The Journal of Chemical Physics

CORE-MD II: A fast, adaptive, and accurate enhanced sampling method

Cite as: J. Chem. Phys. **155**, 104114 (2021); https://doi.org/10.1063/5.0063664
Submitted: 16 July 2021 • Accepted: 26 August 2021 • Published Online: 14 September 2021

Emanuel K. Peter, Dietmar J. Manstein, Joan-Emma Shea, et al.

openaccess

View Online Export Citation CrossMark

ARTICLES YOU MAY BE INTERESTED IN

Machine learning implicit solvation for molecular dynamics
The Journal of Chemical Physics **155**, 084101 (2021); https://doi.org/10.1063/5.0059915

Chemical physics software
The Journal of Chemical Physics **155**, 010401 (2021); https://doi.org/10.1063/5.0059886

Physics-based, neural network force fields for reactive molecular dynamics: Investigation of carbene formation from [EMIM$^+$][OAc$^-$]
The Journal of Chemical Physics **155**, 104112 (2021); https://doi.org/10.1063/5.0063187

AIP Publishing

CORE-MD II: A fast, adaptive, and accurate enhanced sampling method

Cite as: J. Chem. Phys. 155, 104114 (2021); doi: 10.1063/5.0063664
Submitted: 16 July 2021 • Accepted: 26 August 2021 •
Published Online: 14 September 2021

Emanuel K. Peter,[1,2,a)] Dietmar J. Manstein,[1,2,3] Joan-Emma Shea,[4] and Alexander Schug[5,6,b)]

AFFILIATIONS

[1] Institute for Biophysical Chemistry, Fritz-Hartmann-Centre for Medical Research, Hannover Medical School, Carl-Neuberg-Str. 1, Hannover 30625, Germany
[2] Division for Structural Biochemistry, Hannover Medical School, 30625 Hannover, Germany
[3] RESiST, Cluster of Excellence 2155, Medizinische Hochschule Hannover, 30625 Hannover, Germany
[4] Department of Chemistry and Biochemistry, Department of Physics, University of California, Santa Barbara, California 93106, USA
[5] John von Neumann Institute for Computing and Jülich Supercomputing Centre, Institute for Advanced Simulation, Forschungszentrum Jülich, 52425 Jülich, Germany
[6] Faculty of Biology, University of Duisburg-Essen, 45141 Essen, Germany

[a)] **Electronic mail:** peter.emanuel@mh-hannover.de
[b)] **Author to whom correspondence should be addressed:** al.schug@fz-juelich.de

ABSTRACT

In this paper, we present a fast and adaptive correlation guided enhanced sampling method (CORE-MD II). The CORE-MD II technique relies, in part, on partitioning of the entire pathway into short trajectories that we refer to as instances. The sampling within each instance is accelerated by adaptive path-dependent metadynamics simulations. The second part of this approach involves kinetic Monte Carlo (kMC) sampling between the different states that have been accessed during each instance. Through the combination of the partition of the total simulation into short non-equilibrium simulations and the kMC sampling, the CORE-MD II method is capable of sampling protein folding without any *a priori* definitions of reaction pathways and additional parameters. In the validation simulations, we applied the CORE-MD II on the dialanine peptide and the folding of two peptides: TrpCage and TrpZip2. In a comparison with long time equilibrium Molecular Dynamics (MD), 1 μs replica exchange MD (REMD), and CORE-MD I simulations, we find that the level of convergence of the CORE-MD II method is improved by a factor of 8.8, while the CORE-MD II method reaches acceleration factors of ~120. In the CORE-MD II simulation of TrpZip2, we observe the formation of the native state in contrast to the REMD and the CORE-MD I simulations. The method is broadly applicable for MD simulations and is not restricted to simulations of protein folding or even biomolecules but also applicable to simulations of protein aggregation, protein signaling, or even materials science simulations.

I. INTRODUCTION

Molecular Dynamics (MD) and Monte Carlo (MC) simulations are important theoretical tools for the investigation of biological systems on a molecular level. For the last two decades, theoretical improvements, the improvement of forcefield parameter sets, and a substantial rise of the performance of hardware established MD and MC as methods that are complementary to experiments. MD and MC strongly contributed to a better understanding of the underlying molecular processes of protein folding,[1–4] protein aggregation,[5,6] and protein signaling.[7] Both methods play an essential role in the modeling of drug–host interactions,[8] protein self-assembly,[9] and protein membrane interactions.[10] Despite the importance of both techniques, the timescales of biological processes can exceed the accessible computable time-ranges by many orders of magnitude due to the complexity of the underlying energy landscapes. This problem has been tackled through the improvement of the computer hardware,[11] a reduction in the complexity through the development of coarse-grained models,[12–14] and algorithmic improvements that raised the performance of MD and MC.[15–19]

Leaving aside advances based on coarse-grained modeling and the development of efficient hardware and software, the group of umbrella sampling methods solves the timescale problem through projections of the trajectory space to underlying energy landscapes that contain the dimensionality of specific collective variables. Several groups and sub-groups of methods have emerged that belong to the class of umbrella sampling methods.[20,21] Prominent members of that class are the metadynamics method,[22] λ-dynamics,[23,24] adaptive bias MD,[25] the hyperdynamics method,[26] conformational flooding,[27] and the local elevation technique.[28] Regardless of the aforementioned classes of methods, techniques that accelerate the sampling in the trajectory space are adaptive and in most cases do not contain the need of *a priori* definitions of reaction coordinates. This group of methods relies on first-principles conjectures, such as approximations of the geometric degrees of freedom of a molecule or propagation methods that coarse-grain fast thermal fluctuations of a protein. Here, we refer to specific classes of constraint methods, while the use of constraints is not interpreted as an enhanced sampling in the community, although bonded, non-bonded, and angular constraints can contribute to 2–15-fold speed-ups of the MD and MC sampling.[29,30] More broadly, constraining the internal degrees of motion will enable even much higher accelerations, such as in simulations of the Brownian motion of protein complexes or the extended capture of drug–host interactions.[31] Langevin and Brownian dynamics in combination with constraints contain stochastic terms in the propagation scheme and can be applied for the simulation of protein-aggregation phenomena on timescales ranging from micro- to milli-seconds. Hybrid techniques use a combination of the stochastic nature of MC with the deterministic propagation in MD and have the potential of further improvements of the sampling efficiency. Hybrid MC approaches range from the hybrid MC methodologies[32–35] to hybrid kinetic Monte Carlo/MD (kMC/MD)[36–38] approaches that apply an event-driven acceleration. In particular, kMC/MD approaches allow for larger propagations along the time-axis dependent on the nature of the move, which has been shown in simulations of protein folding and protein signaling.[39] Although the classes, groups, and sub-groups of methods we discuss here have proven their performance in simulations, a larger fraction of the methods applies biases or modifications in the energy space of the system, which leads to the occurrence of non-equilibrium states and the need to define appropriate reaction coordinates.[40,41] Considerable efforts have been made in recent work to address the problem of adaptive definitions of collective variables, most notably using artificial intelligence (AI)-based techniques.[42–46] The new approaches use projections to the principle components of the system or generalize the trajectory-dependent variables to a dimensionless space, which is suitable for supervised machine learning methods. Other approaches employ the Riemannian geometry for a suitable transformation of the trajectory space and the identification of collective variables. Although many of these approaches have been deemed successful for the systems to which they have been applied, the new techniques are mathematically complex, difficult to implement, or require large amounts of suitable training data. In contrast to these very complex methods, trajectory space enhanced sampling methods are based on simple conjectures, which resemble an alternative to very complex AI-driven approaches, while they remain computationally easy to handle and lead to reproducible results. In a recent work, we developed a correlation-dependent enhanced sampling method (CORE-MD I)[47] that only depends on one single energy parameter and does not need *a priori* definitions of reaction coordinates. The CORE-MD I method relies on a projection of the complex energy landscape on the dimensionless space of the path-dependent correlation functions. The CORE-MD I method accelerates the sampling through the formulation of history-dependent correlation-dependent bias potentials and an additional statistical bias, which also depends on the autocorrelation function of the adaptive paths. We successfully validated the method on the folding of TrpCage and the conformational landscape of dialanine.

In this work, we developed and implemented a novel correlation-dependent MD method (CORE-MD II), which is parameter-free and does not require an *a priori* definition of reaction coordinates. Using a kinetic Monte Carlo (kMC) formalism, the method performs a statistical sampling between configurations that have been accessed previously. While the CORE-MD I methodology is based on a correlation-dependent probability, the CORE-MD II method partitions the simulation into sub-trajectories that we refer to as instances. In an adaptive way, a correlation-dependent rate is used as a sampling parameter for the selection of the instances *on the fly*. In contrast to CORE-MD I, the CORE-MD II method uses a path-dependent metadynamics formalism and a statistical bias to accelerate the MD sampling within each instance. The set of techniques that we apply in the CORE-MD II method allow for a faster and more accurate sampling of protein folding than in the CORE-MD I simulations. We validate the CORE-MD II method on the conformational landscape of dialanine and the folding of two peptides: TrpZip2 and TrpCage.[48,49] In the simulations, we observe a good agreement with the experiment and long time equilibrium MD simulations. We find that the novel algorithm is capable of sampling the systems with a high level of convergence compared with equilibrium MD data, while the acceleration factors range from 20 to 120.

II. METHODS

The CORE-MD II method uses a central path-definition and calculates a path-dependent autocorrelation function of the increments along the pathway. Using a correlation-dependent path-definition, a metadynamics algorithm samples the system along adaptive pathways. The CORE-MD II method partitions the total trajectory into instances with independent path-definitions and correlations. The states that are obtained after each instance are sampled using a hybrid kinetic Monte Carlo (kMC) algorithm. The CORE-MD II method can be understood as a non-equilibrium sampling strategy relying on an adaptive rate-dependent selection of short enhanced MD trajectories. The total trajectory consists then of a large number of adaptive and path-dependent non-equilibrium simulations. In these terms, the CORE-MD II method can be interpreted as an adaptive rate-dependent state-to-state dynamics sampling method between Markovian states.[50–52] The hybrid kinetic Monte Carlo/MD (kMC/MD) method does not require any *a priori* information on reaction coordinates or collective variables and does not need additional input parameters.

A. Theory

We start with the definition of the global probability $P(x_i(t))$ that can be defined over N time-slices or sub-trajectories k with length τ_k, which are described by local probability densities $\rho_k(x_i(t))$,

$$P(x_i(t)) = \lim_{N\to\infty} \prod_k^N \rho_k(x_i(t)), \tag{1}$$

where $x_i(t)$ stands for the coordinate of an atom with the index i. As a result, we divide the total trajectory into slices with periods of τ_k, where we observe local pathways $L_{i_k}(t)$ and local correlation functions $C_{i_k}(t)$ (see Fig. 1). In analogy, we express the partition of the total pathway into instances with the index k,

$$L_i(t) = \lim_{N\to\infty} \sum_k^N L_{i_k}(t). \tag{2}$$

If we then consider the averaging process of a trajectory-dependent quantity $X(t)$, the partition into small trajectories allows for a faster formation of time-averages than the determination of the expectation value of the complete trajectory, which is linked to the timescale problem of MD simulations. Therefore, the expectation value of the complete trajectory can be approximated as

$$\langle X(t)\rangle = X(t)P(x_i(t)) \approx X(t)\prod_k^K \rho_k(x_i(t)), \tag{3}$$

which states that the partition of the complete trajectory into a finite number of K sub-trajectories is approximately sufficient for the sampling of the expectation value $\langle X(t)\rangle$. We define the number of configurations K by a minimal set of the number of atoms N_a in the system, which guarantees a fast forward propagation. In the following, we introduce the expressions for the fragmented pathway and the local correlation function. Within each sub-trajectory k, the local pathway is described by the reduced action $L_{i_k}(t)$,

$$L_{i_k}(t) = \sum_{t<\tau_k} p_i(t)\Delta x_i(t), \tag{4}$$

where $\Delta x_i(t) = x_i(t) - \langle x_i(t)\rangle$, $p_i(t)$ is the momentum, and t is the time. The local path $L_{i_k}(t)$ is used to define the local autocorrelation function C_{i_k},[47]

$$C_{i_k}(t) = \frac{1}{\tau_k}\sum_{t\le\tau_k} \frac{(L_{i_k}(t') - \langle L_{i_k}(t)\rangle)(L_{i_k}(t) - \langle L_{i_k}(t)\rangle)}{|L_{i_k}(t') - \langle L_{i_k}(t)\rangle\|L_{i_k}(t) - \langle L_{i_k}(t)\rangle|}, \tag{5}$$

where $L_{i_k}(t')$ is determined with a frequency equal to 1 ps^{-1} and $\langle\cdots\rangle$ denotes the time-average. In our implementation, we define a period $\tau_k(t)$ that separates each individual instance k from the preceding instance [see Fig. 2, where we show an example of the index k and the correlation function $C_{i_k}(t)$ as a function of time in a CORE-MD II simulation of dialanine]. The CORE-MD II algorithm samples the system along a correlation-dependent probability between states with an index k using a kinetic Monte Carlo (kMC) algorithm. We limit the number of kMC configurations by a minimal set of the number of atoms N_a in the system, which guarantees a fast forward propagation within a small window of possible selections in each kMC-step. With a frequency of τ_k^{-1}, we perform a kMC-step and express a rate r_k for each instance k as

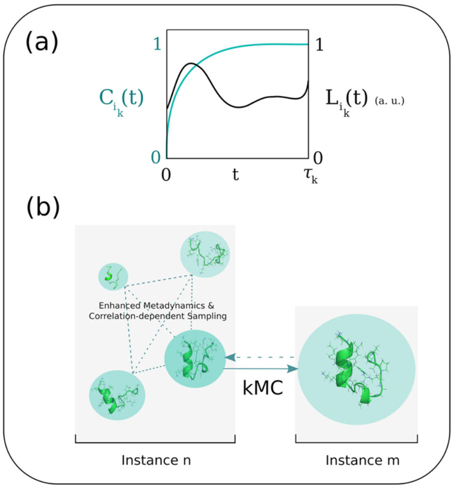

FIG. 1. Schematic description of the CORE-MD II algorithm. The CORE-MD II algorithm relies on two components: (1) a local path-dependent accelerated metadynamics sampling and (2) a kinetic Monte Carlo (kMC) sampling in between the states that are obtained in instances k. (a) Schematic diagram of the local pathway $L_{i_k}(t)$ that is calculated from momenta $p_i(t)$ and positions $\Delta x_i(t)$, which defines a correlation function $C_{i_k}(t)$. In this graph, the correlation function $C_{i_k}(t)$ is displayed on the left y-axis, while the local pathway $L_{i_k}(t)$ is displayed on the right y-axis. The CORE-MD II formalism uses the correlation function $C_{i_k}(t)$ to define the statistical bias-function $\lambda_{i_k}(t)$, the collective variable $\sigma_{i_k}(t)$ for the path-dependent metadynamics simulation and the correlation-dependent time-period $\tau_k(t)$. (b) Schematic description of the state-to-state dynamics between the instances n and m using a kinetic Monte Carlo (kMC) algorithm. Depending on the correlation-dependent rates $r_k(t)$, transitions are sampled in between the states that arise in each instance. (The possible backward transition is indicated by a dashed arrow. The selected kMC-step is indicated by a solid arrow.)

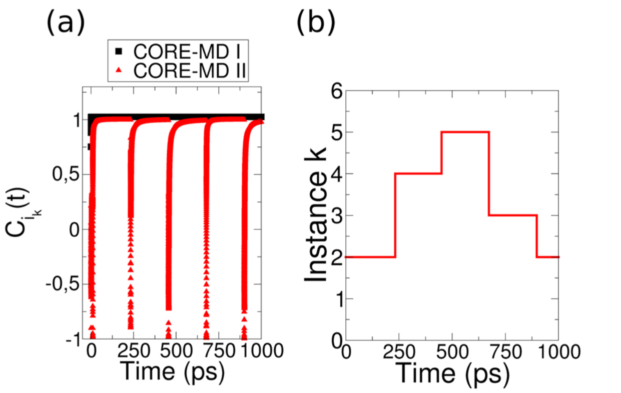

FIG. 2. (a) Correlation function $C_{i_k}(t)$ from a CORE-MD II and CORE-MD I simulation of dialanine as a function of MD-time. (b) Instances k of the CORE-MD II simulation as a function of MD-time. The correlation patterns of the CORE-MD II technique differ from the CORE-MD I technique because the CORE-MD II algorithm sub-divides the trajectory into sub-instances k and performs a kinetic Monte Carlo sampling between the states. The separation of the trajectory into instances k yields an improvement in the sampling of equilibrium properties with an error that is 4.2 times lower than in the CORE-MD I simulation of dialanine.

$$r_k(t) = \nu e^{-\epsilon \Delta E_k(t)}, \tag{6}$$

where ν is a frequency factor (we apply $\nu = N_a \tau_0$, where N_a is the number of atoms and τ_0 is the minimal period equal to 10 ps^{-1}, which is a relation that connects to the friction terms in the prefactors of Kramer's rate theory[53,54]), $\Delta E_k(t) = E_k(t) - E_{k-1}(t)$, and $E_k(t)$ is the energy for the instance k,

$$E_k(t) = E_{\text{pot}_k}(t) + V_k(t), \tag{7}$$

where $E_{\text{pot}_k}(t)$ stands for the potential energy in the instance k, $V_k(t)$ is the bias potential, and $\epsilon = \frac{1}{RT}$, where $R = 8.314\ \frac{J}{K\ \text{mol}}$ and T stands for the temperature. We then define the period τ_k as

$$\tau_k(t) = \frac{1}{r_k(t)}, \tag{8}$$

which is the timescale for the instance k. We then calculate the cumulative rates $R_k(t) = \sum_{j=1}^{k} r_j(t)$ and $R_N(t) = \sum_{j=1}^{N} r_j(t)$ and apply the kinetic Monte Carlo algorithm[36,55,56] for the selection of a configuration k with which a configuration is used for the subsequent trajectory instance,

$$R_{k-1}(t) < R_N(t) \times \xi \le R_k(t), \tag{9}$$

where ξ stands for a random number ranging from 0 to 1. The kMC sampling guides the trajectory between equilibrium configurations of the system, where each instance k resembles a state that resides close to the equilibrium state. We continue with the description of the second component of the CORE-MD II algorithm that applies the local biases. (1) At each initialization of a new trajectory-fragment, the velocities are selected from a random distribution. (2) In order to accelerate the sampling within each instance, we apply a history-dependent bias potential $V_{i_k}(t)$ that is related to metadynamics,[22] while the history dependency is limited by the timescale of each instance. We consider the fragmented pathway that depends on the correlation function and express that the expectation value of a reaction coordinate $\langle \sigma_{i_k}(t) \rangle$ equals the sum over all local path-increments of individual instances k,

$$\langle \sigma_{i_k}(t) \rangle \sim \sum_{k}^{K} L_{i_k}(t)(1 + \beta_k C_{i_k}(t)) e^{-\beta_k C_{i_k}(t)}, \tag{10}$$

where β_k is a normalization factor ranging from 0 to 1. Therefore, we define a bias potential consisting of an accumulation of Gaussian functions along a collective variable $\sigma_{i_k}(t)$ that we define through the correlated path,[57]

$$\sigma_{i_k}(t) = L_{i_k}(t)(1 + \beta_k C_{i_k}(t)) e^{-\beta_k C_{i_k}(t)}, \tag{11}$$

which we normalize by the maximal correlated path occurring within each instance. The history-dependent potential $V_{i_k}(t)$ is accumulated with a frequency of 1 ps^{-1},

$$V_{i_k}(t) = -W \sum_{t' \le t} \exp\left\{ -\left[\frac{\sigma_{i_k}(t) - \sigma_{i_k}(t')}{\delta\sigma} \right]^2 \right\}, \tag{12}$$

where W is the height of the Gaussian, which we determine using $W = 10 k_B T \frac{1\ \text{ps}}{\tau_k(t)}$, and $\delta\sigma$ is the width of the Gaussians, which we set equal to the width of 2 bins in the histogram (consisting of 100 bins in our implementation). We add the Gaussians to the history-dependent potential using the well-tempered metadynamics technique through a normalization of the added Gaussians by the factor $\exp(-V_{i_k}(t)/\Delta T)$, where we apply $\Delta T = 1000$ kJ/mol.[58] The corresponding bias $\frac{\partial}{\partial \sigma_{i_k}(t)} V_{i_k}(t)$ is added throughout the simulation. Finally, we accelerate the sampling within each instance and apply the statistical bias as described in our recent work on the CORE-MD algorithm.[47] We implemented the correlation-dependent bias by a factorization with the variable $\lambda_{i_k}(t)$ with which we scale the gradient of all atoms in the system. The factor $\lambda_{i_k}(t)$ is described by

$$\lambda_{i_k}(t) = (1 + \beta_k C_{i_k}(t)) e^{-\beta_k C_{i_k}(t)}\ . \tag{13}$$

This statistical bias enhances the decay of the correlation function and accelerates the access of new states by the system.[47]

B. Algorithm

- Calculate the path using expression (4) for each time step and evaluate the correlation function given in (5).
- Accumulate the history-dependent potential $V_{i_k}(t)$ as described in Eq. (12) and apply the statistical bias from (13).
- With a frequency of τ_k^{-1}, apply the kinetic Monte Carlo formalism and re-initialize the correlation function:
    - Using the kMC formalism described in the expression (9), select a new configuration and re-initialize a new sub-trajectory, while the path-dependent quantities and the bias potential $V_{i_k}(t)$ are set to values equal zero. Assign the present configuration to the array of configurations and determine the rate r_k.

C. Simulation parameters and system setup

For the simulations and parts of the trajectory analysis, we used the GROMACS-4.5.5 simulation package.[59] We implemented the CORE-MD II method into the same package. In all simulations, we used the AMBER99SB forcefield[60,61] and the generalized Born implicit solvent model using the Still algorithm with a continuum dielectric constant equal to 80.[62] The electrostatics and van der Waals interactions were treated using the twin-range cutoff equal to 1.0/1.2 nm with a neighborlist cutoff equal to 1.0 nm. The neighborlist was updated every integration step. We applied the Nosé–Hoover thermostat with a coupling time of $\tau_T = 1.0$ ps and a target temperature of 300 K. We modeled the starting structures in an extended conformation using the ribosome code that we downloaded from *http://folding.chemistry.msstate.edu/raj/Manuals/ribosome.html*. In validation simulations, we modeled dialanine (Ace-Ala-NMe), tryptophan cage minipeptide (TrpCage) (NLYIQWLKDGG-PSSGRPPPS),[49] and tryptophan zypper peptide 2 (TrpZip2) (SWTWENGKWTWKX).[48] We capped both peptides N- and C-terminal with an acetyl- and a methyl-group. For the comparison of the CORE-MD II results, we applied the CORE-MD I algorithm in simulations of TrpCage, TrpZip2, and dialanine using a parameter $\alpha = 1.0$ kJ/mol.[47] For the generation of equilibrium MD data, we ran an equilibrium MD simulation over 2 μs of dialanine and parallel tempering replica exchange MD (REMD) simulation of TrpCage and TrpZip2 over 1 μs using 24 replicas in the NVT ensemble within a temperature range from 300 to 396 K.[63] An exchange of configurations was attempted every 1000 integration steps in the conventional REMD simulation. We ran a total simulation time of 200 ns for dialanine using CORE-MD I/II and simulated the TrpCage and TrpZip2 minipeptides over 200 ns using the same algorithms. For the calculation of the root-mean square deviation of the backbone atoms Cα to the native structure ($RMSD_{C\alpha-C\alpha}$), we used the NMR-model No. 1 from the protein data bank (PDB) structure: 1l2y (TrpCage).[49] In parts, the trajectory analysis was performed using in-house programs. For the determination of the free energies ΔF, we used

$$\Delta G = -k_B T \ln\left(\frac{P}{P_{\min}}\right), \tag{14}$$

where k_B is Boltzmann's constant, T is the temperature, P is the probability, and $P_{\min}$ is the minimal non-zero reference value of the same function. We determined the level of convergence $\langle \Delta\Delta G \rangle$ by the average difference value of the free energies from equilibrium MD and the CORE-MD I/II simulations. We analyzed the frequency ν of transitions of the Φ-angle of dialanine from values lower than zero to positive values through counting the numbers of transitions N_Φ from negative to positive values and normalizing by the total number of frames N_t,

$$\nu = \frac{N_\Phi}{N_t}. \tag{15}$$

We clustered the structures using $RMSD_{C\alpha-C\alpha}$ to the native structure. We applied an RMSD threshold of ~0.1 nm for the clustering of the structures. For TrpCage, we coarse-grained the total configuration space into eight different clusters, while we divided the conformation space of TrpZip2 into seven clusters. We define the acceleration time by the approximate central processing unit (CPU)-time that is required to sample the identical free energy partition in relation to the CPU-time in conventional MD and REMD simulations.

D. Program

The CORE-MD II simulation code is implemented into the GROMACS-4.5.5 simulation package. The code is available at www.github.com/epeter455/.

III. RESULTS AND DISCUSSION

A. Simulations of dialanine

We validated the CORE-MD II algorithm on the dialanine system and compared our results with long time equilibrium MD and with results from a CORE-MD I simulation. In Fig. 3, we show the free energy landscapes as a function of dihedral angles Φ and Ψ (FEL$_{\Phi-\Psi}$) and our results related to the dynamical behavior of dialanine. In the following, we define the regions along FEL$_{\Phi-\Psi}$ as follows: C7$_{eq}$ ($-180^\circ < \Phi < -160^\circ$, $135^\circ < \Psi < 170^\circ$) [panels (1) and (2)], α ($-110^\circ < \Phi < -60^\circ$, $-50^\circ < \Psi < 5^\circ$) [panel (5)], and C7$_{ax}$ ($40^\circ < \Phi < 70^\circ$, $-5^\circ < \Psi < 40^\circ$) [panel (4)]. We define the interfacial regions $C7_{eq} \between C7_{ax}$ [panel (3)] ($-20^\circ < \Phi < 30^\circ$, $\Psi \approx 90^\circ$), $\alpha \between C7_{ax}$ [panel (6)] ($-30^\circ < \Phi < 30^\circ$, $\Psi \approx -80^\circ$), and $\alpha \between C7_{eq}$ [panel (7)] ($-60^\circ < \Phi < -90^\circ$, $\Psi \approx -80^\circ$). In the FEL$_{\Phi-\Psi}$ averaged over 2 μs equilibrium MD, we find major minima at the C7$_{eq}$ position [(1) and (2)] (-11 to -12 k_BT) at the region α (5) (-11 to -12 k_BT) and the C7$_{ax}$ position (4) (-6 to -8 k_BT) [see Fig. 3(a)]. The histogram of the CORE-MD I simulation is widened at the interfacial regions [(6), (7), and (3)] by $\sim 10\circ$ to $20\circ$, while $C7_{eq}$ and the α-positions are approximately identical to the equilibrium MD result at -11 to -12 k_BT [see Figs. 3(e) and 3(g)]. We find a larger deviation from the equilibrium MD result at the $C7_{ax}$ position (4), where the minimum is widened and the regions for values of Ψ above and below position (4) are populated with approximately -8 k_BT. In the CORE-MD II result, we find populations at (6), (7), and (3), with energy values ranging from -11 to -12 k_BT. At the C7$_{ax}$ position (4), we observe energy values of -6 k_BT, which are approximately equivalent to the equilibrium MD result [see Figs. 3(b) and 3(g)]. We then measured the free energy differences $\Delta\Delta G_{\Phi-\Psi}$ between the simulations using CORE-MD I/II and the equilibrium MD result. In a comparison between CORE-MD I and II, we find that the minima at the $C7_{eq}$ position (1) and (2) are shifted only in the CORE-MD I result, where we find a positive shift of ~0.2 k_BT [see Figs. 3(c), 3(f), and 3(g)]. At that position, the CORE-MD II result agrees very well with equilibrium MD. We observe an identical behavior for the α-position (5), while we find the largest differences at the interfaces at (3), (6), and (7), where we observe deviations of up to $-3.5k_BT$ in the CORE-MD I simulation. In contrast, the CORE-MD II result shows almost no deviation. At the $C7_{ax}$ position (4), we observe that the CORE-MD I result is up to 0.5–1k_BT lower compared with equilibrium MD. The CORE-MD II simulation result shows approximately identical energy values in the sampling of the $C7_{ax}$ position (4). We then looked at the average deviation in energy $\langle \Delta\Delta G_{\Phi-\Psi} \rangle$, where we find a value of -0.67 k_BT for CORE-MD I and $-0.16k_BT$ for the CORE-MD II simulation. We further investigated the comparatively strong improvement in the CORE-MD II simulation and measured the autocorrelation function of the Φ-angle $C_\Phi(t)$ for all simulations [see Figs. 4(a), 4(b), and 4(g)]. In the equilibrium MD simulation, the value of $C_\Phi(t)$ decays from 1 to a value of 0.7 and resides at this value up to a lag time of 1 μs. The correlation function $C_\Phi(t)$ of the CORE-MD I simulation indicates a less correlated behavior with average values of $C_\Phi(t)$ in the range from 0.6 to 0.65. In contrast, the CORE-MD II simulation shows a higher Φ-correlation, where $C_\Phi(t)$ remains at values ranging from 0.72 to 0.73. This indicates that the CORE-MD II approach enhances correlation effects within dialanine, in contrast to the CORE-MD I technique. If we consider the average correlation $\langle C_\Phi(t) \rangle$ and compare the different approaches, the correlation behavior of the CORE-MD II simulation agrees better with equilibrium MD than the CORE-MD I simulation. CORE-MD I yields the highest transition rate of the Φ-angle with $\nu = 4 \times 10^{-4}$ ps^{-1}, while the transition rates for CORE-MD II reside at $\sim 2 \times 10^{-4}$ ps^{-1} [see Fig. 3(d)]. In a comparison with the equilibrium MD result, we observe an acceleration factor of ~20 for the CORE-MD II algorithm, while the level of convergence of the free energy landscapes is 4.2 times higher than the CORE-MD I algorithm.

In the validation simulations on dialanine, we compared 2 μs equilibrium MD with CORE-MD I/II simulations. The CORE-MD II REMD simulation shows an optimal sampling behavior with a level of convergence that is 4.2 higher than in the CORE-MD I simulation, while we observe an acceleration factor of ~20.

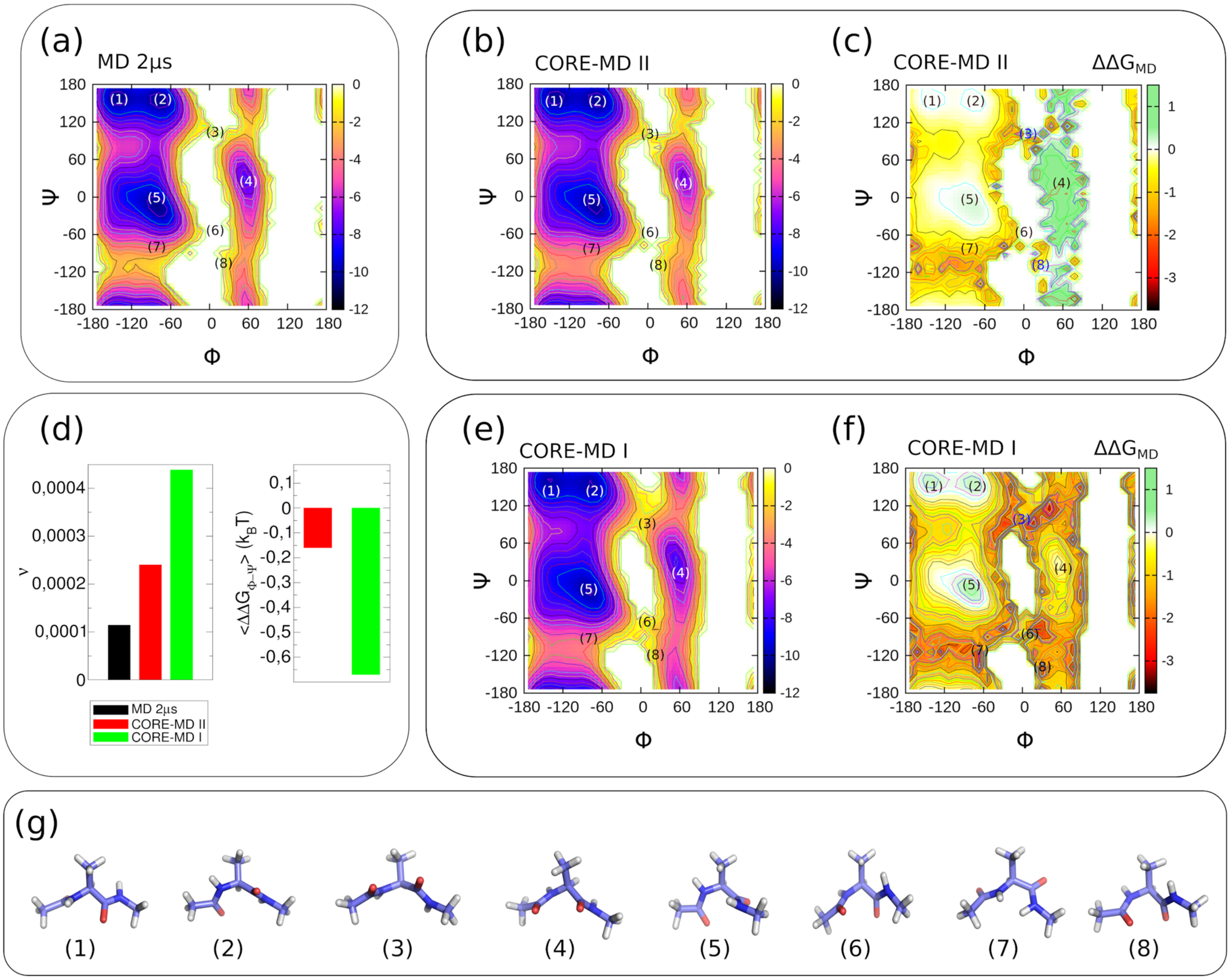

FIG. 3. Results from validation simulations on the dialanine system using 2 μs equilibrium MD, CORE-MD I,[47] and CORE-MD II simulations over 200 ns. (a) Free energy landscape of dialanine as a function of dihedral angles Φ and Ψ (FEL$_{\Phi-\Psi}$) averaged over 2 μs MD (units on all color bars are given in k_BT). (b) FEL$_{\Phi-\Psi}$ averaged over a simulation of dialanine over 200 ns using CORE-MD II. (c) $\Delta\Delta G$ difference plot between FEL$_{\Phi-\Psi}$ in the single CORE-MD II simulation of dialanine and 2 μs MD as a function of Φ and Ψ ($\Delta\Delta G_{\Phi-\Psi}$). (d) (Left) Transition frequency ν of the Φ-angle for all simulations and (right) average free energy differences of the CORE-MD I/II simulations and 2 μs equilibrium MD. (e) FEL$_{\Phi-\Psi}$ averaged over a 200 ns CORE-MD I simulation. (f) $\Delta\Delta G_{\Phi-\Psi}$ from the single CORE-MD I simulation and the 2 μs MD result. The different regions in the FEL plots are indicated with numbers (1–8) [conformers are shown in panel (g)]. The results indicate that the CORE-MD II method with a deviation of 0.15k_BT is 4.2 times more sensitive than CORE-MD I ($\langle\Delta G_{\Phi-\Psi}\rangle = -0.67k_BT$). In general, the CORE-MD II technique yields an improved sampling along the Φ-dihedral angle with an acceleration factor of 20 compared to the 2 μs MD simulation, while the CORE-MD I simulation result shows a factor of 38 with a four times lower level of convergence. (g) Representative conformations indexed with the numbers (1–8) from the simulation of dialanine.

The differences between the CORE-MD II and CORE-MD I sampling are given by the separation into instances k, the kMC sampling between the instances, a re-evaluation of the correlation function $C_{i_k}(t)$ at each instance, and the flexible biasing expression (see Fig. 2). In general, the CORE-MD II method samples the free energy landscapes with higher internal correlations in the system than CORE-MD I, where the system is strongly decorrelated. From that correlation behavior, we deduce that the separation of the system into sub-instances k in CORE-MD II yields a more realistic description of the underlying collective variables that guide the system along its reaction pathways.

B. TrpZip2 folding

As a first validation example, we performed 1 μs REMD simulations with 24 replicas and enhanced MD simulations of TrpZip2

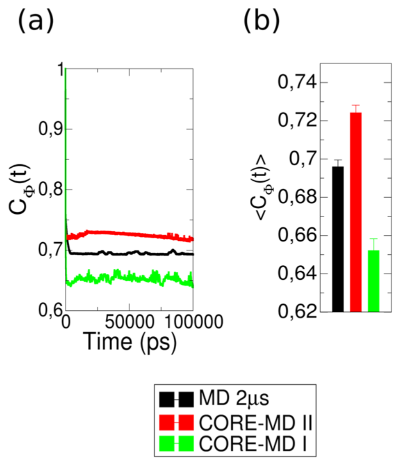

FIG. 4. Autocorrelation function of the Φ-dihedral angle from simulations of dialanine. (a) Autocorrelation function as a function of time. (b) Average autocorrelation value from the different simulations of dialanine using equilibrium MD and the CORE-MD I/II techniques. CORE-MD II shows a higher correlated behavior than the CORE-MD I technique, which explains the higher level of convergence of the CORE-MD II approach.

using CORE-MD I and II (see Fig. 5). We first analyzed $RMSD_{C\alpha-C\alpha}$ to the native structure as a function of simulation time. In the REMD simulation, we observe a fast drop of the RMSD-value from 1.2 nm to $RMSD_{C\alpha-C\alpha}$ of 0.14 nm within the first 1 ps. In this simulation, $RMSD_{C\alpha-C\alpha}$ migrates between extremum values of 0.8 and 0.14 nm with population maxima within 0.3–0.42 and 0.42–0.6 nm [see Fig. 5(a)]. The REMD sampling does not form conformations with $RMSD_{C\alpha-C\alpha}$ values below 0.22 nm, which are only accessed within rare-event fluctuations. The CORE-MD II result on TrpZip2 shows a different behavior [see Fig. 5(b)]. The CORE-MD II simulation forms a hairpin structure within the first 10 ps that remains stable for 13 ns ($RMSD_{C\alpha-C\alpha} \approx 0.22$ nm). The event of the first collapse is followed by a re-opening of the hairpin and a re-orientation of the Trp-sidechains as we find through a visual analysis of the conformers. This reopening is followed by a subsequent collapse at 30 ns, where the hairpin remains stable over the next 5 ns. The remaining CORE-MD II simulation follows the order of hairpin opening and a subsequent closure within a hydrophobic collapse, where the Trp-sidechains reorient. The CORE-MD II simulation result on TrpZip2 is the only trajectory in which we observe the formation of a stable hydrophobic core consisting of four stacked Trp-sidechains[48] [conformation (1), see Fig. 5(j)]. $RMSD_{C\alpha-C\alpha}$ in the CORE-MD I simulation shows a similar fluctuation behavior as the REMD simulation on TrpZip2. Although the fluctuations in the CORE-MD I simulation are the strongest of all three different simulations, the peptide accesses RMSD-values below 0.22 nm only in rare fluctuations, while TrpZip2 mainly resides around 0.3–0.6 nm [see Fig. 5(c)]. Next, we analyzed the free energy landscapes (FEL) as a function of $RMSD_{C\alpha-C\alpha}$ and the radius of gyration R_g for each of the three applied techniques [see Figs. 5(d)–5(f)]. The FEL averaged over 1 μs REMD results in a population ranging from $0.13 < RMSD_{C\alpha-C\alpha} < 0.8$ nm and $0.56 < R_g < 0.92$ nm. We observe minor populations ranging from 0 to -2 k_BT in the range $0.13 < RMSD_{C\alpha-C\alpha} < 0.24$ nm, $0.56 < R_g < 0.8$ nm and $0.7 < RMSD_{C\alpha-C\alpha} < 0.8$ nm, $0.65 < R_g < 0.8$ nm. We find higher populations with energies from -2 to -6 k_BT in the range $0.24 < RMSD_{C\alpha-C\alpha} < 0.4$ nm [conformations (2) and (3)], $0.56 < R_g < 0.8$ nm and $0.6 < RMSD_{C\alpha-C\alpha} < 0.7$ nm, $0.65 < R_g < 0.8$ nm [conformations (4) and (5)] [see Fig. 5(j)]. We locate the maximal population within the range $0.4 < RMSD_{C\alpha-C\alpha} < 0.6$ nm, $0.56 < R_g < 0.8$ nm, where the energy ranges from -8 to -9 k_BT. In contrast to CORE-MD I and the 1 μs REMD simulation, the FEL in the CORE-MD II simulation contains the only minimum in the collapsed state with a stable hydrophobic core [see Fig. 5(e), conformation (1)]. In the CORE-MD II FEL, we observe minor populations ranging from 0 to -4 k_BT in the range $0.11 < RMSD_{C\alpha-C\alpha} < 0.2$ nm, $0.56 < R_g < 0.8$ nm and $0.7 < RMSD_{C\alpha-C\alpha} < 0.9$ nm, $0.9 < R_g < 1.05$ nm. We find higher populations with energies from -2 to -6 k_BT in the range $0.6 < RMSD_{C\alpha-C\alpha} < 0.7$ nm and $0.65 < R_g < 0.8$ nm. We locate two maxima in the population within the range $0.2 < RMSD_{C\alpha-C\alpha} < 0.25$ nm, $0.56 < R_g < 0.8$ nm corresponding to the collapsed native state, and a near native state at $0.4 < RMSD_{C\alpha-C\alpha} < 0.6$ nm, $0.56 < R_g < 0.8$ nm, where the energy ranges from -8 to -9 k_BT. The CORE-MD I result shows the widest population range and no energy minimum corresponding to the native state [see Fig. 5(f)]. In that FEL averaged over the CORE-MD I simulation, we find minor populations with energies ranging from 0 to -2 k_BT within the range $0.13 < RMSD_{C\alpha-C\alpha} < 0.22$ nm, $0.6 < R_g < 0.8$ nm and $0.8 < RMSD_{C\alpha-C\alpha} < 1$ nm, $0.9 < R_g < 1.1$ nm. In that validation example, the population rises toward a main maximum in the range $0.4 < RMSD_{C\alpha-C\alpha} < 0.6$ nm and $0.65 < R_g < 0.8$ nm. We then calculated the relative differences in the free energies between the CORE-MD I/II results and the 1 μs REMD simulation of TrpZip2 [see Figs. 5(g)–5(i)]. For the CORE-MD II result, the differences $\Delta\Delta G$ in the populations range from ~5 to -2.5 k_BT, where the largest differences can be found for radii of gyration R_g below 0.65 nm and above 0.87 nm. In the center of the FEL, we find a good agreement of the CORE-MD II simulation and the REMD result. The CORE-MD I result differs only slightly from the CORE-MD II difference plot [see Fig. 5(i)], where we find approximately the same difference pattern. With -0.038 k_BT, the average deviation $\langle\Delta\Delta G\rangle$ of the CORE-MD II result shows that the CORE-MD II method is 1.4 times more precise than the CORE-MD I sampling with an average deviation of -0.052 k_BT [see Fig. 5(g)]. In a final analysis, we performed a $RMSD_{C\alpha-C\alpha}$-dependent clustering of the conformations in the CORE-MD II simulation [see Fig. 5(j)]. In contrast to the REMD and the CORE-MD I simulation, CORE-MD II samples the formation of the native state with a correct arrangement of the Trp-residues. As a general observation, we find that the folding pathways in the CORE-MD II simulation follow a first hydrophobic collapse, after which a re-ordering of the Trp-sidechain occurs leading to the final formation of the native fold. The CORE-MD II simulation shows that only the near native states can access the native state of TrpZip2, while coiled conformers have to pass through the collapsed state of that peptide toward folding events with a correct hydrophobic core. These results agree with prior simulation studies and the observation of a hydrophobic collapse mechanism of folding of TrpZip2.[64–68] The approximate acceleration factor of the CORE-MD II method compared with the 1 μs REMD simulation is ~120, while the REMD simulation did not sample the formation of the native fold correctly.

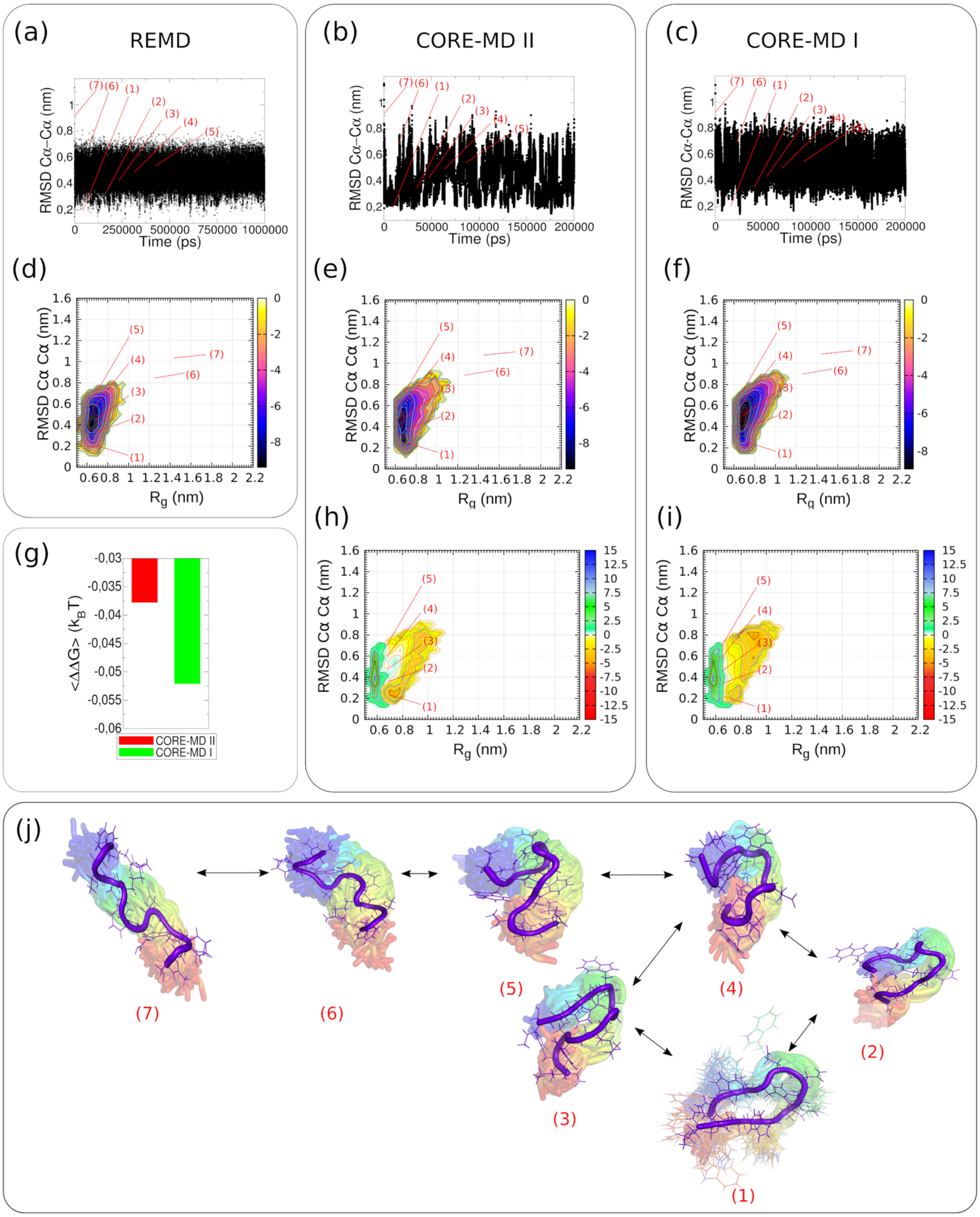

FIG. 5. Results from simulations of the TrpZip2 minipeptide using 1 μs replica exchange MD (REMD) with 24 replicas and CORE-MD I/II simulations over 200 ns. (a) $RMSD_{C\alpha-C\alpha}$ to the native structure (NMR-model No. 1, pdb-code: 1le1[48]) as a function of time from the REMD simulation of TrpZip2. (b) $RMSD_{C\alpha-C\alpha}$ as a function of time from the CORE-MD II simulation of TrpZip2. (c) $RMSD_{C\alpha-C\alpha}$ as a function of time from the CORE-MD I simulation of TrpZip2. (d) Free energy landscape (FEL) as a function of $RMSD_{C\alpha-C\alpha}$ and the radius of gyration R_g averaged over a 1 μs REMD simulation. Energies in the color bar are given in units of k_BT. (e) Free energy landscape (FEL) as a function of $RMSD_{C\alpha-C\alpha}$ and the radius of gyration R_g averaged over a 200 ns CORE-MD II simulation. (f) Free energy landscape (FEL) as a function $RMSD_{C\alpha-C\alpha}$ and the radius of gyration R_g averaged over a 200 ns CORE-MD I simulation. (g) Average free energy difference $\langle \Delta\Delta G \rangle$ to the 1 μs REMD result for the CORE-MD I and the CORE-MD II result. (h) Free energy difference $\Delta\Delta G$ between the CORE-MD II result and 1 μs REMD as a function of $RMSD_{C\alpha-C\alpha}$ and the radius of gyration R_g. (i) Free energy difference $\Delta\Delta G$ between the CORE-MD I result and 1 μs REMD as a function of $RMSD_{C\alpha-C\alpha}$ and the radius of gyration R_g. (j) Kinetic network of $RMSD_{C\alpha-C\alpha}$-dependent clusters obtained from the CORE-MD II simulation of TrpZip2. The cluster indexes are given below each cluster, which are displayed in each of the plots above.

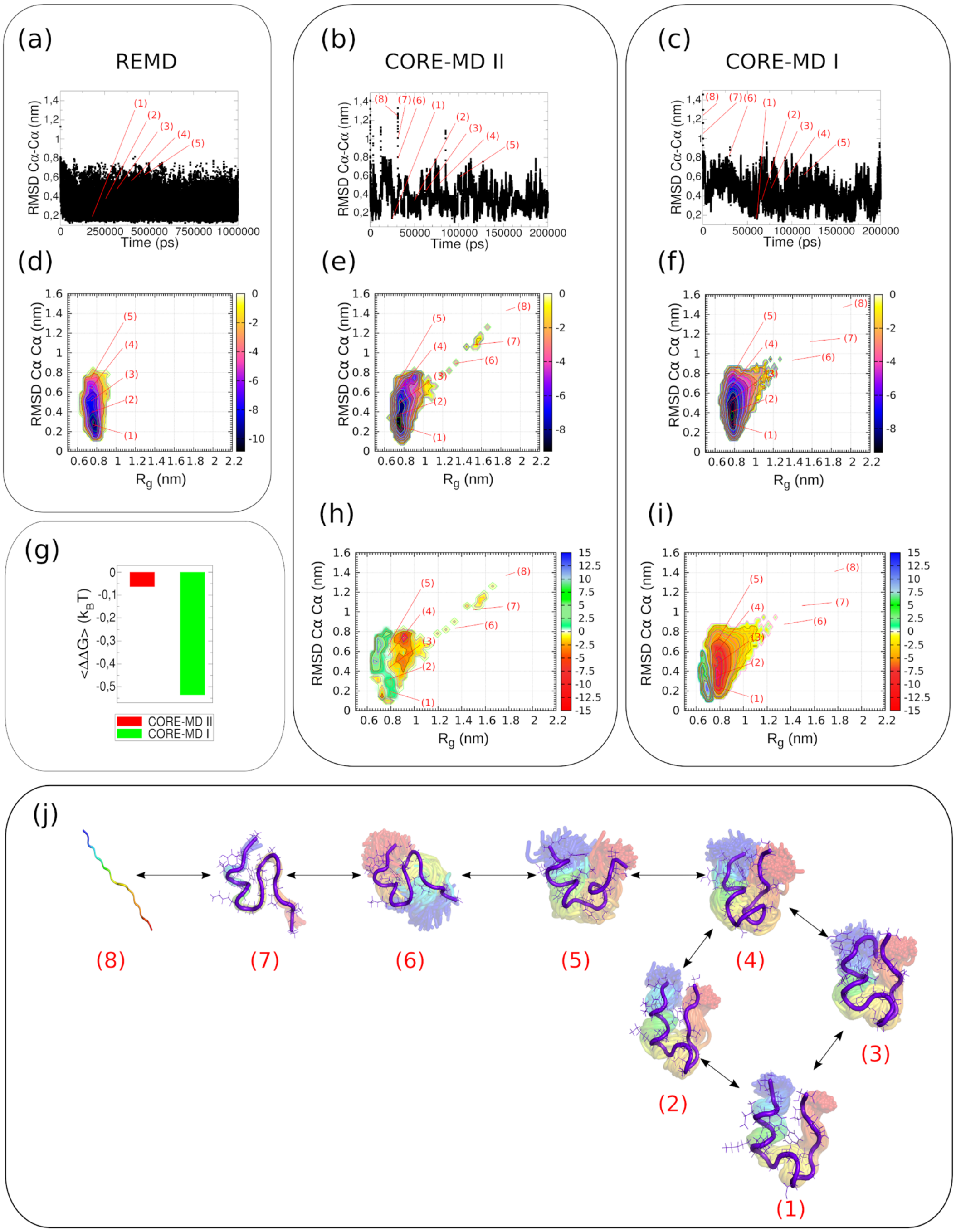

FIG. 6. Results from simulations of the TrpCage minipeptide using 1 μs replica exchange MD (REMD) with 24 replicas and CORE-MD I/II simulations over 200 ns. (a) $RMSD_{C\alpha-C\alpha}$ to the native structure (NMR-model No. 1, pdb-code: 1l2y[49]) as a function of time from the REMD simulation of TrpCage. (b) $RMSD_{C\alpha-C\alpha}$ as a function of time from the CORE-MD II simulation of TrpCage. (c) $RMSD_{C\alpha-C\alpha}$ as a function of time from the CORE-MD I simulation of TrpCage. (d) Free energy landscape (FEL) as a function of $RMSD_{C\alpha-C\alpha}$ and the radius of gyration R_g averaged over a 1 μs REMD simulation. Energies in the color bar are given in units of k_BT. (e) Free energy landscape (FEL) as a function of $RMSD_{C\alpha-C\alpha}$ and the radius of gyration R_g averaged over a 200 ns CORE-MD II simulation. (f) Free energy landscape (FEL) as a function of $RMSD_{C\alpha-C\alpha}$ and the radius of gyration R_g averaged over a 200 ns CORE-MD I simulation. (g) Average free energy difference $\langle\Delta\Delta G\rangle$ to the 1 μs REMD result for the CORE-MD I and the CORE-MD II result. (h) Free energy difference $\Delta\Delta G$ between the CORE-MD II result and 1 μs REMD as a function of $RMSD_{C\alpha-C\alpha}$ and the radius of gyration R_g. (i) Free energy difference $\Delta\Delta G$ between the CORE-MD I result and 1 μs REMD as a function of $RMSD_{C\alpha-C\alpha}$ and the radius of gyration R_g. (j) Kinetic network of $RMSD_{C\alpha-C\alpha}$-dependent clusters obtained from the CORE-MD II simulation of TrpCage. The cluster indexes are given below each cluster, which are displayed in each of the plots above.

C. TrpCage folding

As a second validation example, we performed 1 μs REMD simulations with 24 replicas and enhanced MD simulations of TrpCage using CORE-MD I and II (see Fig. 6). In a first analysis, we measured $RMSD_{C\alpha-C\alpha}$ to the native structure as a function of simulation time. In the REMD simulation, the RMSD-value decreases from 1.6 nm to $RMSD_{C\alpha-C\alpha}$ of 0.2 nm within the first 50 ns. $RMSD_{C\alpha-C\alpha}$ fluctuates between extremum values of 0.7 and 0.13 nm with a population maximum at 0.2–0.32 nm [see Fig. 5(a)]. The CORE-MD II result on TrpCage shows an almost identical behavior [see Fig. 6(b)]. The CORE-MD II simulation forms a helical structure within the first 1–2 ns that remains stable for 9.5 ns ($RMSD_{C\alpha-C\alpha}$ $\approx$ 0.22 nm). The event of the first collapse is followed by a re-opening of the native fold and a re-organization of the sidechain involving Trp6 and Tyr3 as we find through a visual analysis of the conformers. This reopening is followed by a subsequent collapse at 25 ns, where the peptide remains stable over the next 56 ns. The remaining CORE-MD II simulation follows the order of opening and a subsequent closure of the PPII helix and the N-terminal α-helix within a hydrophobic collapse. The CORE-MD II simulation result on TrpCage differs from the CORE-MD I result, where we do not observe the formation of a stable hydrophobic core [see Fig. 6(c)]. $RMSD_{C\alpha-C\alpha}$ in the CORE-MD I simulation shows the strongest fluctuations in contrast to the CORE-MD II simulation. As the fluctuations in the CORE-MD I simulation are the strongest of all three different simulations, the peptide accesses RMSD-values below 0.22 nm only through rare-event fluctuations, while TrpCage mainly resides around 0.3–0.6 nm [see Fig. 6(c)]. Next, we analyzed the free energy landscapes (FEL) as a function of $RMSD_{C\alpha-C\alpha}$ and the radius of gyration R_g for each of the three applied techniques [see Figs. 6(d)–6(f)]. The FEL averaged over 1 μs REMD results in a population ranging from $0.13 < RMSD_{C\alpha-C\alpha} < 0.8$ nm and $0.63 < R_g < 0.92$ nm. We observe minor populations ranging from 0 to −3 k_BT in the range $0.13 < RMSD_{C\alpha-C\alpha} < 0.17$ nm, $0.63 < R_g < 0.85$ nm and $0.7 < RMSD_{C\alpha-C\alpha} < 0.93$ nm, $0.63 < R_g < 0.8$ nm. We find higher populations with energies from −3 to −8 k_BT in the range $0.17 < RMSD_{C\alpha-C\alpha} < 0.2$ nm, $0.63 < R_g < 0.85$ nm and $0.4 < RMSD_{C\alpha-C\alpha} < 0.93$ nm, $0.65 < R_g < 0.8$ nm. We locate the maximal population within the range $0.2 < RMSD_{C\alpha-C\alpha} < 0.3$ nm, $0.63 < R_g < 0.8$ nm, where the energy ranges from −9 to −10.5 k_BT. In agreement with the 1 μs REMD simulation, the FEL in the CORE-MD II simulation contains a single minimum in the collapsed state with a stable hydrophobic core [see Fig. 6(e)]. In the CORE-MD II FEL, we observe minor populations ranging from 0 to −4 k_BT in the range $0.11 < RMSD_{C\alpha-C\alpha} < 0.2$ nm, $0.63 < R_g < 0.9$ nm and $0.6 < RMSD_{C\alpha-C\alpha} < 0.9$ nm, $0.9 < R_g < 1.1$ nm. We find higher populations with energies from −2 to −6 k_BT in the range $0.5 < RMSD_{C\alpha-C\alpha} < 0.8$ nm and $0.65 < R_g < 0.9$ nm. We locate the maximum in the population within the range $0.19 < RMSD_{C\alpha-C\alpha} < 0.42$ nm and $0.63 < R_g < 0.7$ nm corresponding to the collapsed native state and a near native state, where the energy ranges from −8 to −8.5 k_BT. The CORE-MD I result shows the widest population range and an energy minimum corresponding to the native state [see Fig. 6(f)]. In that FEL averaged over the CORE-MD I simulation, we find minor populations with energies ranging from 0 to −2 k_BT within the range $0.11 < RMSD_{C\alpha-C\alpha} < 0.24$ nm, $0.63 < R_g < 0.9$ nm and $0.6 < RMSD_{C\alpha-C\alpha} < 1$ nm, $0.9 < R_g < 1.1$ nm. In that validation example, the population rises toward a main maximum in the range $0.22 < RMSD_{C\alpha-C\alpha} < 0.4$ nm and $0.65 < R_g < 0.8$ nm. We then calculated the relative differences in the free energies between the CORE-MD I/II results and the 1 μs REMD simulation of TrpCage [see Figs. 6(g)–6(i)]. For the CORE-MD II result, the differences $\Delta\Delta G$ in the populations range from ∼5 to −5 k_BT, where the largest differences can be found for radii of gyration R_g below 0.8 nm and above 0.9 nm. In the center of the FEL, we find a good agreement of the CORE-MD II simulation and the REMD result. The CORE-MD I result differs strongly from the CORE-MD II difference plot [see Fig. 6(i)], where we observe that the CORE-MD I result shows $\Delta\Delta G$ values of up to −10 k_BT at R_g values of 0.83 nm. With −0.063 k_BT, the average deviation $\langle\Delta\Delta G\rangle$ of the CORE-MD II result shows that the CORE-MD II method is 8.8 times more precise than the CORE-MD I sampling with an average deviation of −0.53 k_BT [see Fig. 6(g)]. In a final analysis, we performed a $RMSD_{C\alpha-C\alpha}$-dependent clustering of the conformations in the CORE-MD II simulation [see Fig. 6(j)]. As a general observation, we find that the folding pathways in the CORE-MD II simulation follow a first hydrophobic collapse, after which a re-ordering of the N-terminal and the 3_{10}-helical segment occurs leading to the final formation of the native fold [conformations (4) and (5)], while the PPII helical element does not perform a strong internal restructuring. The CORE-MD II simulation shows that only the near native states can access the native state of TrpCage [conformations (2) and (3)], while opened conformers have to pass through the collapsed state of that peptide toward folding events with a correct hydrophobic core [(conformation (1)]. The folding pathways observed in the CORE-MD II validation simulation is dominated by the formation of the secondary structure and an internal reorganization toward the formation of the native fold.[36,69] The CORE-MD I technique samples the formation of the native state correctly, while the fluctuations in the CORE-MD I simulation lead to a stronger relative deviation from the native state as in the CORE-MD II simulation. Related to the total convergence to the folded state, the CORE-MD II algorithm shows an 8.8 times higher convergence than the CORE-MD I technique. The approximate acceleration factor of the CORE-MD II method compared with the 1 μs REMD simulation is ∼120. Our results are in agreement with our previous findings and other theoretical studies.[36,69–78]

IV. CONCLUSIONS

In this paper, we presented a fast and adaptive correlation guided enhanced sampling MD method (CORE-MD II) that raises the performance of the CORE-MD I methodology. The CORE-MD II technique applies a partition of the total pathway into short trajectories that we refer to as instances. Within each instance, the CORE-MD II technique samples independent states using adaptive path-dependent metadynamics. Using a detailed balance criterion, the technique applies a kinetic Monte Carlo (kMC) sampling between the different states that have been accessed in the individual instances. Through the combination of the partition of the total simulation into short non-equilibrium simulations and the kMC sampling, the CORE-MD II method is capable of sampling protein folding in an adaptive and non-parameter-dependent way. In contrast to the CORE-MD I method, the CORE-MD II technique considers the local heterogeneity of correlation patterns and reaction pathways, while the CORE-MD I method applies the global correlation function and the associated probability function. Compared to the CORE-MD I technique, the combination of short path-dependent metadynamics simulations and the kMC sampling of the instances leads to an improvement in the accuracy and the performance of the CORE-MD II algorithm. We applied the CORE-MD II state-to-state dynamics on the dialanine peptide and the folding of two peptides: TrpCage and TrpZip2. In a comparison with long time equilibrium MD and 1 μs REMD simulations, we find that the level of convergence of the CORE-MD II method is up to 8.8 times higher than that of the CORE-MD I method, while the CORE-MD II method reaches acceleration factors of ~120.

ACKNOWLEDGMENTS

This work was supported by the Deutsche Forschungsgemeinschaft (Grant No. MA1081/23-1) (to D.J.M.). D.J.M. is a member of the Cluster of Excellence RESIST (Grant No. EXC 2155) with support from the DFG (Project No. 39 087 428-B11) and of the ERA-Net Joint Project on Rare Diseases Consortium "PredACTINg" with support from the German Federal Ministry of Education and Research under Grant Agreement No. 01 GM1922B. J.E.S. acknowledges support through Grant No. NSF MCB-1716956. Calculations have been carried out at the JUWELS Cluster at the Forschungszentrum Juelich, the Lise Cluster of the Norddeutscher Verbund für Hoch-und Höchstleistungsrechnen (HLRN-IV) (Project No. nib00016), and Nodes of the European Life-science Infrastructure for Biological Information.

The authors have no conflicts to disclose.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] M. Karplus and J. A. McCammon, Nat. Struct. Biol. **9**, 646–652 (2002).
[2] U. H. E. Hansmann and Y. Okamoto, Curr. Opin. Struct. Biol. **9**, 177–183 (1999).
[3] A. Schug, W. Wenzel, and U. H. E. Hansmann, J. Chem. Phys. **122**, 194711 (2005).
[4] P. L. Freddolino, C. B. Harrison, Y. Liu, and K. Schulten, Nat. Phys. **6**, 751–758 (2010).
[5] J. Nasica-Labouze, P. H. Nguyen, F. Sterpone, O. Berthoumieu, N.-V. Buchete, S. Coté, A. De Simone, A. J. Doig, P. Faller, A. Garcia, A. Laio, M. S. Li, S. Melchionna, N. Mousseau, Y. Mu, A. Paravastu, S. Pasquali, D. J. Rosenman, B. Strodel, B. Tarus, J. H. Viles, T. Zhang, C. Wang, and P. Derreumaux, Chem. Rev. **115**, 3518–3563 (2015).
[6] M. Carballo-Pacheco and B. Strodel, J. Phys. Chem. B **120**, 2991–2999 (2016).
[7] B. J. Grant, A. A. Gorfe, and J. A. McCammon, Curr. Opin. Struct. Biol. **20**, 142–147 (2010).
[8] J. D. Durrant and J. A. McCammon, BMC Biol. **9**, 71 (2011).
[9] A. C. Pan, D. Jacobson, K. Yatsenko, D. Sritharan, T. M. Weinreich, and D. E. Shaw, Proc. Natl. Acad. Sci. U. S. A. **116**, 4244–4249 (2019).
[10] J. Loschwitz, O. O. Olubiyi, J. S. Hub, B. Strodel, and C. S. Poojari, "Chapter seven-computer simulations of protein–membrane systems," Prog. Mol. Biol. Transl. Sci. **170**, 273–403 (2020).
[11] J. L. Klepeis, K. Lindorff-Larsen, R. O. Dror, and D. E. Shaw, Curr. Opin. Struct. Biol. **19**, 120–127 (2009).
[12] H. A. Scheraga, M. Khalili, and A. Liwo, Annu. Rev. Phys. Chem. **58**, 57–83 (2007).
[13] S. J. Marrink and D. P. Tieleman, Chem. Soc. Rev. **42**, 6801–6822 (2013).
[14] P. Liu and G. A. Voth, J. Chem. Phys. **126**, 045106 (2007).
[15] A. Irbäck and S. Mohanty, J. Comput. Chem. **27**, 1548–1555 (2006).
[16] J. C. Phillips, D. J. Hardy, J. D. C. Maia, J. E. Stone, J. V. Ribeiro, R. C. Bernardi, R. Buch, G. Fiorin, J. Hénin, W. Jiang, R. McGreevy, M. C. R. Melo, B. K. Radak, R. D. Skeel, A. Singharoy, Y. Wang, B. Roux, A. Aksimentiev, Z. Luthey-Schulten, L. V. Kalé, K. Schulten, C. Chipot, and E. Tajkhorshid, J. Chem. Phys. **153**, 044130 (2020).
[17] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten, J. Comput. Chem. **26**, 1781–1802 (2005).
[18] B. R. Brooks, C. L. Brooks, A. D. MacKerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus, J. Comput. Chem. **30**, 1545–1614 (2009).
[19] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods, J. Comput. Chem. **26**, 1668–1688 (2005).
[20] G. M. Torrie and J. P. Valleau, J. Comput. Phys. **23**, 187–199 (1977).
[21] D. Hamelberg, J. Mongan, and J. A. McCammon, J. Chem. Phys. **120**, 11919 (2004).
[22] A. Laio and M. Parrinello, Proc. Natl. Acad. Sci. U. S. A. **99**, 12562–12566 (2002).
[23] X. Kong and C. L. Brooks III, J. Chem. Phys. **105**, 2414 (1996).
[24] J. L. Knight and C. L. Brooks III, J. Chem. Theory Comput. **7**, 2728–2739 (2011).
[25] J. Comer, J. C. Gumbart, J. Hénin, T. Lelièvre, A. Pohorille, and C. Chipot, J. Phys. Chem. B **119**, 1129–1151 (2015).
[26] A. F. Voter, Phys. Rev. Lett. **78**, 3908 (1997).
[27] H. Grubmüller, Phys. Rev. E **52**, 2893 (1995).
[28] T. Huber, A. E. Torda, and W. F. van Gunsteren, J. Comput.-Aided Mol. Des. **8**, 695–708 (1994).
[29] R. Elber, J. Chem. Phys. **144**, 060901 (2016).
[30] Q. Ma and J. A. Izaguirre, Multiscale Model. Simul. **2**, 1–21 (2003).
[31] R. R. Gabdoulline and R. C. Wade, Methods **14**, 329–341 (1998).
[32] K. Druart, J. Bigot, E. Audit, and T. Simonson, J. Chem. Theory Comput. **12**, 6035–6048 (2016).
[33] J. Hu, A. Ma, and A. R. Dinner, J. Comput. Chem. **27**, 203–216 (2006).
[34] Y. Chen and B. Roux, J. Chem. Theory Comput. **11**, 3572–3583 (2015).
[35] D. Suh, B. K. Radak, C. Chipot, and B. Roux, J. Chem. Phys. **148**, 014101 (2018).
[36] E. K. Peter and J.-E. Shea, Phys. Chem. Chem. Phys. **16**, 6430–6440 (2014).
[37] G. Kabbe, C. Wehmeyer, and D. Sebastiani, J. Chem. Theory Comput. **10**, 4221–4228 (2014).
[38] B. Rennekamp, F. Kutzki, A. Obarska-Kosinska, C. Zapp, and F. Gräter, J. Chem. Theory Comput. **16**, 553–563 (2020).
[39] E. Peter, B. Dick, and S. A. Baeurle, J. Chem. Phys. **136**, 124112 (2012).
[40] M. Miao, H. Fu, H. Zhang, X. Shao, C. Chipot, and W. Cai, Mol. Simul. **47**, 390–394 (2020).
[41] G. Ciccotti and M. Ferrario, Entropy **16**, 233–257 (2014).
[42] J. Smiatek and A. Heuer, J. Comput. Chem. **32**, 2084–2096 (2011).
[43] L. Donati and B. G. Keller, J. Chem. Phys. **149**, 072335 (2018).
[44] F. Noé, G. De Fabritiis, and C. Clementi, Curr. Opin. Struct. Biol. **60**, 77–84 (2020).
[45] J. A. Morrone, A. Perez, J. MacCallum, and K. A. Dill, J. Chem. Theory Comput. **13**, 870–876 (2017).
[46] J. M. Jumper, N. F. Faruk, K. F. Freed, and T. R. Sosnick, PLoS Comput. Biol. **14**, e1006578 (2018).
[47] E. K. Peter, J.-E. Shea, and A. Schug, J. Chem. Phys. **153**, 084114 (2020).
[48] A. G. Cochran, N. J. Skelton, and M. A. Starovasnik, Proc. Natl. Acad. Sci. U. S. A. **98**, 5578–5583 (2001).
[49] J. W. Neidigh, R. M. Fesinmeyer, and N. H. Andersen, Nat. Struct. Biol. **9**, 425–430 (2002).

[50] J. D. Chodera and F. Noé, Curr. Opin. Struct. Biol. **25**, 135–144 (2014).

[51] U. Sengupta and B. Strodel, Philos. Trans. R. Soc., B **373**, 20170178 (2018).

[52] V. S. Pande, K. Beauchamp, and G. R. Bowman, Methods **52**, 99–105 (2010).

[53] H. A. Kramers, Physica **7**, 284 (1940).

[54] P. Haenggi, P. Talkner, and M. Borkovec, Rev. Mod. Phys. **62**, 251–332 (1990).

[55] A. B. Bortz, M. H. Kalos, and J. L. Lebowitz, J. Comput. Phys. **17**, 10–18 (1975).

[56] D. T. Gillespie, J. Comput. Phys. **22**, 403–434 (1976).

[57] E. K. Peter, J. Chem. Phys. **147**, 214902 (2017).

[58] M. Bonomi and M. Parrinello, Phys. Rev. Lett. **104**, 190601 (2010).

[59] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, J. Chem. Theory Comput. **4**, 435–447 (2008).

[60] P. A. Kollman, Acc. Chem. Res. **29**, 461–469 (1996).

[61] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, Proteins **65**, 712–725 (2006).

[62] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson, J. Am. Chem. Soc. **112**, 6127–6129 (1990).

[63] K. Hukushima and K. Nemoto, J. Phys. Soc. Jpn. **65**, 1604–1608 (1996).

[64] Y. Xiao, C. Chen, and Y. He, Int. J. Mol. Sci. **10**, 2838–2848 (2009).

[65] C. Chen and Y. Xiao, Bioinformatics **24**, 659–665 (2008).

[66] G. H. Zerze, B. Uz, and J. Mittal, Proteins **83**, 1307–1315 (2015).

[67] J. Juraszek and P. G. Bolhuis, J. Phys. Chem. B **113**, 16184–16196 (2009).

[68] T. Wu, R. Zhang, H. Li, L. Yang, and W. Zhuang, J. Chem. Phys. **140**, 055101 (2014).

[69] H. Meuzelaar, K. A. Marino, A. Huerta-Viga, M. R. Panman, L. E. J. Smeenk, A. J. Kettelarij, J. H. van Maarseveen, P. Timmerman, P. G. Bolhuis, and S. Woutersen, J. Phys. Chem. B **117**, 11490–11501 (2013).

[70] L. Qiu, S. A. Pabit, A. E. Roitberg, and S. J. Hagen, J. Am. Chem. Soc. **124**, 12952–12953 (2002).

[71] H. Neuweiler, S. Doose, and M. Sauer, Proc. Natl. Acad. Sci. U. S. A. **102**, 16650–16655 (2005).

[72] R. M. Culik, A. L. Serrano, M. R. Bunagan, and F. Gai, Angew. Chem. **123**, 11076–11079 (2011).

[73] J. Juraszek and P. G. Bolhuis, Proc. Natl. Acad. Sci. U. S. A. **103**, 15859–15864 (2006).

[74] J. Juraszek and P. G. Bolhuis, Biophys. J. **95**, 4246–4257 (2008).

[75] F. Marinelli, F. Pietrucci, A. Laio, and S. Piana, PLoS Comput. Biol. **5**, e1000452 (2009).

[76] C. D. Snow, B. Zagrovic, and V. S. Pande, J. Am. Chem. Soc. **124**, 14548–14549 (2002).

[77] H. Ren, Z. Lai, J. D. Biggs, J. Wang, and S. Mukamel, Phys. Chem. Chem. Phys. **15**, 19457–19464 (2013).

[78] R. Zhou, Proc. Natl. Acad. Sci. U. S. A. **100**, 13280–13285 (2003).