated with the i -th atom and ρ is the equilibrium probability density function (PDF). In the case of our linear CV, M can simply be defined using the equipartition theorem:

$$M \langle \dot{\mathbf{q}} \dot{\mathbf{q}}^T \rangle = \beta^{-1} \mathbf{I},$$

where $\beta = (k_B T)^{-1}$, k_B is the Boltzmann constant, T is the temperature, and $\mathbf{I}$ is the N -dimensional identity matrix. For details regarding the equivalence of the two formulæ, see Lange and Grubmüller [31]. We note that the mass matrix may also be approximated as a function of the CV, so that the GLE consists of only coarse-grained terms [30].

Given a mass matrix M , the momentum $\mathbf{p}(t) : [0, \infty) \mapsto \mathbb{R}^N$ is defined as

$$\mathbf{p}(t) = M \dot{\mathbf{q}}(t) \quad (1)$$

and the GLE can be written as

$$\dot{\mathbf{p}} = \mathbf{F}(\mathbf{q}) - \int_0^t \mathbf{K}(t - \tau) \dot{\mathbf{q}}(\tau) d\tau + \mathbf{R}(t), \quad (2)$$

where $\mathbf{F}(\mathbf{q}) : [0, \infty) \mapsto \mathbb{R}^N$ is the conservative force, $\mathbf{K}(t) : [0, \infty) \mapsto \mathbb{R}^N \times \mathbb{R}^N$ is the time-dependent memory kernel function, and $\mathbf{R}(t) : [0, \infty) \mapsto \mathbb{R}^N$ is the random noise modeled as a stationary Gaussian process with zero mean that satisfies the second FDT:

$$\langle \mathbf{R}(t) \mathbf{R}(t')^T \rangle = \beta^{-1} \mathbf{K}(t - t').$$

2.2. Constructing a rational approximation to the memory kernel

We define the correlation matrices $\mathbf{G}(t) : [0, \infty) \mapsto \mathbb{R}^N \times \mathbb{R}^N$ and $\mathbf{H}(t) : [0, \infty) \mapsto \mathbb{R}^N \times \mathbb{R}^N$ as

$$\begin{aligned} \mathbf{G}(t) &= \langle \dot{\mathbf{p}}(t) \mathbf{q}(0)^T - \mathbf{F}(\mathbf{q}(t)) \mathbf{q}(0)^T \rangle \\ \mathbf{H}(t) &= \langle \dot{\mathbf{q}}(t) \mathbf{q}(0)^T \rangle. \end{aligned} \quad (3)$$

Right-multiplying by $\mathbf{q}(0)^T$, the GLE (Eq. (2)) becomes

$$\mathbf{G}(t) = - \int_0^t \mathbf{K}(t - \tau) \mathbf{H}(\tau) d\tau \quad (4)$$

with the assumption $\langle \mathbf{R}(t) \mathbf{q}(0)^T \rangle = 0$; see [32] for details.

With $\mathbf{G}(t)$ and $\mathbf{H}(t)$ defined by Eq. (3), we can solve Eq. (4) by transferring this integral equation into frequency space using the Laplace transform [33]:

$$\hat{\mathbf{G}}(\lambda) = \int_0^\infty \mathbf{G}(t) e^{-t/\lambda} dt, \quad \hat{\mathbf{H}}(\lambda) = \int_0^\infty \mathbf{H}(t) e^{-t/\lambda} dt, \quad \hat{\mathbf{K}}(\lambda) = \int_0^\infty \mathbf{K}(t) e^{-t/\lambda} dt, \quad (5)$$

such that Eq. (4) becomes

$$\hat{\mathbf{G}}(\lambda) = -\hat{\mathbf{K}}(\lambda) \hat{\mathbf{H}}(\lambda). \quad (6)$$

Taking $\lambda \rightarrow \infty$ of Eq. (5) gives

$$\hat{\mathbf{G}}(\infty) = \int_0^\infty \mathbf{G}(t) dt, \quad \hat{\mathbf{H}}(\infty) = \int_0^\infty \mathbf{H}(t) dt, \quad \hat{\mathbf{K}}(\infty) = \int_0^\infty \mathbf{K}(t) dt. \quad (7)$$

We note that the definitions of $\mathbf{G}$ and $\mathbf{H}$ differ from our previous work [22], where $\mathbf{H}$ was defined as the velocity correlation matrix, i.e., $\langle \dot{\mathbf{q}}(t) \dot{\mathbf{q}}(0)^T \rangle$. However, our previous choice led to numerical instability in the construction of $\hat{\mathbf{K}}(\lambda)$. In particular, the Markovian limit condition requires $\lim_{\lambda \rightarrow \infty} \hat{\mathbf{G}}(\lambda) = -\lim_{\lambda \rightarrow \infty} \hat{\mathbf{K}}(\lambda) \hat{\mathbf{H}}(\lambda)$. If we choose $\mathbf{H}(t) = \langle \dot{\mathbf{q}}(t) \dot{\mathbf{q}}(0)^T \rangle$, we need to evaluate the term $\lim_{\lambda \rightarrow \infty} \hat{\mathbf{H}}(\lambda) = \int_0^{+\infty} \langle \dot{\mathbf{q}}(t) \dot{\mathbf{q}}(0)^T \rangle dt$. For the gradient system considered in the present work, we note that

$$\begin{aligned}
\lim_{t \rightarrow \infty} \langle \mathbf{q}(t) \dot{\mathbf{q}}(0)^T \rangle &= \lim_{t \rightarrow \infty} \int \rho(\mathbf{q}(t) = \mathbf{q}' | \dot{\mathbf{q}}(0) = \mathbf{v}_0) \rho_0^v(\mathbf{v}_0) \mathbf{q}' \mathbf{v}_0^T d\mathbf{v}_0 \\
&= \int \rho_{\text{eq}}^q(\mathbf{q}') \rho_0^v(\mathbf{v}_0) \mathbf{q}' \mathbf{v}_0^T d\mathbf{v}_0 \\
&\propto \int e^{-\beta U(\mathbf{q}')} e^{-\beta \mathbf{v}_0^T \mathbf{M}^{-1} \mathbf{v}_0 / 2} \mathbf{q}' \mathbf{v}_0^T d\mathbf{v}_0 \equiv 0
\end{aligned}$$

where $\rho(\mathbf{q}(t) = \mathbf{q}' | \dot{\mathbf{q}}(0) = \mathbf{v}_0)$ represents the conditional probability of $\mathbf{q}(t)$ with the initial condition $\dot{\mathbf{q}}(0) = \mathbf{v}_0$ and we take $\rho_0^v(\mathbf{v}_0)$, the probability density function of $\mathbf{v}_0$, to be the equilibrium density. Accordingly, $\lim_{\lambda \rightarrow \infty} \hat{H}(\lambda) = 0$, yielding the ill-conditioning of the Markovian limit of Eq. (6). Using $H(t)$ as defined in Eq. (3) does not result in this ill-conditioning. (We note that for the numerical cases considered in [22], the dynamic equation does not contain the term $U(\mathbf{q})$ and we do not encounter such difficulty).

With $\hat{G}(\lambda)$ and $\hat{H}(\lambda)$ sampled from MD simulations, we construct the memory kernel $\hat{K}(\lambda)$ in the form of

$$\hat{K}(\lambda) \approx \left(\mathbf{I} - \sum_{m=1}^M \mathbf{B}_m \lambda^m \right)^{-1} \left(\sum_{m=1}^M \mathbf{A}_m \lambda^m \right), \quad (8)$$

where the terms of the expression are matrices $\mathbf{A}_m, \mathbf{B}_m \in \mathbb{R}^{N \times N}$. The highest-order coefficients of an M -order expansion can be found through the limit of Eq. (8):

$$\lim_{\lambda \rightarrow \infty} \hat{K}(\lambda) = -\mathbf{B}_M^{-1} \mathbf{A}_M, \quad (9)$$

as $\hat{K}(\infty) = -\hat{G}(\infty) \hat{H}(\infty)^{-1} = -\left(\int_0^\infty \mathbf{G}(t) dt \right) \left(\int_0^\infty \mathbf{H}(t) dt \right)^{-1}$ by taking $\lambda \rightarrow \infty$. Note that $\hat{K}(\infty)$ recovers the friction tensor in Markovian limit, i.e., the Markovian approximation is the zeroth-order GLE approximation,

$$\dot{\mathbf{p}} = \mathbf{F}(\mathbf{q}) - \hat{K}(\infty) \dot{\mathbf{q}}(t) + \mathbf{R}(t), \quad (10)$$

where $\hat{K}(\infty)$ is the friction tensor and is proportional to the diffusion tensor $\mathbf{D}$ by the Einstein relation $\hat{K}(\infty) = k_B T \mathbf{D}^{-1}$.

Eq. (9) allows us to solve for either $\mathbf{A}_M$ or $\mathbf{B}_M$. To solve for the remaining $M - 1$ unknown coefficients, there are two approaches. If high-order derivatives of $H(t)$ and $G(t)$ are available from the data at $t = 0$, then $\hat{K}(\lambda)$ can be (semi-analytically) constructed by the first approach described below. Alternatively, as the numerical evaluation of higher-order terms may introduce significant numerical error in the coefficient calculations, $\hat{K}(\lambda)$ can be numerically constructed by a regression approach using $\hat{G}(\lambda)$ and $\hat{H}(\lambda)$ at interpolation points.

Approach 1. The first approach involves coefficient matching and differentiation. First, we perform a Taylor expansion of $\hat{K}(\lambda)$

$$\hat{K}(\lambda) = \sum_{n=1}^{\infty} \frac{\hat{K}^{(n)}(0)}{n!} \lambda^n. \quad (11)$$

Substituting this expression into the left-hand side of Eq. (8) and matching with respect to λ , we obtain the formula

$$\frac{\hat{K}^{(n)}(0)}{n!} = \mathbf{A}_n + \sum_{l+m=n} \mathbf{B}_l \frac{\hat{K}^{(m)}(0)}{m!}. \quad (12)$$

We can then determine the terms $\hat{K}^{(i)}$ by differentiating Eq. (6). As an example, we compute the first-order coefficients. In this case, we use Eq. (12) to match coefficients with respect to λ^1 :

$$\hat{K}^{(1)}(0) = \mathbf{A}_1, \quad (13)$$

where we have used the fact that $\hat{K}^{(0)}(0) = 0$. To find an expression for the derivative $\hat{K}^{(1)}(0)$, we differentiate Eq. (6) and let $\lambda \rightarrow 0$

$$\begin{aligned}
\hat{G}^{(3)}(0) &= -[\hat{K}(0) \hat{H}^{(3)}(0) + 3\hat{K}^{(1)}(0) \hat{H}^{(2)}(0) + 3\hat{K}^{(2)}(0) \hat{H}^{(1)}(0) + \hat{K}^{(3)}(0) \hat{H}(0)] \\
&= -3\hat{K}^{(1)}(0) \hat{H}^{(2)}(0),
\end{aligned} \quad (14)$$

noting that $\lim_{\lambda \rightarrow 0} \hat{K}(\lambda) = \lim_{\lambda \rightarrow 0} \hat{H}(\lambda) = \lim_{\lambda \rightarrow 0} \hat{H}^{(1)}(\lambda) = 0$ since $\mathbf{q}$ and $\dot{\mathbf{q}}$ are uncorrelated. Integrating Eq. (5) by parts and letting $\lambda \rightarrow 0$ gives [22]

$$\hat{G}^{(i)}(0) = i! \cdot \mathbf{G}^{(i-1)}(0), \quad \hat{H}^{(i)}(0) = i! \cdot \mathbf{H}^{(i-1)}(0), \quad \hat{K}^{(i)}(0) = i! \cdot \mathbf{K}^{(i-1)}(0). \quad (15)$$

Combining Eq. (15) with Eq. (14), we arrive at the following expressions for the first-order coefficients:

$$A_1 = -G^{(2)}(0)[H^{(1)}(0)]^{-1} \quad (16)$$

$$B_1 = -A_1 \hat{K}(\infty)^{-1}. \quad (17)$$

Approach 2. The second approach to solve for unknown coefficients also starts with Eq. (9). However, in this approach, the memory kernel is constructed using regression at discrete values of λ . For an M -order approximation, we choose a set of points $\lambda_1, \lambda_2, \dots, \lambda_{2M-1} \in (0, \infty)$ and solve for the coefficients such that the approximate memory kernel (Eq. (8)) interpolates the exact memory kernel at the chosen set of points. This results in $M - 1$ nonlinear equations. With Eq. (9), these equations comprise a nonlinear system of M equations to be solved, which we can express as

$$\mathbf{F}(\lambda_1, \dots, \lambda_{2M-1}; A_1, \dots, A_M, B_1, \dots, B_M) = 0. \quad (18)$$

Continuing with our first-order example, this approach would result in the following $\mathbf{F}$:

$$\mathbf{F}(\lambda_1; A_1, B_1) = \begin{cases} (I - B_1 \lambda_1)^{-1} (A_1 \lambda_1) - \hat{K}(\lambda_1) \\ B_1 + A_1 \hat{K}(\infty)^{-1} \end{cases} \quad (19)$$

where λ_1 is the user-chosen point and we are solving for unknowns A_1 and B_1 . Any nonlinear solver can be employed to solve Eq. (18); we used the default trust-region algorithm available with MATLAB. Our previous work used the first approach since the correlation matrices G and H were defined differently, allowing access to higher-order derivatives at $t = 0$. In contrast, for the new correlation matrices defined in Eq. (3), high-order information is no longer available. As such, we use the second approach in this current work. We note that the choice of the interpolation points is somewhat *ad hoc*: for the present study, we choose points to capture the peak and asymptotic values of $\hat{K}(\lambda)$. As shown in Fig. 3, we see a pronounced peak in $\hat{K}(\lambda)$ which is indicative of dynamics in the BnBr time domain fluctuating more prominently toward the origin (for more on the relationship between $K(t)$ and $\hat{K}(\lambda)$, see [34]). Thus, we chose points close to $\lambda = 0$ so that we could more accurately capture this behavior. This choice of points does affect the quality of the approximation; an example is shown in Section 3.2.1. The optimal choice of the λ requires further investigation.

2.3. Representing the GLE with extended dynamics driven by white noise

Once we have determined coefficients of the rational memory term, we can construct a new approximate GLE system [22]. We illustrate this construction by deriving the extended system in the first-order case. We know from Eq. (8) that the first-order rational approximation for $\hat{K}$ is given by

$$\hat{K}(\lambda) \approx (I - B_1 \lambda)^{-1} (A_1 \lambda).$$

Taking the inverse Laplace transform, $\mathcal{L}^{-1}$, we obtain

$$K(t) = \mathcal{L}^{-1} \{ \hat{K}(\lambda) \} \approx A_1 e^{B_1 t}.$$

Let us define the auxiliary variable $\mathbf{d}(t) : [0, \infty) \mapsto \mathbb{R}^{MN}$, where M is the order of the rational approximation, as

$$\mathbf{d}(t) = - \int_0^t K(t - \tau) \dot{\mathbf{q}}(\tau) d\tau + \mathbf{R}(t).$$

With $M = 1$ for this derivation, we next define $\mathbf{d}_1(t)$, the auxiliary variable in the first-order case, as

$$\mathbf{d}_1(t) = - \int_0^t A_1 e^{B_1(t-\tau)} \dot{\mathbf{q}}(\tau) d\tau + \mathbf{R}(t). \quad (20)$$

Using the Leibniz integral rule, we can differentiate $\mathbf{d}_1(t)$:

$$\dot{\mathbf{d}}_1(t) = -A_1 \dot{\mathbf{q}}(t) - B_1 \int_0^t A_1 e^{B_1(t-\tau)} \dot{\mathbf{q}}(\tau) d\tau + \dot{\mathbf{R}}(t).$$

Note that $\mathbf{R}(t)$ is assumed to be colored noise and not white noise, and as such is differentiable. As discussed in Section 2.4, $\mathbf{R}(t)$ obeys the FDT. Deferring the details to Section 2.4, the detailed colored noise can be expressed in terms of the initial condition $\mathbf{d}_1(0)$ and a simple white noise term $\mathbf{W}_1(t)$:

$$\mathbf{R}(t) = \int_0^t e^{B_1(t-\tau)} \mathbf{W}_1(\tau) d\tau + e^{B_1 t} \mathbf{d}_1(0),$$

which is further discussed in the next section. We remark that B_1 is negative and the first term of $\mathbf{R}(t)$ is well-behaved for large t . Using the Leibniz integral rule again, we can write $\dot{\mathbf{d}}_1(t)$:

$$\begin{aligned}\dot{\mathbf{d}}_1(t) = & -A_1\dot{\mathbf{q}}(t) - B_1 \int_0^t A_1 e^{B_1(t-\tau)} \dot{\mathbf{q}}(\tau) d\tau \\ & + \mathbf{W}_1(t) + B_1 \int_0^t e^{B_1(t-\tau)} \mathbf{W}_1(\tau) d\tau + B_1 e^{B_1 t} \mathbf{d}_1(0).\end{aligned}\quad (21)$$

Note that

$$\begin{aligned}B_1 \int_0^t e^{B_1(t-\tau)} \mathbf{W}_1(\tau) d\tau + B_1 e^{B_1 t} \mathbf{d}_1(0) &= B_1 \mathbf{R}(t) \\ B_1 \int_0^t A_1 e^{B_1(t-\tau)} \dot{\mathbf{q}}(\tau) d\tau &= B_1 \mathbf{d}_1(t) - B_1 \mathbf{R}(t)\end{aligned}$$

and so Eq. (21) can be written as

$$\dot{\mathbf{d}}_1(t) = B_1 \mathbf{d}_1(t) - A_1 \dot{\mathbf{q}}(t) + \mathbf{W}_1(t)$$

to obtain the first-order approximate GLE system:

$$\begin{aligned}\dot{\mathbf{q}} &= \mathbf{M}^{-1} \mathbf{p}, \\ \dot{\mathbf{p}} &= \mathbf{F}(\mathbf{q}) + \mathbf{d}_1, \\ \dot{\mathbf{d}}_1 &= B_1 \mathbf{d}_1 - A_1 \dot{\mathbf{q}} + \mathbf{W}_1.\end{aligned}\quad (22)$$

Higher-order approximations are obtained by generalizing the procedure above to obtain

$$\begin{aligned}\dot{\mathbf{q}} &= \mathbf{M}^{-1} \mathbf{p}, \\ \dot{\mathbf{p}} &= \mathbf{F}(\mathbf{q}) + \mathbf{Z}^T \mathbf{d}_M, \\ \dot{\mathbf{d}}_M &= \mathbf{B} \mathbf{d}_M - \mathbf{Q} \mathbf{Z} \dot{\mathbf{q}} + \mathbf{W}_M,\end{aligned}\quad (23)$$

where the symmetric positive definite matrix $\mathbf{Q} \in \mathbb{R}^{MN \times MN}$ and the matrix $\mathbf{Z} \in \mathbb{R}^{MN \times N}$ are determined by matching Eq. (8) with the equation $\mathbf{K}(t) = \mathbf{Z}^T e^{\mathbf{B}t} \mathbf{Q} \mathbf{Z}$. $\mathbf{Q}$ is the covariance matrix of the auxiliary vector $\mathbf{d}_M$ under equilibrium. The matrix $\mathbf{B}$ is dependent on order; e.g., a fourth-order approximation would have the form

$$\mathbf{B} = \begin{pmatrix} 0 & 0 & 0 & B_4 \\ 0 & 0 & 0 & B_3 \\ 0 & 0 & 0 & B_2 \\ 0 & 0 & 0 & B_1 \end{pmatrix}.$$

2.4. Initial and noise conditions to satisfy the second FDT

Recall that $\mathbf{R}(t)$ in Eq. (2) simulates system noise as a colored noise that must satisfy the second FDT. Through our extended GLE system, we can replace $\mathbf{R}(t)$ with a simpler white noise term $\mathbf{W}(t)$ and choose the initial and noise conditions for $\mathbf{W}(t)$ and $\mathbf{d}(t)$ to ensure that the colored noise generated by these extended dynamics also satisfies the second FDT [22]. For the first-order approximation, the initial and noise conditions are

$$\begin{aligned}\langle \mathbf{d}_1(0) \mathbf{d}_1(0)^T \rangle &= \beta^{-1} A_1 \\ \langle \mathbf{W}_1(t) \mathbf{W}_1(t')^T \rangle &= -\beta^{-1} (B_1 A_1 + A_1 B_1^T) \delta(t - t'),\end{aligned}\quad (24)$$

and for higher-order approximations, the initial and noise conditions are

$$\begin{aligned}\langle \mathbf{d}_M(0) \mathbf{d}_M(0)^T \rangle &= \beta^{-1} \mathbf{Q} \\ \langle \mathbf{W}_M(t) \mathbf{W}_M(t')^T \rangle &= -\beta^{-1} (\mathbf{B} \mathbf{Q} + \mathbf{Q} \mathbf{B}^T) \delta(t - t')\end{aligned}\quad (25)$$

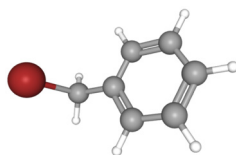


Fig. 1. The benzyl bromide (BnBr) molecule with carbon atoms shown in gray, hydrogen in white, and bromine in red. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

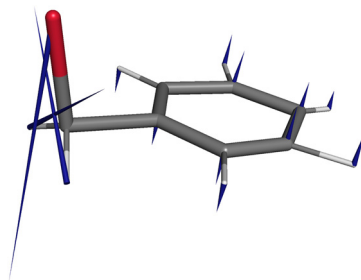


Fig. 2. Porcupine plot of first eigenvector showing dominant motion of benzyl bromide (BnBr).

For other work on GLEs for systems driven by white noise, see [35] and [36]. The extended GLE system approximations eliminate costly integration of the exact memory term, which depends on the system history, by replacing it with an extended system of stochastic differential equations. The accuracy of this approximation improves with increasing order which involves reformulation of the matrix B and recalculation of the matrices Q and Z . These computations are relatively simple to perform for the low-order approximations needed to model small molecular systems (see Section 3 of this manuscript). Overall, our method provides substantial dimension-reduction and results in significant increase in computational tractability.

2.5. Simulation setup

Simulations were run using GROMACS [37] and the general AMBER force field [38]. We performed 360 simulations of a single BnBr molecule (Fig. 1), which is comprised of 15 atoms, in a solvent consisting of 1011 water molecules in a $(3.14216 \text{ nm})^3$ domain. We used a constant number-volume-temperature (NVT) ensemble with a Nosé-Hoover thermostat [39] at 300 K. Each simulation ran for 10 ns with a time step of 2 fs. The particle-mesh Ewald method [40] was used for long-range electrostatics. All BnBr bond lengths were constrained using the LINCS algorithm [41]; we note that this significantly reduces the dimension of the molecular conformational space. The BnBr positions were stored at every time step.

As a post-processing step, translational and rotational degrees of freedom were removed from the trajectory using the GROMACS function *trajconv*. Averaging was then done over the 360 total trajectories. We performed PCA on these trajectories and checked for convergence by splitting the post-processed trajectories into equal halves and calculating the PDFs of each half for the first few principal components. These PDFs of both halves matched well with each other, indicating the simulation had converged. As the first principal component accounted for 63% of the observed variance, we found it sufficient to use this as our single CV for the purposes of illustration in this paper. To physically interpret and visualize the results of PCA, we can generate a porcupine plot showing the motion along an eigenvector. In particular, the porcupine plot of the first eigenvector, Fig. 2, shows dominant motions of BnBr, with the direction and length of each “quill” showing the direction and magnitude of motion, respectively. In particular, we can see the bromomethyl group contributes prominently to the motion of the first mode.

From this, we constructed the correlation matrices and solved for the unknown rational coefficients for zeroth- to fourth-order approximations as described above. Note that a zero-order approximation is simply a Markovian approximation, with the integral term in Eq. (2) simplifying to $\hat{K}(\infty)\dot{\mathbf{q}}(t)$; see Ma et al. for details [42].

3. Results

The following section presents results testing our approximation against exact methods. In particular, we assess comparisons using the memory kernel, position autocorrelation, velocity autocorrelation, and mean first-passage time.

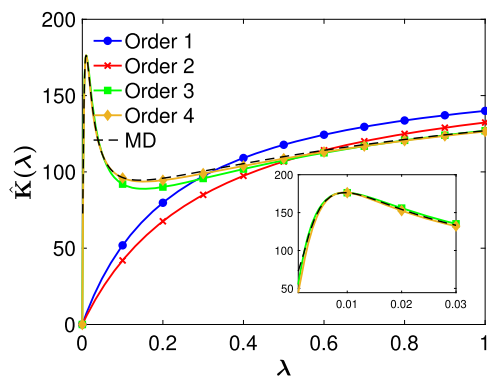


Fig. 3. Memory kernel in Laplace space from MD simulation versus kernels constructed using data-driven GLE approximations of varying orders. Inset: Close-up of third- and fourth-order approximations capturing the pronounced peak of $\hat{K}(\lambda)$.

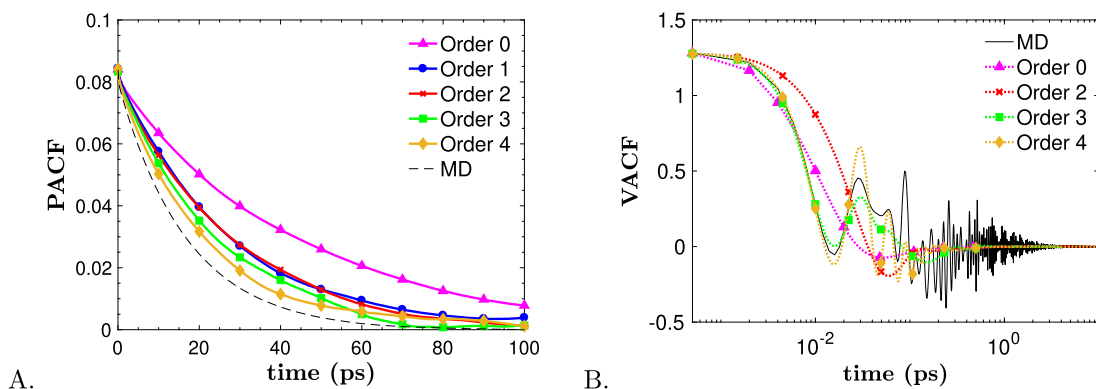


Fig. 4. PACF (A) and VACF (B) for exact MD data compared to approximate GLE simulations.

3.1. Memory kernel

The PDF $\rho(\mathbf{q})$ of the CV defined in Section 2.1 can be calculated using kernel density estimation on samples from the MD trajectory. This PDF can be used to calculate the free energy $U(\mathbf{q})$

$$U(\mathbf{q}) = -\beta^{-1} \ln(\rho(\mathbf{q})) \quad (26)$$

which, in turn, can be used to calculate the mean force

$$\mathbf{F}(\mathbf{q}) = -\nabla U(\mathbf{q}). \quad (27)$$

With $\mathbf{F}(\mathbf{q})$, we are able to sample $G(t)$ and $H(t)$ and construct the Laplace transform of the memory term $\hat{K}(\lambda)$ based on the numerical approach introduced in Section 2.2.

The exact $\hat{K}(\lambda)$ calculated from our simulations shows a pronounced peak near $\lambda = 0.01$. Since this peak is indicative of oscillations in $K(t)$ (i.e. oscillations back in the time domain), good approximation of this peak is important for capturing system dynamics. As shown in Fig. 3, the first- and second-order approximations do not reproduce the peak; however, the third- and fourth-order rational functions have enough interpolation points for an accurate model.

3.2. Autocorrelation functions

The PDF tests the equilibrium properties of the approximation. To test dynamic properties, we computed both the position autocorrelation function (PACF) $\langle \mathbf{q}(t)\mathbf{q}(0)^T \rangle$ as well as the velocity autocorrelation function (VACF) $\langle \dot{\mathbf{q}}(t)\dot{\mathbf{q}}(0)^T \rangle$ and compared the resulting approximate trajectories to data calculated directly from the original MD simulation. The results are shown in Fig. 4. The accuracy of the PACF increases with increasing order of the GLE approximation, with all performing better than the zero-order Markovian approximation. Likewise, the accuracy of the VACF also increases with increasing order of the GLE approximation. Long-time oscillations occur in the VACF of BnBr; we found such behavior is due to the following:

- the strong intramolecular covalent bond interactions and the dominant motions of BnBr.

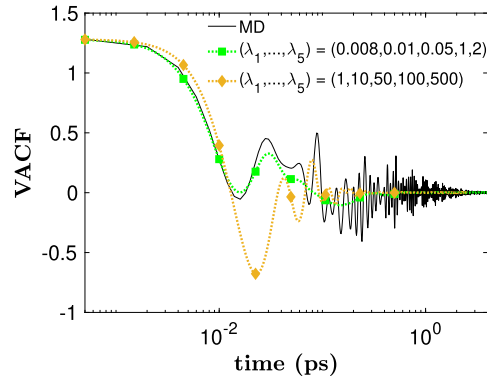


Fig. 5. Third-order VACF approximation using two different sets of interpolation points λ_i , $i = 1, \dots, 5$.

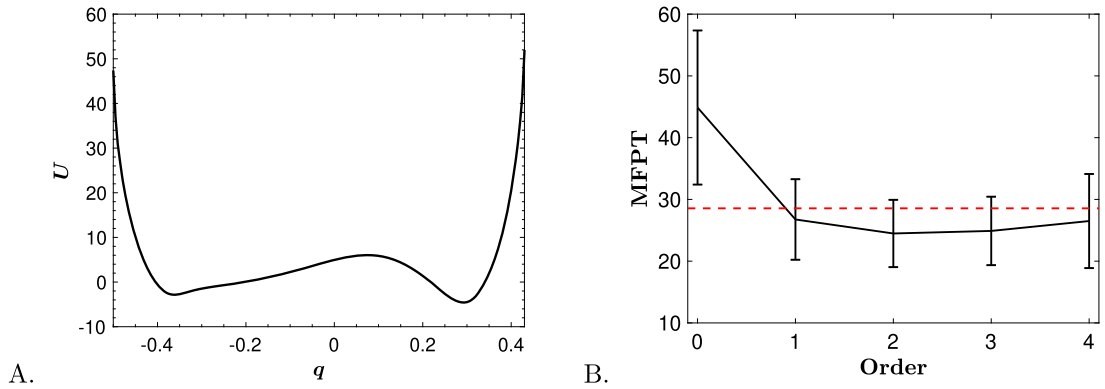


Fig. 6. Mean first-passage time between states in a double-well potential of mean force. A. Double-well potential $U(\mathbf{q})$ calculated from the PDF $\rho(\mathbf{q})$. B. MFPT (ps) from the approximate GLE compared to the exact MD data (red dotted line), shown with a 95% confidence interval.

- the number of peaks in the FFT of the velocity autocorrelation is of comparable order to the number of peaks in the vibrational spectrum of BnBr.

These oscillations make it particularly challenging to approximate; inaccuracies in these autocorrelations may lead to misinterpretation of the underlying nature of the system dynamics. The third- and fourth-order approximations reproduce both the PACF and VACF fairly well. Recall that we applied LINCS constraints to all BnBr bond lengths, which reduced the amount of noise in the VACF of the principal components, and likely allowed for easier approximation of the GLE terms.

3.2.1. Selecting interpolation points

Recall that we construct an order- M rational memory term through regression with user-selected values $\lambda_1, \lambda_2, \dots, \lambda_{2M-1} \in (0, \infty)$. Fig. 5 compares the VACFs of two third-order approximations constructed using two different sets of interpolation points. We see that the approximation using shorter-time interpolation points more accurately reproduces the VACF. As shown in Fig. 3, this increase in accuracy is due to the regression model sufficiently capturing the pronounced peak in $\hat{K}(\lambda)$, which occurs close to $\lambda = 0$.

3.3. Mean first-passage time

Predicting non-equilibrium properties such as mean first-passage time (MFPT) between states is a challenging test for the GLE approximation since this statistic from the original BnBr MD simulations was not known or used *a priori* in the data-driven parametrization of the GLE. From the density ρ , we can calculate the potential of our first principal component $U(\mathbf{q}) = -\beta^{-1} \ln(\rho)$ which has two wells, as shown in Fig. 6A. Denoting the left potential well as state “A” and the right as state “B”, we define the MFPT as the mean time for a particle starting at an initial state to cross the peak maximum into the other state. In this example, this maximum occurs at $q = 0.075$; thus state A is defined as $q < 0.075$ and state B is defined as $q > 0.075$. Fig. 6B shows a comparison for all orders of the MFPT and the exact MD trajectory. The Markovian approximation fails to accurately reproduce the MFPT, while the higher-order GLE approximations show significantly better agreement with the MD results.

4. Conclusion

As full MD simulations often require intensive computational resources and long run times to achieve ergodicity, researchers have increasingly relied on reduced-order modeling for simulation. In particular, the GLE has seen a resurgence in popularity, as it provides a convenient description of coarse-grained dynamics. While the exact GLE can significantly reduce problem size and difficulty, the memory kernel of the GLE relies on past-system history and is often hard to characterize and compute. To mitigate this, previous work introduced a data-driven approximation to the GLE. While there is cost associated with sampling from the exact system, it can be computationally cheaper than running the full MD simulation for very long times. Directly sampling from correlation functions of exact system dynamics, we replaced the memory kernel with a rational approximation and carefully introduced an auxiliary variable and white-noise term to convert the GLE into an extended system that does not rely on past-system history. Additionally, accuracy is adaptively affected by the chosen order of the rational approximation. This current work extends a data-driven approximation of the GLE [22] to more complex and realistic molecules. Using BnBr as our test case, our comparison of exact MD simulation against the approximation shows observables are reproduced well using relatively low orders for the rational term.

There are multiple avenues for future work that further develops modeling capability of complex systems. While we were able to represent BnBr system dynamics with a single CV, accurate construction of the memory term and reduced dynamics remains a challenging task as the complexity of the system increases, since the condition number of B may become large. Thus, it would be necessary to test for robustness on systems where the CV dimension is higher than one, as is done in [30]. Towards this end, to alleviate such difficulties, we are developing a regularization approach for the GLE approximation by formulating the memory kernel construction as an optimization problem. Furthermore, while we were able to use an unbiased density to compute the force term in this current work, it is more difficult to compute this term with respect to higher-dimensional spaces. To ensure adequate sampling of the energy surface, enhanced sampling methods may need to be paired with the data-driven GLE approximation in order to give robust results.

Acronyms

ASCR	Advanced Scientific Computing Research
BnBr	benzyl bromide
CV	collective variable
DOE	Department of Energy
FDT	fluctuation-dissipation theorem
GLE	generalized Langevin equation
MFPT	mean first-passage time
MD	molecular dynamics
NVT	constant number-volume-temperature
PACF	position autocorrelation function
PCA	principal component analysis
PDF	probability density function
VACF	velocity autocorrelation function

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Peiyuan Gao for helpful discussions. This work was performed using resources through Research Computing at Pacific Northwest National Laboratory. FG acknowledges support from the Department of Energy (DOE) Office of Advanced Scientific Computing Research (ASCR) through the ASCR Distinguished Computational Mathematics Postdoc Project under ASCR Project 71268. HL and NAB acknowledge support from NIH grant GM069702.

References

- [1] Anna V. Gubskaya, Vladyslav Kholodovych, Doyle Knight, Joachim Kohn, William J. Welsh, Prediction of fibrinogen adsorption for biodegradable polymers: integration of molecular dynamics and surrogate modeling, *Polymer* 48 (19) (2007) 5788–5801, <https://doi.org/10.1016/j.polymer.2007.07.007>.
- [2] James Daniel Nance, *Investigating Molecular Dynamics with Sparse Grid Surrogate Models*, PhD thesis, North Carolina State University, 2015.
- [3] H. Lei, X. Yang, B. Zheng, G. Lin, N.A. Baker, Constructing surrogate models of complex systems with enhanced sparsity: quantifying the influence of conformational uncertainty in biomolecular solvation, *SIAM Multiscale Model. Simul.* 13 (4) (2015) 1327–1353.
- [4] M. Razi, A. Narayan, R.M. Kirby, D. Bedrov, Fast predictive models based on multi-fidelity sampling of properties in molecular dynamics simulations, *Comput. Mater. Sci.* 152 (2018) 125–133, <https://doi.org/10.1016/j.commatsci.2018.05.029>.
- [5] H. Lei, J. Li, P. Gao, P. Stinis, N.A. Baker, A data-driven framework for sparsity-enhanced surrogates with arbitrary mutually dependent randomness, *Comput. Methods Appl. Mech. Eng.* (ISSN 0045-7825) 350 (2019) 199–227.

- [6] Pei Zhao, Song Han, Xiaoxia Li, Tong Zhu, Xiaofang Tao, Li Guo, Comparison of RP-3 pyrolysis reactions between surrogates and 45-component model by ReaxFF molecular dynamics simulations, *Energy Fuels* 33 (8) (2019) 7176–7187, <https://doi.org/10.1021/acs.energyfuels.9b01321>.
- [7] Hazime Mori, Transport collective motion, and Brownian motion, *Prog. Theor. Phys.* 33 (3) (1965) 423–455, <https://doi.org/10.1143/ptp.33.423>.
- [8] Robert Zwanzig, Nonlinear generalized Langevin equations, *J. Stat. Phys.* 9 (3) (1973) 215–220, <https://doi.org/10.1007/bf01008729>.
- [9] S.A. Adelman, Generalized Langevin equations and many-body problems in chemical dynamics, in: *Advances in Chemical Physics*, John Wiley & Sons, Inc., 2007, pp. 143–253.
- [10] Pierre Turq, Frédéric Lantelme, Harold L. Friedman, Brownian dynamics: its application to ionic solutions, *J. Chem. Phys.* 66 (7) (1977) 3039–3044, <https://doi.org/10.1063/1.434317>.
- [11] Andrés Córdoba, Tsutomu Indei, Jay D. Schieber, Elimination of inertia from a generalized Langevin equation: applications to microbead rheology modeling and data analysis, *J. Rheol.* 56 (1) (2012) 185–212, <https://doi.org/10.1122/1.3675625>.
- [12] Vincent Démery, Olivier Bénichou, Hugo Jacquin, Generalized Langevin equations for a driven tracer in dense soft colloids: construction and applications, *New J. Phys.* 16 (5) (2014) 053032, <https://doi.org/10.1088/1367-2630/16/5/053032>.
- [13] Yu-Wen Wu, Hsiu-Yu Yu, Adhesion of a polymer-grafted nanoparticle to cells explored using generalized Langevin dynamics, *Soft Matter* 14 (48) (2018) 9910–9922, <https://doi.org/10.1039/c8sm01579a>.
- [14] Gil Ariel, Eric Vanden-Eijnden, Testing transition state theory on Kac-Zwanzig model, *J. Stat. Phys.* 126 (1) (2007) 43–73.
- [15] E. Darve, J. Solomon, A. Kia, Computing generalized Langevin equations and generalized Fokker-Planck equations, *Proc. Natl. Acad. Sci.* 106 (27) (2009) 10884–10889, <https://doi.org/10.1073/pnas.0902633106>.
- [16] Xiantao Li, A coarse-grained molecular dynamics model for crystalline solids, *Int. J. Numer. Methods Eng.* 83 (8–9) (2010) 986–997, <https://doi.org/10.1002/nme.2892>.
- [17] Gerhard Jung, Martin Hanke, Friederike Schmid, Iterative reconstruction of memory kernels, *J. Chem. Theory Comput.* 13 (6) (2017) 2481–2488, <https://doi.org/10.1021/acs.jctc.7b00274>.
- [18] E. Guàrdia, J.A. Padró, Generalized Langevin dynamics simulation of interacting particles, *J. Chem. Phys.* 83 (4) (1985) 1917–1920, <https://doi.org/10.1063/1.449379>.
- [19] Huan Lei, Bruce Caswell, George Em Karniadakis, Direct construction of mesoscopic models from microscopic simulations, *Phys. Rev. E* 81 (2) (2010), <https://doi.org/10.1103/physreve.81.026704>.
- [20] Mary Shugard, John C. Tully, Abraham Nitzan, Dynamics of gas-solid interactions: calculations of energy transfer and sticking, *J. Chem. Phys.* 66 (6) (1977) 2534–2544, <https://doi.org/10.1063/1.434249>.
- [21] Yuanran Zhu, Jason M. Dominy, Daniele Venturi, On the estimation of the Mori-Zwanzig memory integral, *J. Math. Phys.* 59 (10) (2018) 103501, <https://doi.org/10.1063/1.5003467>.
- [22] Huan Lei, Nathan A. Baker, Xiantao Li, Data-driven parameterization of the generalized Langevin equation, *Proc. Natl. Acad. Sci.* 113 (50) (2016) 14183–14188, <https://doi.org/10.1073/pnas.1609587113>.
- [23] R. Kubo, The fluctuation-dissipation theorem, *Rep. Prog. Phys.* 29 (1) (1966) 255–284, <https://doi.org/10.1088/0034-4885/29/1/306>.
- [24] Kevin K. Lin, Fei Lu, Data-driven model reduction, Wiener projections, and the Mori-Zwanzig formalism, *arXiv preprint*, arXiv:1908.07725, 2019.
- [25] Fei Lu, Kevin Lin, Alexandre Chorin, Comparison of continuous and discrete-time data-based modeling for hypoelliptic systems, *Commun. Appl. Math. Comput. Sci.* 11 (2) (2016) 187–216.
- [26] Chao Ma, Jianchun Wang, et al., Model reduction with memory and the machine learning of dynamical systems, *arXiv preprint*, arXiv:1808.04258, 2018.
- [27] Antonio Russo, Miguel A. Durán-Olivencia, Ioannis G. Kevrekidis, Serafim Kalliadas, Deep learning as closure for irreversible processes: a data-driven generalized Langevin equation, *arXiv preprint*, arXiv:1903.09562, 2019.
- [28] Yuanran Zhu, Daniele Venturi, Faber approximation of the Mori-Zwanzig equation, *J. Comput. Phys.* 372 (2018) 694–718.
- [29] Yuanran Zhu, Daniele Venturi, Generalized Langevin equations for systems with local interactions, *J. Stat. Phys.* (2020), <https://doi.org/10.1007/s10955-020-02499-y>.
- [30] Hee Sun Lee, Surl-Hee Ahn, Eric F. Darve, The multi-dimensional generalized Langevin equation for conformational motion of proteins, *J. Chem. Phys.* 150 (17) (2019), <https://doi.org/10.1063/1.5055573>.
- [31] Oliver F. Lange, Helmut Grubmüller, Collective Langevin dynamics of conformational motions in proteins, *J. Chem. Phys.* 124 (21) (2006) 214903, <https://doi.org/10.1063/1.2199530>.
- [32] Minxin Chen, Xiantao Li, Chun Liu, Computation of the memory functions in the generalized Langevin models for collective dynamics of macromolecules, *J. Chem. Phys.* 141 (6) (2014) 064112, <https://doi.org/10.1063/1.4892412>.
- [33] P. Linz, Numerical methods for Volterra integral equations of the first kind, *Comput. J.* 12 (4) (1969) 393–397, <https://doi.org/10.1093/comjnl/12.4.393>.
- [34] Brian Davies, *Integral Transforms and Their Applications*, Springer, 2010.
- [35] Thomas Hudson, Xingjie Helen Li, Coarse-graining of overdamped Langevin dynamics via the Mori-Zwanzig formalism, *arXiv e-prints*, arXiv:1810.08175, 2018.
- [36] Yuanran Zhu, Daniele Venturi, Hypocoellipticity and the Mori-Zwanzig formulation of stochastic differential equations, *arXiv e-prints*, arXiv:2001.04565, 2020.
- [37] H.J.C. Berendsen, D. van der Spoel, R. van Drunen, GROMACS: a message-passing parallel molecular dynamics implementation, *Comput. Phys. Commun.* 91 (1–3) (1995) 43–56, [https://doi.org/10.1016/0010-4655\(95\)00042-e](https://doi.org/10.1016/0010-4655(95)00042-e).
- [38] Junmei Wang, Romain M. Wolf, James W. Caldwell, Peter A. Kollman, David A. Case, Development and testing of a general AMBER force field, *J. Comput. Chem.* 25 (9) (2004) 1157–1174, <https://doi.org/10.1002/jcc.20035>.
- [39] William G. Hoover, Canonical dynamics: equilibrium phase-space distributions, *Phys. Rev. A* 31 (3) (1985) 1695–1697, <https://doi.org/10.1103/physreva.31.1695>.
- [40] Tom Darden, Darrin York, Lee Pedersen, Particle mesh Ewald: an $N \cdot \log(N)$ method for Ewald sums in large systems, *J. Chem. Phys.* 98 (12) (1993) 10089–10092, <https://doi.org/10.1063/1.464397>.
- [41] Berk Hess, Henk Bekker, Herman J.C. Berendsen, Johannes G.E.M. Fraaije, LINCS: a linear constraint solver for molecular simulations, *J. Comput. Chem.* 18 (12) (1997), [https://doi.org/10.1002/\(sici\)1096-987x\(199709\)18:12<1463::aid-jcc4>3.0.co;2-h](https://doi.org/10.1002/(sici)1096-987x(199709)18:12<1463::aid-jcc4>3.0.co;2-h).
- [42] Lina Ma, Xiantao Li, Chun Liu, The derivation and approximation of coarse-grained dynamics from Langevin dynamics, *J. Chem. Phys.* 145 (20) (2016) 204117, <https://doi.org/10.1063/1.4967936>.