Supplemental Material: Data-driven learning of the generalized Langevin equation with state-dependent memory

Pei Ge,[1] Zhongqiang Zhang,[2] and Huan Lei[1,3,*]

[1]*Department of Computational Mathematics, Science & Engineering, Michigan State University, MI 48824, USA*
[2]*Department of Mathematical Sciences, Worcester Polytechnic Institute, MA 01609, USA*
[3]*Department of Statistics & Probability, Michigan State University, MI 48824, USA*

I. MASS MATRIX OF THE REDUCED MODEL

In this study, we focus on the effect of the state-dependent non-Markovian memory on the collective behavior of complex systems. We choose the coarse-grained resolved variables such that the corresponding mass matrix is a constant. On the other hand, the mass matrix should further depend on the resolved variables for the general cases, and we refer to Refs. [1, 2] for further discussions and the reduced dynamics with position-dependent mass. Specifically, we define $q = \|\mathbf{Q}_1 - \mathbf{Q}_2\|$, where $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are the atom coordinates of the full model (see Fig. 1 and Sec. V for details). Accordingly, we have $\dot{q} = \mathbf{Q}_{12}^T \dot{\mathbf{Q}}_{12}/q$ and its covariance follows

$$\begin{aligned}
\langle \dot{q}\dot{q} \rangle &= \left\langle \frac{1}{q^2} \mathbf{Q}_{12}^T \dot{\mathbf{Q}}_{12} \dot{\mathbf{Q}}_{12}^T \mathbf{Q}_{12} \right\rangle \\
&= \left\langle \frac{1}{q^2} \mathrm{Tr}\left[(\mathbf{Q}_{12}\mathbf{Q}_{12}^T)(\dot{\mathbf{Q}}_{12}\dot{\mathbf{Q}}_{12}^T) \right] \right\rangle \\
&= \left\langle \frac{1}{q^2} \mathrm{Tr}(\mathbf{Q}_{12}\mathbf{Q}_{12}^T) \right\rangle \left(M_1^{-1} + M_2^{-1} \right) k_B T \\
&= \left(M_1^{-1} + M_2^{-1} \right) k_B T,
\end{aligned} \tag{1}$$

where M_1 and M_2 represent the mass of two atoms and we have used the fact that the distribution of $\mathbf{Q}_{12}$ and $\dot{\mathbf{Q}}_{12}$ are independent. Therefore, the mass matrix of q is a constant $M \equiv M_1 M_2/(M_1 + M_2)$.

II. COHERENT NOISE AND INVARIANT DENSITY OF THE REDUCED MODEL

We construct the reduced model

$$\begin{aligned}
\dot{\mathbf{q}} &= \mathbf{M}^{-1}\mathbf{p}, \\
\dot{\mathbf{p}} &= -\nabla U(\mathbf{q}) - \int_0^t \phi_t^T \Theta(t-\tau) \phi_\tau \mathbf{v}(\tau)\, \mathrm{d}\tau + \boldsymbol{\mathcal{R}}_t,
\end{aligned} \tag{2}$$

where $\phi_t = [\phi_1(\mathbf{q}_t), \cdots, \phi_n(\mathbf{q}_t)]$ and $\Theta(t)$ represent the spatial features and the kernel to be learned. In particular, $\Theta(t)$ is directly constructed in the Fourier space, i.e., $\Theta(t) = \mathrm{e}^{-\alpha t} \sum_{k=0}^{N_\omega} \hat{\Theta}_k \cos(\omega_k t)$, where $\omega_k = \frac{2\pi}{T_c} k$ and T_c is the time domain cut-off of the kernel. We note that $\mathrm{e}^{-\alpha t}$ should not be viewed as the bases to approximate $\Theta(t)$ (e.g., $\{\mathrm{e}^{-\alpha_i t}\cos(\beta_i t), \mathrm{e}^{-\alpha_i t}\sin(\beta_i t)\}_{i=1}^{N_\alpha}$; see Refs. [2, 3]). Rather, $\Theta(t)$ is mainly characterized by the Fourier series expansion on $[0, T]$, and the exponential term $\mathrm{e}^{-\alpha t}$ is essentially a regularization term to eliminate the periodicity while maintaining the semi-positive definiteness condition.

For the fluctuation term $\boldsymbol{\mathcal{R}}_t$, we represent it as a noise in the form of $\boldsymbol{\mathcal{R}}_t = \phi_t^T \widetilde{\mathbf{R}}(t)$, where $\widetilde{\mathbf{R}}(t)$ is a Gaussian random process whose covariance function determined by $\Theta(t)$, i.e., $\langle \widetilde{\mathbf{R}}(t)\widetilde{\mathbf{R}}(\tau)^T \rangle = k_B T \Theta(t-\tau)$. This choice avoids dealing with the orthogonal dynamics to calculate the fluctuation term. Furthermore, we can show that this choice enables the reduced model to retain a consistent invariant density function.

* leihuan@msu.edu

Proposition II.1. *For reduced model* (2) *with* $\Theta(t) = \mathrm{e}^{-\alpha t}\sum_{k=0}^{N_\omega}\hat{\Theta}_k\cos(\omega_k t)$*, by choosing the fluctuation term* $\boldsymbol{\mathcal{R}}_t = \phi_t^T\widetilde{\mathbf{R}}(t)$*, where* $\widetilde{\mathbf{R}}(t)$ *is a Gaussian random process satisfying*

$$\langle\widetilde{\mathbf{R}}(t)\widetilde{\mathbf{R}}(\tau)^T\rangle = k_BT\Theta(t-\tau), \tag{3}$$

the reduced model has an invariant distribution

$$\rho_{\mathrm{eq}}(\mathbf{q},\mathbf{p}) \propto \exp\left\{-\left[U(\mathbf{q}) + \mathbf{p}^T\mathbf{M}^{-1}\mathbf{p}/2\right]/k_BT\right\}. \tag{4}$$

Proof. Let us introduce auxiliary variables

$$\begin{aligned}
\mathbf{z}_{k,1} &= -\int_0^t \mathrm{e}^{-\alpha(t-\tau)}\Gamma_k\cos(\omega_k(t-\tau))\phi_\tau\mathbf{v}(\tau)\,\mathrm{d}\tau + \mathbf{R}_{k,1}(t),\\
\mathbf{z}_{k,2} &= -\int_0^t \mathrm{e}^{-\alpha(t-\tau)}\Gamma_k\sin(\omega_k(t-\tau))\phi_\tau\mathbf{v}(\tau)\,\mathrm{d}\tau + \mathbf{R}_{k,2}(t),
\end{aligned} \tag{5}$$

where $\Gamma_k^T\Gamma_k = \hat{\Theta}_k$ and $\mathbf{R}_{k,1}(t)$ is a Gaussian random process satisfying

$$\left\langle\mathbf{R}_{j,1}(t)\mathbf{R}_{k,1}(\tau)^T\right\rangle = k_BT\delta_{jk}\mathrm{e}^{-\alpha(t-\tau)}\cos(\omega_k(t-\tau)), \tag{6}$$

where δ_{jk} is the Kronecker delta. Accordingly, the second equation of Eq. (2) can be written as

$$\dot{\mathbf{p}} = -\nabla U(\mathbf{q}) + \phi(\mathbf{q})^T\sum_k\Gamma_k^T\mathbf{z}_{k,1}, \tag{7}$$

and $\mathbf{R}_{j,2}(t)$ will be specified later.

Let $\mathbf{z}_k = [\mathbf{z}_{k,1},\mathbf{z}_{k,2}]$ and $\mathbf{R}_k = [\mathbf{R}_{k,1},\mathbf{R}_{k,2}]$, we can rewrite Eq. (5) by

$$\begin{aligned}
\mathbf{z}_k &= -\int_0^t \mathrm{e}^{-\alpha(t-\tau)}\begin{pmatrix}\cos(\omega_k(t-\tau))I & \sin(\omega_k(t-\tau))I\\ -\sin(\omega_k(t-\tau))I & \cos(\omega_k(t-\tau))I\end{pmatrix}\begin{pmatrix}\Gamma_k\phi_\tau\mathbf{v}(\tau)\\ 0\end{pmatrix}\mathrm{d}\tau + \mathbf{R}_k(t)\\
&= -\int_0^t \exp\left[\begin{pmatrix}-\alpha I & \omega_k I\\ -\omega_k I & -\alpha I\end{pmatrix}(t-\tau)\right]\begin{pmatrix}\Gamma_k\phi_\tau\mathbf{v}(\tau)\\ 0\end{pmatrix}\mathrm{d}\tau + \mathbf{R}_k(t).
\end{aligned} \tag{8}$$

By taking the time derivative of Eq. (8) with respect to t, we have

$$\frac{\mathrm{d}\mathbf{z}_k}{\mathrm{d}t} = \underbrace{\begin{pmatrix}-\alpha I & \omega_k I\\ -\omega_k I & -\alpha I\end{pmatrix}}_{\triangleq\mathbf{J}}\mathbf{z}_k - \begin{pmatrix}\Gamma_k\phi_\tau\mathbf{v}(t)\\ 0\end{pmatrix} + \frac{\mathrm{d}\mathbf{R}_k}{\mathrm{d}t} - \mathbf{J}\mathbf{R}_k(t). \tag{9}$$

Furthermore, we note that $\mathbf{R}_k(t)$ can be modeled as a generalized Ornstein–Uhlenbeck process and $\frac{\mathrm{d}\mathbf{R}_k}{\mathrm{d}t} - \mathbf{J}\mathbf{R}_k(t)$ can be represented by

$$\frac{\mathrm{d}\mathbf{R}_k}{\mathrm{d}t} - \mathbf{J}\mathbf{R}_k(t) = \Lambda_k\dot{\mathbf{W}}_{k,t}, \tag{10}$$

where $\dot{\mathbf{W}}_{k,t}$ is the standard white noise and $\Lambda_k\Lambda_k^T = -k_BT(\mathbf{J}+\mathbf{J}^T)$. With this choice, the covariance of $\mathbf{R}_k(t) = [\mathbf{R}_{k,1},\mathbf{R}_{k,2}]$ is given by

$$\left\langle\mathbf{R}_k(t)\mathbf{R}_k(\tau)^T\right\rangle = k_BT\mathrm{e}^{-\alpha(t-\tau)}\begin{pmatrix}\cos(\omega_k(t-\tau))I & \sin(\omega_k(t-\tau))I\\ -\sin(\omega_k(t-\tau))I & \cos(\omega_k(t-\tau))I\end{pmatrix}$$

such that Eq. (6) remains valid. Using Eqs. (5)(7)(9), we can write the reduced model (2) in the form of

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}\mathbf{q}\\ \mathbf{p}\\ \cdots\\ \mathbf{z}_{k,1}\\ \mathbf{z}_{k,2}\\ \cdots\end{pmatrix} &= \begin{pmatrix}0 & I & \cdots & 0 & 0 & \cdots\\ -I & 0 & \cdots & \phi(\mathbf{q})^T\Gamma_k^T & 0 & \cdots\\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots\\ 0 & -\Gamma_k\phi(\mathbf{q}) & \cdots & -\alpha I & -\omega_k I & \cdots\\ 0 & 0 & \cdots & \omega_k I & -\alpha I & \cdots\\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots\end{pmatrix}\begin{pmatrix}\nabla U(\mathbf{q})\\ \mathbf{v}\\ \cdots\\ \mathbf{z}_{k,1}\\ \mathbf{z}_{k,2}\\ \cdots\end{pmatrix} + \begin{pmatrix}0\\ 0\\ \cdots\\ \Lambda_k\dot{\mathbf{W}}_{k,t}\\ \cdots\end{pmatrix}\\
&\triangleq \mathbf{K}\nabla F(\mathbf{q},\mathbf{p},\cdots,\mathbf{z}_{k,1},\mathbf{z}_{k,2},\cdots) + \Lambda\dot{\mathbf{W}}_t,
\end{aligned} \tag{11}$$

where $\mathbf{K}$ is the first matrix of the right-hand-side of Eq. (11), $F(\mathbf{q},\mathbf{p},\cdots,\mathbf{z}_{k,1},\mathbf{z}_{k,2},\cdots) = U(\mathbf{q}) + \frac{1}{2}\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p} + \frac{1}{2}\sum_{k=1}^{N_\omega}\left(\mathbf{z}_{k,1}^T\mathbf{z}_{k,1} + \mathbf{z}_{k,2}^T\mathbf{z}_{k,2}\right)$ is the total free energy of the extended system, and $\Lambda = \mathrm{diag}(0,0,\cdots,\Lambda_k,\cdots)$. Using (10), it is easy to show $\Lambda\Lambda^T = -k_BT(\mathbf{K}+\mathbf{K}^T)$. Therefore, the gradient system (11) (i.e., the reduced model (2)) has the invariant density function

$$\rho_{\mathrm{eq}}(\mathbf{q},\mathbf{p},\mathbf{z}) \propto \exp\left[-F(\mathbf{q},\mathbf{p},\mathbf{z})/k_BT\right].$$

□

III. TRAINING DETAILS OF THE REDUCED MODEL

We represent the state encoders $\phi(\mathbf{q}) = [\phi_1(\mathbf{q}),\cdots,\phi_n(\mathbf{q})]$ and $\Theta(t)$ in reduced model (2) by

$$\begin{aligned}\phi_i(\mathbf{q}) &= \mathbf{H}_i^T\psi(\mathbf{q})\\ \Theta(t) &= \mathrm{e}^{-\alpha t}\sum_{k=0}^{N_\omega}\hat{\Theta}_k\cos(\omega_k t),\end{aligned} \tag{12}$$

where $\psi(\mathbf{q}) = [\psi_1(\mathbf{q}),\cdots,\psi_{N_b}(\mathbf{q})]$ is a set of sparse bases, $\mathbf{H} = \left[\mathbf{H}_1^T,\cdots,\mathbf{H}_n^T\right]$ are trainable coefficients. $\omega_k = 2\pi k/T_c$ is the frequency, where T_c is the time domain cut-off of the kernel and δt is the step size of the discrete samples. For the present study, ψ is chosen as the uniform piecewise linear basis function defined on $[2.8, 4.1]$ with $N_b = 66$, $T_c = 200$ and $N_\omega = 2000$.

To train the reduced model, we use the correlation functions conditional with differential initial states such that the state-dependent nature can be manifested. Specifically, we right-multiply the second equation of Eq. (2) by $\mathbf{v}(0)$ and take the conditional expectation on $\mathbf{q}_0 = \mathbf{q}^*$, which defines $\mathbf{g}(t;\mathbf{q}^*) = \left\langle[\dot{\mathbf{p}}_t + \nabla U(\mathbf{q}_t)]\mathbf{v}_0^T|\mathbf{q}_0 = \mathbf{q}^*\right\rangle$. Accordingly, we define the empirical loss function

$$\begin{aligned}L &= \sum_{l=1}^{N_q}\sum_{k=1}^{N_t}\left\|\widetilde{\mathbf{g}}(t_k;\mathbf{q}^{(l)}) - \mathbf{g}(t_k;\mathbf{q}^{(l)})\right\|^2,\\ \widetilde{\mathbf{g}}(t_k;\mathbf{q}^{(l)}) &= \sum_{j=1}^{k}\mathrm{Tr}\left[\Theta(t_k-t_j)\mathbf{H}\mathbf{C}_{\psi,\psi}(t_k,t_j;\mathbf{q}^{(l)})\mathbf{H}^T\right]\delta t,\end{aligned}$$

where $\widetilde{\mathbf{g}}(\cdot)$ represents the prediction by the reduced model; l and k correspond to various initial states and discrete time, respectively. $\mathbf{C}_{\psi,\psi}(t,\tau;\mathbf{q}^*) = \left\langle\psi_\tau\mathbf{v}_\tau\mathbf{v}_0^T\psi_t^T|\mathbf{q}_0 = \mathbf{q}^*\right\rangle$ is a three-point correlation characterizing the coupling among the bases. We use 2×10^6 short trajectories extracted from the full MD simulation (see Sec. V for details) where each one consists of 300 time-series samples to sample the correlation for $\mathbf{q}^*$ at the saddle point. We use $2\times10^6 \sim 8\times10^6$ short trajectories for other individual points.

Besides the conditional correlation functions, we can also introduce the loss function with respect to the overall correlation function, i.e.,

$$\begin{aligned}L_2 &= \sum_{k=1}^{N_t}\left\|\widetilde{\mathbf{g}}_2(t_k) - \mathbf{g}_2(t_k)\right\|^2\\ \widetilde{\mathbf{g}}_2(t_k) &= \sum_{j=1}^{k}\mathrm{Tr}\left[\Theta(t_k-t_j)\mathbf{H}\overline{\mathbf{C}}_{\psi,\psi}(t_k,t_j)\mathbf{H}^T\right]\delta t,\end{aligned}$$

where $\mathbf{g}_2(t) = \left\langle[\dot{\mathbf{p}}_t + \nabla U(\mathbf{q}_t)]\mathbf{v}_0^T\right\rangle$ is the overall correlation and $\overline{\mathbf{C}}_{\psi,\psi}(t,\tau) = \left\langle\psi_\tau\mathbf{v}_\tau\mathbf{v}_0^T\psi_t^T\right\rangle$. In particular, if there is scale separation between $c_{vv}(t)$ and $c_{qq}(t)$ (e.g., $c_{vv}(t)$ decays much faster than $c_{qq}(t)$; see Fig. 2 and Fig. 4 in the main manuscript), we may approximate $\overline{\mathbf{C}}_{\psi,\psi}(t,\tau)$ by two-point correlations, i.e., $\overline{\mathbf{C}}_{\psi,\psi}(t,\tau) \approx \left\langle\psi_\tau\otimes\psi_t^T\right\rangle : \left\langle\mathbf{v}_\tau\mathbf{v}_0^T\right\rangle$.

Efficient training is achieved by using the following numerical methods to evaluate $\widetilde{\mathbf{g}}$ and $\widetilde{\mathbf{g}}_2$. Specifically, $\psi_\tau\psi_t^T$ can be efficiently pre-computed with $O(1)$ complexity by using the sparse piecewise linear basis functions. Furthermore, we can use the low-rank representation (e.g., based on the singular value decomposition) of $\mathbf{C}_{\psi,\psi}$ and $\overline{\mathbf{C}}_{\psi,\psi}$ to accelerate the matrix production $\mathbf{H}\mathbf{C}_{\psi,\psi}\mathbf{H}^T$. In addition, the convolution on index j can be efficiently evaluated by the Fast Fourier Transform algorithm [4].

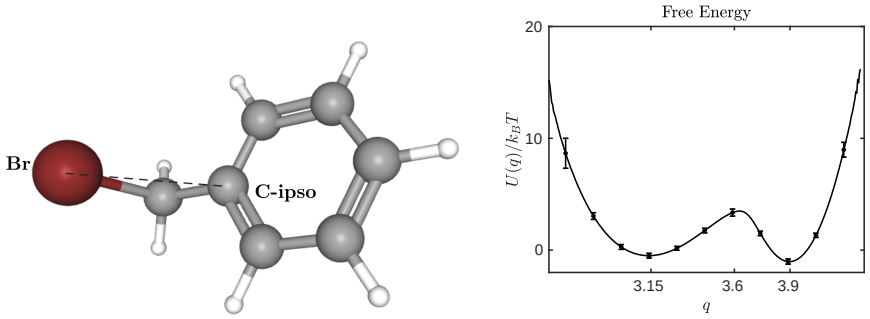

FIG. 1. Left: A sketch of the molecule benzyl bromide. The resolved variable is defined as the distance between the bromine atom and the ipso-carbon atom. Right: The free energy of the resolved variable q. The error bar represent the 95% confidence interval.

While L_2 alone is insufficient to characterize the emergence of the state-dependent memory, it serves as a necessary condition and can facilitate the learning of the reduced model. In practice, we can use both loss functions to train the reduced model with $N_q = 65$, $N_t = 300$ for L, and $N_t = 30000$ for L_2. Specifically, the training is conducted by the Adam [5] optimization method in three stages with 2000, 6000, and 6000 steps respectively. For the first stage, we only use L_2 to train the model with a constant learning rate of 0.04. For each step of the following two stages, 16 initial states (i.e., $q^{(l)}$) are randomly selected as one training batch to evaluate the total loss $L_t = L + L_2$. For both stages, the initial learning rate is 1×10^{-2} and the exponential decay rate is 0.9 per 150 steps. In practice, the number of N_b and N_q can be further reduced. As shown in the GitHub repository, with $N_b = 8$, $N_q = 26$, the constructed reduced model can accurately recover the MD results.

IV. SIMULATION OF THE REDUCED MODEL

We simulate the reduced model (2) by generating the noise term using the constructed Fourier modes rather than Markovian embedding (e.g., see Ref. [6]). Following Prop. II.1, we generate $\widetilde{\mathbf{R}}(t)$ on $[0, T]$ similar to Refs. [7, 8] by

$$\widetilde{\mathbf{R}}(t) = \beta^{-1/2} \sum_{k=0}^{2N} \widetilde{\Theta}_k^{1/2} \left[\cos(\omega_k t)\xi_k + \sin(\omega_k t)\eta_k\right], \tag{13}$$

where $\beta^{-1} = k_B T$, $\widetilde{\Theta}_k$ are the Fourier (essentially cosine) modes of $\Theta(|t|)$ on $[-T, T]$, ξ_k and η_k are independent Gaussian random vectors, and N is the total number of simulation step (see also Ref. [9]).

Specifically, for large simulation time T, the Fourier modes $\widetilde{\Theta}_k$ is given by

$$\begin{aligned}
\widetilde{\Theta}_k &= \sum_{j=1}^{N_\omega} \int_{-T}^{T} \mathrm{e}^{-\alpha|t|} \hat{\Theta}_j \cos(\omega_j t) \cos(\omega_k t)\, \mathrm{d}t \\
&= \sum_{j=1}^{N_\omega} \int_{0}^{T} \mathrm{e}^{-\alpha t} \hat{\Theta}_j \left(\cos\left((\omega_j - \omega_k)t\right) + \cos\left(\omega_j + \omega_k)t\right)\right) \\
&\approx \sum_{j=1}^{N_\omega} \left(\frac{\alpha \hat{\Theta}_j}{\alpha^2 + (\omega_j - \omega_k)^2} + \frac{\alpha \hat{\Theta}_j}{\alpha^2 + (\omega_j + \omega_k)^2} \right).
\end{aligned} \tag{14}$$

Therefore $\widetilde{\mathbf{R}}(t)$ can be generated using the Fast Fourier Transform algorithm [4] using $O(N \log N)$ complexity. Also, the convolution term $\int_0^t \phi_t^T \Theta(t-\tau)\phi_\tau \mathbf{v}(\tau)\, \mathrm{d}\tau$ in Eq. (2) can be computed using the fast convolution method developed in Ref. [10] with $O(N \log N)$ complexity.

V. FULL ATOMISTIC MODEL

We consider the full micro-scale model of benzyl bromide (see Fig. 1 for a sketch of the molecule structure) in an aqueous environment. The general AMBER [11] force field is used for the benzyl bromide molecule and the

partial charges of molecule atoms were set by the restrained electrostatic potential (RESP) approach [12]. The rigid TIP3P water model [13] is used for the water molecules and the bond lengths and angles were held constant through the SHAKE algorithm [14, 15]. Long-range electrostatic interactions were calculated using a Particle Mesh Ewald summation with a relative error set to be 10^{-4}. The full system consists of one benzyl bromide molecule and 2400 water molecules with the periodic boundary condition imposed along each direction. The isothermal-isobaric thermostat [16] is used to equilibrate the system for 16 ns at 298K and 1 bar using a time step of 1 fs. Following the equilibration, the box size is scaled to be near $41.5 \times 41.5 \times 41.5$ Å^3. The simulation was run for a production period of 1.5 μs in a canonical ensemble with a Nosé-Hoover thermostat [17, 18]. The numerical results of this work (both main manuscript and SM) are presented in Å for length, picosecond for time, and gram per mole for mass.

The resolved variable q is defined as the distance between the bromine atom and the ipso-carbon atom. The free energy is obtained from the probability density function $\rho(q)$ (see the inset plot of Fig. 2(a) in the main manuscript), i.e., $U(q) = -k_B T \ln \rho(q)$, where $\rho(q)$ is directly obtained from the full MD samples using the kernel density estimation. To verify the accuracy of the constructed $U(q)$, we calculate the expectation of $q\nabla U(q)$ on the sample. The numerical result gives $0.996 k_B T$ and is close to the theoretical prediction $\langle q\nabla U(q)\rangle = \int q\nabla U(q) \mathrm{e}^{-U(q)/k_B T}\, \mathrm{d}q \equiv k_B T$.

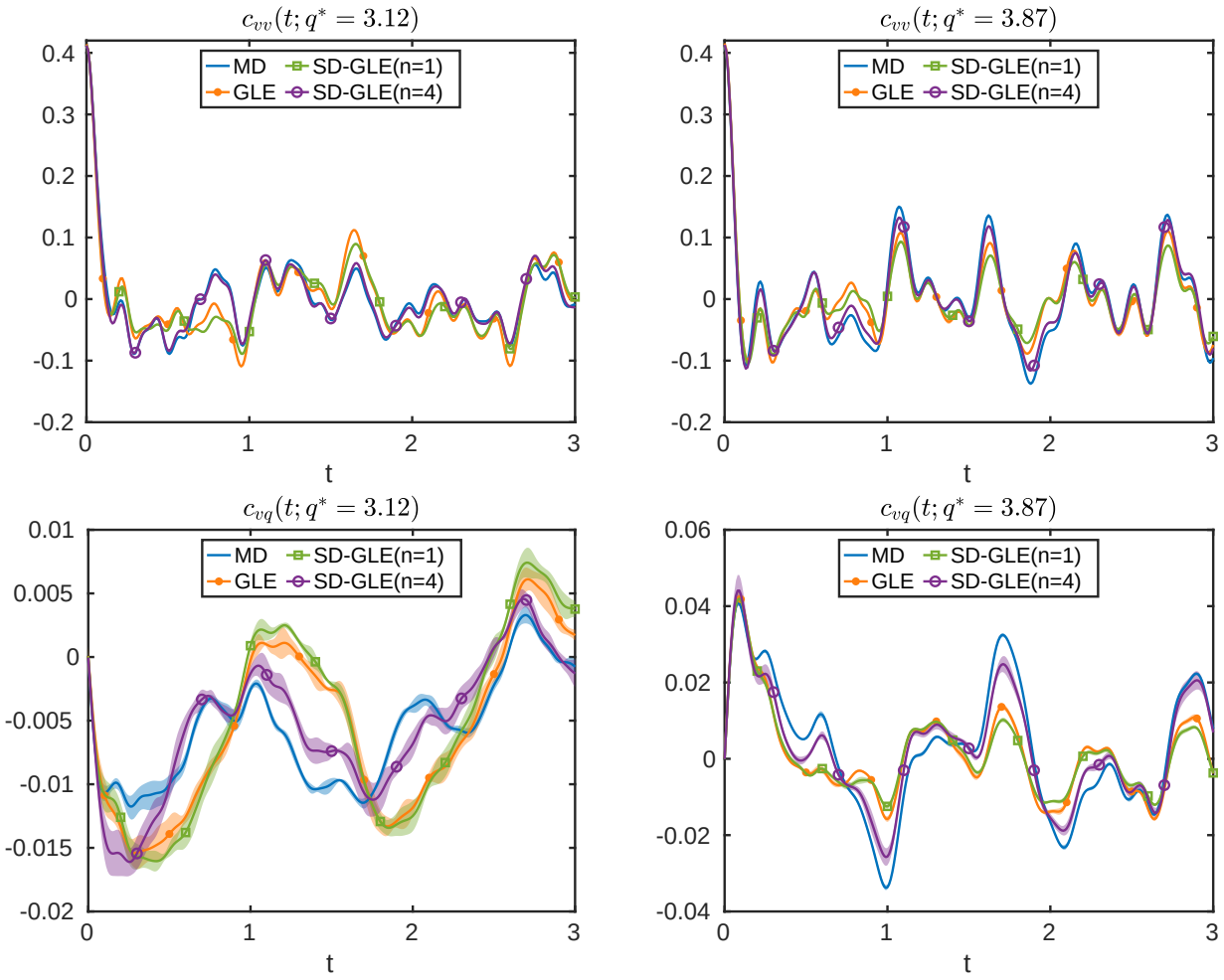

FIG. 2. The conditional correlation functions $c_{vq}(t, q^*)$ and $c_{vv}(t, q^*)$ for the two local minima predicted by the full MD, the standard GLE, and the present model (SD-GLE) constructed using one and four spatio-features. Left: $q^* = 3.07$; Right: $q^* = 3.87$. The predictions by the standard GLE show apparent discrepancies with the full MD results. Shaded regimes represent the 95% confidence interval.

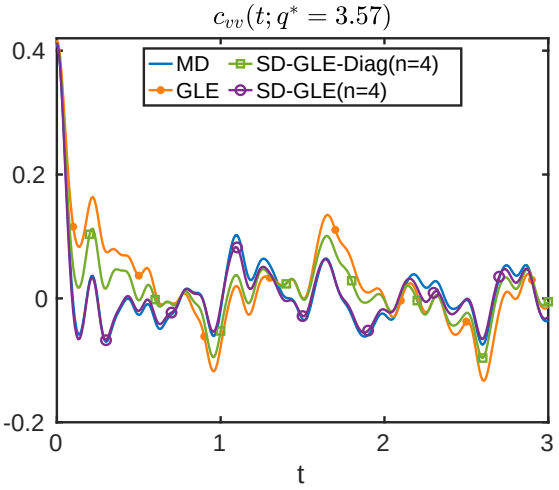

FIG. 3. The conditional correlation functions $c_{vv}(t, q^*)$ for the saddle point predicted by the full MD, the standard GLE, the reduced models constructed using four state-features with diagonal $\Theta(t)$ (SD-GLE-Diag) and full $\Theta(t)$ (SD-GLE). The large discrepancy between the SD-GLE-Diag model and the full MD results implies the complexity of the state-dependency of the memory term, which can not be well represented by the coupling of independent bath variables. The non-Markovian interactions among the state-features are essential to capture the heterogeneous energy dissipation process.

VI. ADDITIONAL NUMERICAL RESULTS

A. Limitations of the standard GLE near the local minima

Fig. 2 shows the predictions of the conditional correlations $c_{qv}(t, q^*)$ and $c_{vv}(t, q^*)$ for q^* representing the two local minima. Similar to the results of the saddle point (see Fig. 3(b) in the main manuscript), the predictions of the standard GLE show apparent deviations from the full MD results due to the ignorance of the state-dependent memory nature. In contrast, the predictions of the present model with four features can accurately recover the MD predictions.

B. Other forms of the state-dependent memory term

For comparison, we also consider other forms of the reduced model. In particular, we retain the encoders ϕ in Eq. (2) but set $\Theta(t)$ to be diagonal, i.e., we ignore the non-Markovian coupling among the different state features. The reduced model is trained using four features. Fig. 3 shows the conditional correlation $c_{vv}(t; q^*)$ obtained from the full MD and different reduced models. The prediction of the constructed model (labeled by "SD-GLE-Diag") shows apparent deviations from the full MD result with incremental improvement over the stand GLE. The large discrepancy reveals the complex state-dependent nature; the non-Markovian effect can be neither approximated by ansatz like $\gamma(q)\theta(t)$ as a simple generalization/re-scaling of the initial value at $t = 0$, nor represented by the coupling with the independent bath variables. Instead, the non-Markovian coupling among the various state-features retained in the present model plays a crucial for accurately modeling the heterogeneous energy dissipation arising from the unresolved intramolecular interactions and reproducing the collective dynamics.

Furthermore, we note that Ref. [19] develops an efficient approach to compute the memory function based on the finite-rank approximation of the Zwanzig's projection formalism. The method can be used to efficiently extract the state-dependent memory and probe the physics insights from the trajectory samples of the full MD model. The present work focuses on training a reduced model with heterogeneous memory that enables generating coherent noise and conducting stochastic simulations.

Specifically, the memory term takes the form $K(q(\tau), t-\tau)$ (with a simple change of variable $s = t - \tau$ following the notation) in Ref. [19] and $\tilde{K}(q(t), q(\tau), t-\tau) = \phi(q(t))^T \Theta(t-\tau)\phi(q(\tau))$ in the present work. In particular, by setting $q(t) = q(\tau)$, the memory term in two forms should have similar prediction. Following this argument, we use the method in Ref. [19] to calculate $K(q(\tau), t-\tau)$ and compare it with $\tilde{K}(q(t), q(\tau), t-\tau)$ of the present model with $q(t) = q(\tau) = q^*$. Fig. 4 shows the obtained memory functions for q^* taking the saddle point and two local minima. The prediction by the two approaches show good agreement.

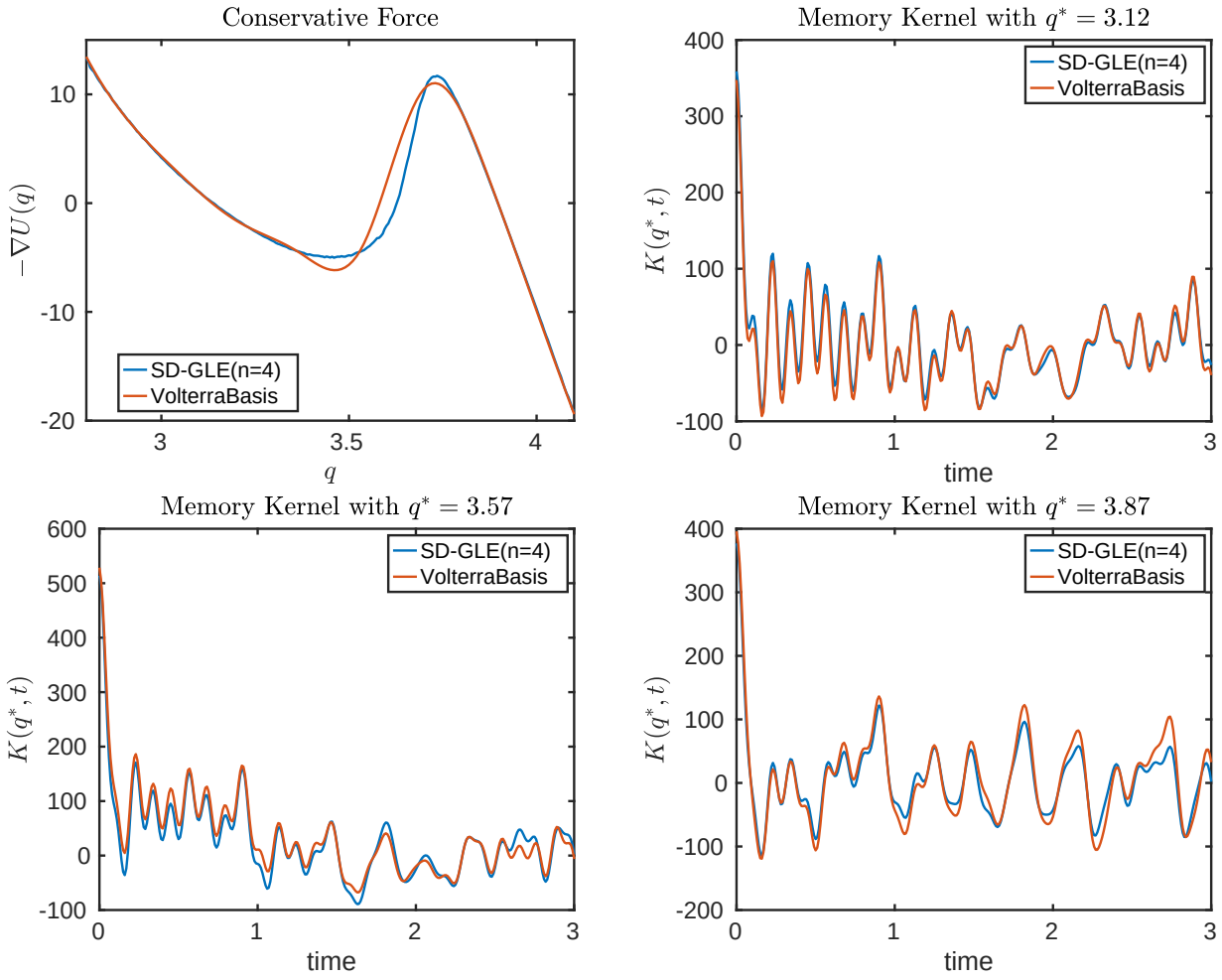

FIG. 4. The conservative force and memory kernel obtained from Ref. [19] (labeled as Volterra Basis) and the present model. The memory term takes the form $K(q(\tau), t-\tau)$ and $\tilde{K}(q(t), q(\tau), t-\tau) = \phi(q(t))^T \Theta(t-\tau)\phi(q(\tau))$, respectively. Specifically, we set $q(t) = q(\tau) = q^*$; the predictions of the two models show good agreement for the saddle point and the two local minima.

VII. GENERALIZATION OF THE PRESENT REDUCED MODEL FORMULATION

So far, we have constructed the reduced model (2) by assuming the matrix-valued kernel $\Theta(t)$ is symmetry. In fact, this form can be generalized by introducing an anti-symmetry part, i.e.,

$$\Theta(t) = \mathrm{e}^{-\alpha t} \sum_{k=0}^{N_\omega} (\Gamma_{k,1}^T \Gamma_{k,1} + \Gamma_{k,2}^T \Gamma_{k,2}) \cos(\omega_k t) + (\Gamma_{k,1}^T \Gamma_{k,2} - \Gamma_{k,2}^T \Gamma_{k,1}) \sin(\omega_k t), \tag{15}$$

where $\Gamma_{k,1}$ and $\Gamma_{k,2}$ are lower-triangular matrices representing the Fourier modes of $\Theta(t)$. The form is general non-symmetric except for $t = 0$ and satisfies $\Theta(-t) = \Theta(t)^T$.

Similar to the symmetry form, we can model the fluctuation term $\boldsymbol{\mathcal{R}}_t$ as a noise in the form of $\boldsymbol{\mathcal{R}}_t = \phi_t^T \widetilde{\mathbf{R}}(t)$, where $\widetilde{\mathbf{R}}(t)$ is a Gaussian random process satisfying $\langle \widetilde{\mathbf{R}}(t) \widetilde{\mathbf{R}}(\tau)^T \rangle = k_B T \mathrm{e}^{-\alpha(t-\tau)} \Theta(t-\tau)$. Similar to Prop. II.1, we can show that this choice retains a consistent invariant density function. In practice, we can generate the noise term $\widetilde{\mathbf{R}}(t)$ on $[0, T]$ by

$$\widetilde{\mathbf{R}}(t) = \beta^{-1/2} \sum_{k=0}^{2N} \left[\widetilde{\Theta}_{k,1}^{1/2} \cos(\omega_k t) \xi_k + \sin(\omega_k t) (Q_{1,k}^{1/2} \xi_k + Q_{2,k}^{1/2} \eta_k) \right],$$
$$Q_{1,k} = \widetilde{\Theta}_{k,2} \widetilde{\Theta}_{k,1}^{-1} \widetilde{\Theta}_{k,2}^T \quad Q_{2,k} = \widetilde{\Theta}_{k,1} - \widetilde{\Theta}_{k,2} \widetilde{\Theta}_{k,1}^{-1} \widetilde{\Theta}_{k,2}^T,$$

where $\beta^{-1} = k_B T$, $\widetilde{\Theta}_{k,1}$, $\widetilde{\Theta}_{k,2}$ are the Fourier cosine and sine modes on $[-T, T]$ with $\Theta(-t) = \Theta(t)^T$, ξ_k and η_k are independent Gaussian random vectors, and N is the total number of simulation step. Here $\widetilde{\mathbf{R}}(t)$ can still be generated using the Fast Fourier Transform algorithm [4] using $O(N \log N)$ complexity. We will investigate this generalized formulation for model reduction in future studies.

[1] C. Ayaz, L. Scalfi, B. A. Dalton, and R. R. Netz, Physical Review E **105**, 054138 (2022).
[2] H. S. Lee, S.-H. Ahn, and E. F. Darve, The Journal of Chemical Physics **150**, 174113 (2019).
[3] H. Lei, N. A. Baker, and X. Li, Proc. Natl. Acad. Sci. **113**, 14183 (2016).
[4] J. W. Cooley and J. W. Tukey, Mathematics of Computation **19**, 297 (1965).
[5] D. Kingma and J. Ba, International Conference on Learning Representations (ICLR) (2015).
[6] C. Ayaz, L. Tepper, and R. R. Netz, Turkish Journal of Physics **46**, 194 (2022).
[7] M. Berkowitz, J. Morgan, and J. A. McCammon, J. Chem. Phys. **78**, 3256 (1983).
[8] V. A. Ogorodnikov and S. M. Prigarin, *Numerical Modelling of Random Processes and Fields: Algorithms and Applications* (De Gruyter, Berlin, Boston, 1996).
[9] Z. Li, X. Bian, X. Li, and G. E. Karniadakis, The Journal of chemical physics **143**, 243128 (2015).
[10] A. Schädle, M. López-Fernández, and C. Lubich, SIAM Journal on Scientific Computing **28**, 421 (2006).
[11] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, Journal of Computational Chemistry **25**, 1157 (2004).
[12] C. I. Bayly, P. Cieplak, W. Cornell, and P. A. Kollman, The Journal of Physical Chemistry **97**, 10269 (1993).
[13] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, The Journal of Chemical Physics **79**, 926 (1983).
[14] J.-P. Ryckaert, G. Ciccotti, and H. J. Berendsen, Journal of Computational Physics **23**, 327 (1977).
[15] S. Miyamoto and A. Kollman Peter, Journal of Computational Chemistry **13**, 952 (2004).
[16] G. J. Martyna, D. J. Tobias, and M. L. Klein, The Journal of Chemical Physics **101**, 4177 (1994).
[17] S. Nosé, Molecular Physics **52**, 255 (1984).
[18] W. G. Hoover, Physical Review A **31**, 1695 (1985).
[19] H. Vroylandt and P. Monmarché, The Journal of Chemical Physics **156**, 244105 (2022).