The Journal of Chemical Physics

# CORE-MD, a path correlated molecular dynamics simulation method

Cite as: J. Chem. Phys. **153**, 084114 (2020); https://doi.org/10.1063/5.0015398
Submitted: 27 May 2020 . Accepted: 10 August 2020 . Published Online: 26 August 2020

Emanuel K. Peter, Joan-Emma Shea, and Alexander Schug

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

# CORE-MD, a path correlated molecular dynamics simulation method

Cite as: J. Chem. Phys. 153, 084114 (2020); doi: 10.1063/5.0015398
Submitted: 27 May 2020 • Accepted: 10 August 2020 •
Published Online: 26 August 2020



Emanuel K. Peter,[1,a)] Joan-Emma Shea,[2] and Alexander Schug[1,3,b)]

## AFFILIATIONS

[1]John von Neumann Institute for Computing and Julich Supercomputing Centre, Institute for Advanced Simulation, Forschungszentrum Jülich, Jülich, Germany
[2]Department of Chemistry and Biochemistry, Department of Physics, University of California, Santa Barbara, Santa Barbara, California 93106, USA
[3]Faculty of Biology, University of Duisburg-Essen, Duisburg, Germany

**Note:** This paper is part of the JCP Special Topic on Classical Molecular Dynamics (MD) Simulations: Codes, Algorithms, Force Fields, and Applications.
[a)]**Electronic mail:** e.peter@fz-juelich.de
[b)]**Author to whom correspondence should be addressed:** al.schug@fz-juelich.de

## ABSTRACT

We present an enhanced Molecular Dynamics (MD) simulation method, which is free from the requirement of *a priori* structural information of the system. The technique is capable of folding proteins with very low computational effort and requires only an energy parameter. The path correlated MD (CORE-MD) method uses the autocorrelation of the path integral over the reduced action and propagates the system along the history dependent path correlation. We validate the new technique in simulations of the conformational landscapes of dialanine and the TrpCage mini-peptide. We find that the novel method accelerates the sampling by three orders of magnitude and observe convergence of the conformational sampling in both cases. We conclude that the new method is broadly applicable for the enhanced sampling in MD simulations. The CORE-MD algorithm reaches a high accuracy compared with long time equilibrium MD simulations.



## I. INTRODUCTION

The process of protein folding evolves along a multi-dimensional funneled energy landscape. The sequence–structure relationship of a protein is important for the development of novel drug molecules and an understanding of biological or biochemical processes.[1] Molecular Dynamics (MD) simulations are a useful tool for the investigation of the kinetics and thermodynamics of protein folding on a molecular level. However, biologically relevant timescales can range to seconds, or orders of magnitude bigger[2] MD simulations cannot readily reach the timescales necessary for a convergent sampling of that process despite simulations where specific hardware and software are applied.[3] One option is using a coarse-grained description of the biomolecular system to allow sufficient sampling even of large-scale conformational transitions such as those found during folding.[4–7] One can also forego the system dynamics and use Monte Carlo based sampling to obtain free energy landscapes.[8,9] In MD, significant progress has been achieved in enhanced sampling methodologies capable of sampling protein folding landscapes more efficiently. In general, a large number of enhanced sampling MD methods differ in the definitions of the reaction coordinate for the accelerated propagation of the system. The methods can be roughly distinguished between projections in the trajectory space and the space of energetic or geometric degrees of freedom of a biomolecular system. Umbrella sampling methods act on energetic and/or geometric degrees of freedom,[10] where we mention Wang–Landau sampling,[11] meta- and hyperdynamics,[12,13] conformational flooding,[14] local elevation,[15] and energy landscape paving techniques[9,16] that belong to that group. Recent developments couple umbrella sampling techniques with machine learning approaches.[17–19] Other techniques that are part of umbrella sampling methods are variational bias optimization,[20] adaptive

bias reweighting methodologies,[21] frequency adaptive metadynamics,[22] and biasing techniques combined with Bayesian inference.[23–25] Although a large number of enhanced sampling methodologies exist, the generalization of collective variables for the enhanced sampling is still a growing field of research.[21,23–28] An alternative group of methods that accelerate the sampling along the underlying trajectory space have the advantage that they act adaptively and do not require system dependent information.[29] The first developments in that group of methods are bonded and non-bonded constraints, which are capable of reaching twofold to fourfold accelerations.[30–33] Langevin damping, energy based techniques, and a whole group of multiple time stepping methods allowed a 10-fold to 15-fold acceleration of MD simulations.[34–37] More advanced techniques included transition path-sampling techniques, biased path-sampling, milestoning, and weighted ensemble techniques that apply a statistical biasing between the initial and product states or depend on projections onto order parameters related to the underlying free energy landscape (FEL).[38–42]

In this article, we present a novel technique that applies a correlation dependent formalism. The CORE-MD method uses a correlation dependent probability density extracted from the history of the system. The resulting bias is derived adaptively from correlation functions of the path integral over the reduced action. We implemented the path correlated MD (CORE-MD) technique in the GROMACS package.[43] We validate the algorithm on dialanine and a folding simulation of the TrpCage mini-peptide, which have been studied extensively in the literature.[8,9,12,38,44–60] We observe that the novel method accelerates the sampling by many orders of magnitude and observe convergence of the conformational sampling in both cases.

## II. METHODS

The presented method applies a correlation function formalism to the Molecular Dynamics (MD) propagator. In the derivation of our method, we define a propagator function and a correlation dependent probability density function $\rho$. From the history dependent probability density $\rho$, we derive the correlation dependent bias (see Fig. 1). Therefore, the method determines the correlation $C(t)$ from the path integral over momenta $p$ and displacements $dq$. Subsequently, a correlation dependent potential defines the global likelihood gradient depending on the path correlation. The method is applicable with one single additional energy parameter $\alpha$ and does not require *a priori* structural information.

### A. Theory

We start with the definition of the path over the reduced action $L_i$ as a function of the momenta $p_i$ and coordinates $q_i$ for an atom with the index $i$,[58,59,61–63]

$$L_i = \oint p_i dq_i. \tag{1}$$

For the calculation of this integral, we apply a summation over the discretized path over momenta and displacements along the trajectory. For the calculation of the momentum $p_i = m_i v_i$, we use a uniform atomic mass. Then, we define the path-dependent correlation function $C_i(t)$ as

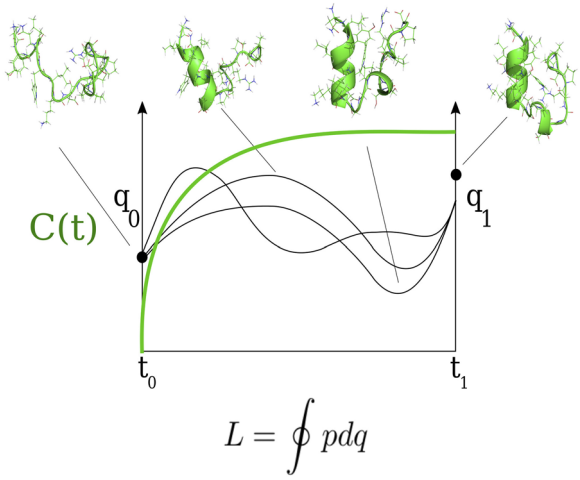


**FIG. 1**. Schematic description of a correlation function $C(t)$ of pathways $L$ of a protein from an initial $q_0$ at a time $t_0$ to a final state $q_1$ at $t_1$. The autocorrelation of the adaptive path is used for the determination of the bias in our simulation using a correlation dependent functional $\rho$.

$$C_i(t) = \frac{1}{\tau}\sum_{t\le\tau}\frac{(L_i' - \langle L_i\rangle)(L_i - \langle L_i\rangle)}{|L_i' - \langle L_i\rangle\|L_i - \langle L_i\rangle|}, \tag{2}$$

where $\langle\cdots\rangle$ denotes the time average and $L_i'$ is determined at a time $t'$ with a probability $\mathcal{P}_i(t')$,

$$\mathcal{P}_i(t') = \frac{1}{1 + e^{-C_i(t')}}, \tag{3}$$

at every time step. In the next step, we discretize the space of the correlation function into a histogram ranging from −1 to +1 and define a probability density $\rho_i(t)$ at the time $t$ for the history dependent number of states $N_{C_i(t)}$ and the total number of states in this state of the correlation function $C_i(t)$,

$$\rho_i(t) = \frac{N_{C_i(t)}}{\sum_i N_{C_i(t)}}. \tag{4}$$

That definition of the local probability density allows the discretization of the path-dependent correlation and the definition of a log likelihood function (see Eq. (3) below). As follows, we define the correlation dependent density $\rho_i(t)$ as a function of the correlation function $C_i(t)$ for an atom with index $i$,

$$\rho_i(t) = \frac{\sum_t \sum_{C_\mu=-1}^{C_\mu=1} e^{-\frac{(C_i(t)-C_\mu)^2}{2\sigma}}}{\sum_i \sum_t \sum_{C_\mu=-1}^{C_\mu=1} e^{-\frac{(C_i(t)-C_\mu)^2}{2\sigma}}}, \tag{5}$$

where $\sigma$ defines the width of the Gaussian function (due to the fact that we apply a histogram over $10^2$ bins, we apply $\sigma = 2 \times 10^{-2}$). [see Fig. 2, where we show the time-dependency of the quantities $C(t)$ and $L$ over a simulation period of 1 ns]. Subsequently, we introduce a log pseudo-likelihood function $l$ of the correlation

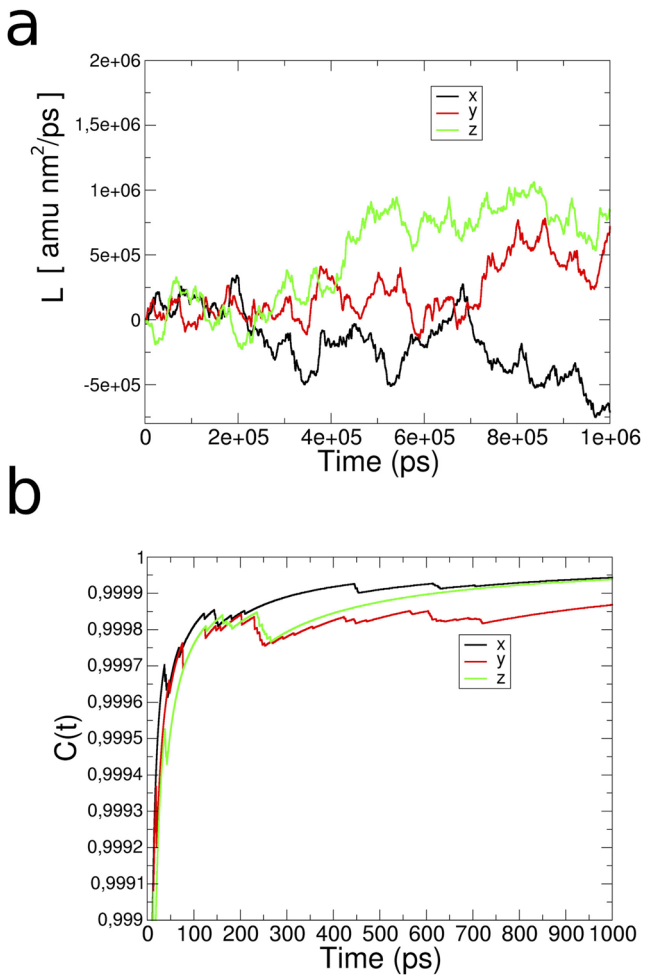


**FIG. 2**. History dependent quantities of the algorithm extracted every 100 time steps from a simulation of TrpCage (extended conformation) over 1 ns (atom: CA, Leu2). (a) Path components $L$ as a function of MD time. (b) Components of correlation function $C(t)$ as a function of MD time.

dependent density. The log likelihood function of the history dependent correlation density describes a form of a correlation dependent potential,

$$l_i(t) = -\log(\rho_i(t)), \tag{6}$$

which defines the corresponding bias $A_i$ with an additional parameter $\alpha$ with the units of energy,

$$A_i = \alpha \nabla l_i(t), \tag{7}$$

as the derivative along a unit vector with a unit length due to the dimensionality of the correlation function. As a consequence, the bias gradient evolves as the gradient of the potential of the history dependent probability density $\rho_i(t)$, which is described by the log functional in Eq. (6). That way, we maximize the correlation dependent likelihood in analogy to the principle of maximum entropy.[64]

We introduce a factorization of the total gradient by a factor $r_i$ as a second segment into the CORE-MD algorithm. The application of the bias gradient using only the bias derived from the path-dependent correlation requires the sufficient sampling of the correlation space. The correlation space of the path correlation is sampled along a first-order rate equation (see the supplementary material),

$$\dot{C}_i(t) = -k_{i_1} t, \tag{8}$$

where $k_1$ stands for the first-order rate constant. In order to reach a sufficient sampling efficiency of the correlation space, we scale the resulting gradient by a correlation dependent factor $r$, in order to enhance the decay of the autocorrelation and achieve a faster access of the folded conformation space.[65,66] As a consequence of the factorization, the time-dependent behavior of the correlation function is, then, described by a second-order rate equation (for the derivation, see Sec. I.A. of the supplementary material),

$$\dot{C}_i(t) = -(k_{i_1} + k_{i_2})t. \tag{9}$$

We define the factor $r_i(t)$ as

$$r_i(t) = e^{-C_i(t)}(1 + C_i(t)). \tag{10}$$

In the global picture, the log likelihood function converges to the global log likelihood of the total correlation dependent density $\Xi$,

$$\lim_{t\to\infty} l_i(t) = -\log(\Xi_i), \tag{11}$$

where we can approximate $\Xi_i$ as the probability function $P_i$ of the path-dependent correlation,

$$P_i \approx \Xi_i. \tag{12}$$

Finally, we conclude that this algorithm samples the global free energy in the infinite time limit due to the definition of $\Delta F_i = -k_B T \log(P_i)$.

### B. Algorithm

- Increment the path integral [see Eq. (1)].
- Calculate correlation function $C(t)$ using Eq. (2).
- Calculate the correlation dependent likelihood from Eq. (5).
- Calculate the new gradient using Eq. (7) and apply the factorization from Eq. (10).

### C. Simulation parameters and system setup

For all simulations, we used a modified version of the GROMACS simulation package, version 4.5.5.[43] The code and the run input files of the validation simulations are available upon request. We used the AMBER99-SB force field for describing the interactions of peptides and ions.[6,7,67,68] We used periodic boundary conditions in xyz. In each simulation, we applied a time step of 1 fs. All simulations have been carried out at a temperature of 300 K. The temperature has been controlled using the velocity rescaling

thermostat with a coupling constant $\tau = 1$ ps.[69] We used the standard generalized Born solvent accessibility (GBSA) parameters, where electrostatics and Lennard-Jones interactions were calculated using a twin-range cutoff of 1.0/1.4 nm (the long-ranged interaction calculated every second time step) (TrpCage). We mention that for explicit solvent systems, the factor $\alpha$ has to be adjusted to smaller values in the range below 0.1 kJ/mol. For dialanine, we used a cutoff of 1.0/1.1 nm. In all simulations, we applied a neighbor list cutoff equal to 1.0 nm with an update frequency every three time steps. For the simulations of dialanine, we centered the extended peptide (Ace-Ala-NMe) in a box with dimensions $2.27 \times 2.27 \times 2.27$ nm$^3$. For the simulations of TrpCage,[70] we centered the extended peptide (NLYIQWLKDGGPSSGRPPPS) in a cubic box with a box-length of 7.022 nm. We propagated the TrpCage system over 100 ns and dialanine over 200 ns using $\alpha = 0.0$. In ten validation simulations over 20 ns, we applied $\alpha$-values ranging from −1000 kJ/mol to 1000 kJ/mol. We performed a third equilibrium MD simulation of dialanine over 2 $\mu$s with the same conditions. [For the determination of the root mean square deviation (RMSD), we used the NMR structure: 1l2y model No. 1 as reference.] We defined the dihedral $\Phi$ by the atoms C-N-CA-C and $\Psi$ using N-CA-C-N of the dialanine peptide. We compared the transition kinetics of the $\Phi$-angle of dialanine in 2 $\mu$s with the CORE-MD result using a normalized number of transitions $\nu$ of the $\Phi$ angle, $N_\Phi(t)$, from values below zero to positive values,

$$\nu = \frac{N_\Phi(t)}{N_{dt}}, \tag{13}$$

which we normalize by the number of time-frames $N_{dt}$. We define the acceleration factor $a$ using the following relation:

$$a = \frac{\nu_b}{\nu_u}, \tag{14}$$

where $\nu_u$ is the unbiased frequency and $\nu_b$ stands for the biased case. For the determination of $a$, we first applied the analysis to the number of timeframes, while we second used a scaling by the simulation time to determine the total acceleration factor. For dialanine, we applied an energy parameter $\alpha$ equal to 0 so that the bias was dependent on the factor $r_i(t)$, which acts on the gradient. In the simulation of TrpCage, we used an energy parameter $\alpha$ equal to 5 kJ/mol. The probability density $\rho_i(t)$ was updated with a frequency equal to 1 ps$^{-1}$ [see Eq. (5)]. For the measurement of the free energies $\Delta F$, we used[71,72]

$$\Delta F = -k_B T \ln \frac{P}{P_{min}}, \tag{15}$$

where $P_{min}$ stands for the minimal probability of the projection on two quantities, $k_B$ is Boltzmann's constant, and $T$ stands for the temperature. We compared the free energy landscapes of the 2 $\mu$s MD simulation with the CORE-MD result through the determination of the difference $\Delta\Delta F$ between the MD and the CORE-MD result. We defined the number of folding transitions $N_{fold}(t)$ as the sum of RMSD-dependent folding events over time. We counted each event as a transition, in which the $RMSD_{C\alpha-C\alpha}$ (i.e. the root mean square deviation of the $C\alpha$-atoms to the native structure, PDB: 1L2Y, NMR model No. 1) changes from values above to values below a given threshold (we varied the cutoff from 0.75 nm to 0.12 nm). We plotted the number $N_{fold}$ as a function of MD time for different cutoffs.

### D. Program

The module has been implemented into the molecular dynamics module of the GROMACS-4.5.5 package, into /src/kernel/md.c. A global call to collect forces and coordinates is performed, and bias forces are re-distributed onto each assigned core within the domain decomposition of the simulation system. All simulations have been carried out on a seven core Desktop with Intel Core(TM) i7-7700 CPUs 3.60 GHz computer using two cores for each simulation. The average computation time of the dialanine systems was 2 h–4 h. We simulated the folding landscape of TrpCage within 3 days.

## III. RESULTS AND DISCUSSION

### A. Dialanine

We applied the algorithm to the sampling of the conformational landscape of dialanine and compared the result with a 2 $\mu$s MD simulation on the same system (see Fig. 3). The basic outline of the CORE-MD algorithm contains two different segments: the history dependent segment, which adds Gaussian functions to determine a pseudo-likelihood $l$ that is, then, applied as a bias to the system using the energy parameter $\alpha = 0$. The second segment applies a correlation dependent shift to the total force contribution through a factorization with the factor $r$. In order to evaluate the effect of the factor $r$ on the first CORE-MD validation simulation, we applied an energy parameter $\alpha = 0$ kJ/mol and conclude that the factorization with the factor $r$ is sufficient for an accurate sampling of the FEL. In ten independent simulations, we applied different $\alpha$-values ranging from −1000 kJ/mol to 1000 kJ/mol to validate the effect of this parameter (see Fig. 1S of the supplementary material). In the simulations (CORE-MD, 200 ns $\alpha = 0.0$ kJ/mol, and 2 $\mu$s MD), we observe populations of $\Delta F = -9k_BT$ at the minima (1, 2), and (3) along the FEL [see Figs. 3(a) and 3(e)]. The minimum (4) is populated with $\sim -6k_BT$. We compared both results from equilibrium MD and CORE-MD through the calculation of the difference $\Delta\Delta F$ between both FELs. A large area of 41.8% of the CORE-MD landscape of dialanine is approximately identical with the result from equilibrium MD within the range of the accuracy. A comparatively big fraction of 31.8% of the FEL differs by $1k_BT$, while an area of 18.6% contains differences in the free energy in the range from 1 to $3k_BT$ [see Fig. 3(c)]. In the difference plot, we observe that the main populations (1, 2, 3, and 4) are sampled within an accuracy of less than $1k_BT$, where the largest difference resides at the minimum (4) due to a slightly larger propensity for the minimum in the CORE-MD simulation ($\Phi \approx 50°$, $0 < \Psi < 50°$). We find larger energy differences in the range from 1 to $3k_BT$ at the regions that lie apart from the main minima, especially at the regions (2a, 3a, and 3b) [see Fig. 3(c)]. Finally, we compared the relative transition counts of the $\Phi$-angle $\nu$, which we normalized by the number of time-frames $N_t$, which provides an estimate for the finite difference between the relative transitions in the CORE-MD simulation and the equilibrium MD run. Using this measure, we find that the relative acceleration of the CORE-MD sampling is 3.897 times faster than the conventional MD simulation [see Fig. 3(g)]. If we

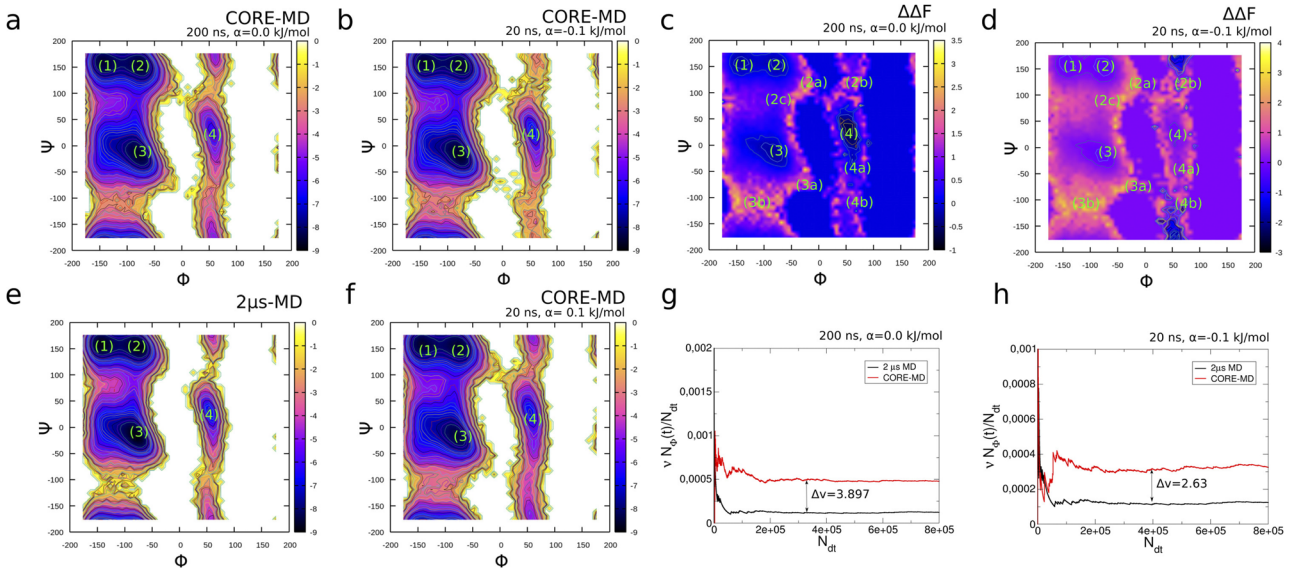


**FIG. 3**. Results from CORE-MD validation simulations of dialanine (Ace-Ala-NMe) and 2 $\mu$s equilibrium MD of the same system. (a) Free energy landscape of dialanine as a function of the dihedral angles $\Phi$ and $\Psi$ from a 200 ns CORE-MD simulation using $\alpha$ = 0.0 kJ/mol. (b) Free energy landscape along the same order parameters averaged over 20 ns CORE-MD with $\alpha$ = −0.1 kJ/mol. (c) $\Delta\Delta F$ plot, given by the difference between the 2 $\mu$s MD result and the FEL of the 200 ns CORE-MD simulation using $\alpha$ = 0.0 kJ/mol. (d) $\Delta\Delta F$ plot, given by the difference between the 2 $\mu$s MD result and the FEL of the 20 ns CORE-MD simulation using $\alpha$ = −0.1 kJ/mol. (e) Free energy landscape of dialanine as a function of the dihedral angles $\Phi$ and $\Psi$ from a 2 $\mu$s equilibrium MD simulation. (f) Free energy landscape of dialanine as a function of the dihedral angles $\Phi$ and $\Psi$ from a 200 ns CORE-MD simulation using $\alpha$ = 0.1 kJ/mol. (g) Normalized number of transitions of the $\Phi$ angle in the 200 ns CORE-MD simulation with $\alpha$ = 0.0 kJ/mol and 2 $\mu$s MD as a function of the number of time-frames $N_{dt}$. (h) Normalized number of transitions of the $\Phi$ angle in the 20 ns CORE-MD simulation with $\alpha$ = −0.1 kJ/mol and 2 $\mu$s MD as a function of the number of time-frames $N_{dt}$. The simulation using $\alpha$ = 0.0 kJ/mol (CORE-MD dynamics are, then, only dependent on the factor $r$) results in an effective acceleration factor of 38.97, while the simulation with $\alpha$ = −0.1 kJ/mol leads to a factor equal to 263.0.

consider the total simulation time and apply it to the transition counts, the factor is ten times larger, and the CORE-MD sampling with $\alpha$ = 0.0 kJ/mol results in a 38.97 times larger transition rate of the $\Phi$-angle.

In the additional set of ten validation simulations over 20 ns, we, then, tested the effect of the factor $\alpha$ with non-zero values ranging from −1000 kJ/mol to 1000 kJ/mol to evaluate the effect of the first segment of the algorithm on the conformational landscape of dialanine. In general, we find that the parameter range at ±1000 kJ/mol contains too large energies, which can induce impeded transitions of the $\Phi$-angle [see Figs. 1S(a) and 1S(b)]. Surprisingly, even values of ±1 kJ/mol lead to a comparatively weak sampling of the $\Phi$-transition [see Figs. 1S(g) and 1S(h)]. The energetic window in which the first segment of the CORE-MD algorithm accelerates the $\Phi$-transition lies in the range from ±0.01 kJ/mol to ±0.1 kJ/mol because we again find a low propensity for $\Phi$-transitions for $\alpha$-values below that specific magnitude of 0.01 kJ/mol (see Fig. 1S). For the two simulations over 20 ns, we find again a very good quantitative agreement of the FEL with the 2 $\mu$s MD result [see Figs. 3(b), 3(d), and 3(f)]. Surprisingly, an $\alpha$-parameter equal to −0.1 leads to an approximate equivalence of the resulting free energy values at the main minima (1, 2, 3), and (4) [see Fig. 3(d)]. In the comparison of the relative transition rate, we observe that the acceleration factor at an $\alpha$-value equal to 0.1 is 2.63, which results in a total acceleration factor equal to 263 in relation to the MD simulation over 2 $\mu$s [see Fig. 3(h)].

We conclude that the CORE-MD algorithm is capable of sampling the conformational landscape of dialanine with a high accuracy in comparison with 2 $\mu$s equilibrium MD. The factorization with the factor $r$ with a value of $\alpha$ = 0.0 kJ/mol is sufficient for the accurate sampling of the system and leads to an effective acceleration equal to a factor of 38.97. The application of non-zero alpha values in the range from −1000 kJ/mol to 1000 kJ/mol shows that the choice of this parameter is crucial for the sampling of the transition along the $\Phi$-angle. Within a suitable parameter range between ±0.01 kJ/mol and ±0.1 kJ/mol, the conformational landscape of dialanine is sampled accurately with a higher acceleration factor of 263, which is one order of magnitude bigger than for $\alpha$ = 0.0 kJ/mol.

## B. TrpCage

We validated the algorithm in a folding simulation of TrpCage starting from an extended conformation (see Fig. 4). We observe folding of the peptide to $RMSD_{C\alpha-C\alpha}$ = 0.14 nm at ~5 ns after an initial collapse within first ns to $RMSD \approx 0.3$ nm [see Fig. 4(a)]. After the population of the minimum with $RMSD < 0.2$ nm, the contact between the N-terminal helix and the poly-proline segment re-opens, and we observe a process of un-refolding over the complete trajectory with a total number of ~1500 folding and un-folding events over the whole trajectory of 100 ns if we define the folded state at $RMSD_{C\alpha-C\alpha} \leq 0.2$ nm [see Fig. 4(b)]. Using the thresholds $RMSD_{C\alpha-C\alpha} \leq 0.15$ nm and $RMSD_{C\alpha-C\alpha} \leq 0.12$ nm, we observe

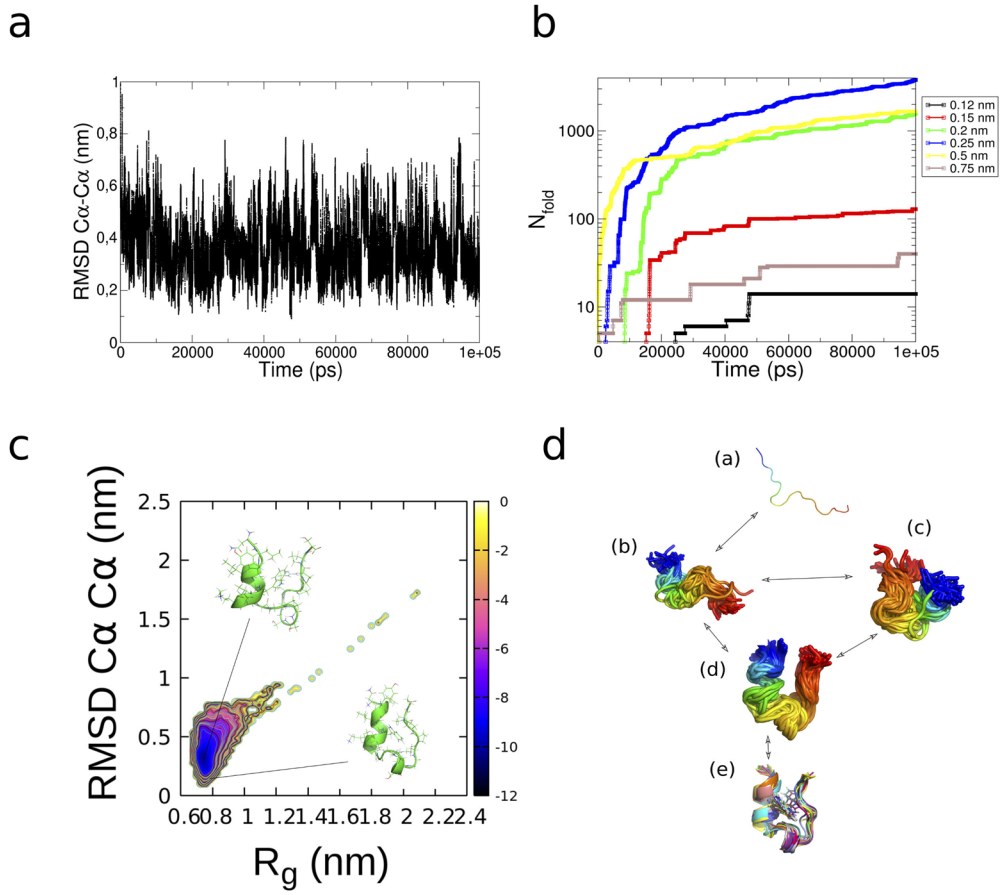


**FIG. 4**. Results from a validation simulation of folding of TrpCage. (a) $RMSD_{C\alpha-C\alpha}$ of TrpCage to the native structure (PDB: 1l2y, model No. 1) as a function of simulation time.[70] (b) Kinetic analysis of the folding simulation of TrpCage. Number of folding transitions for different RMSD values in the range from 0.12 nm to 0.75 nm. (c) Free energy landscape as a function of $RMSD_{C\alpha-C\alpha}$ and the radius of gyration ($R_g$). Energy units on the color bar are in units of $k_BT$. (d) Kinetic network analysis of folding pathways from a RMSD based clustering.

126 (0.15 nm) and 14 folding events to reach RMSD values below 0.12 nm of this peptide. The free energy landscape of that simulation contains a minimum at $Rg$ = 0.75 and $RMSD_{C\alpha-C\alpha}$ = 0.17 nm–0.3 nm, while a region that is slightly higher in energy reaches states below $RMSD$ = 0.10 nm. At ~50 ns, the transition frequency for $RMSD_{C\alpha-C\alpha}$ < 0.12 nm reaches a plateau. We point out that this observation does not indicate that the simulations for the time before 50 ns are out of equilibrium, in principle, but that the global correlation dependent functional has reached convergence within its history dependent potential. We find that the unfolded state in that set of simulations remains unpopulated with energies above $-1k_BT$ [see Fig. 4(c)]. We conclude that TrpCage collapses to RMSD values below 0.2 nm at various timescales, which indicates that the folding pathways of that peptide can strongly vary in the transition time. Using an RMSD based clustering with a cutoff equal to 0.1 nm, we find that the dominant pathways of folding of TrpCage occur via a helix-rich intermediate at 0.25 nm < RMSD <0.5 nm [conformer (d)], which is accessed by conformations with an unfolded alpha-helical part (c) or an ensemble of conformations, which contains a pre-folded 3–10 helix [conformer (b)] [see Fig. 4(d)].

The folding pathway observed in that validation simulation is dominated by the formation of the secondary structure and followed by the formation of the tertiary fold as essential steps along the folding pathway.[50,51] A larger number of simulation studies agree on the existence of two dominant pathways of folding of TrpCage: the prior formation of the N-terminal helical element followed by the closure of the hydrophobic pocket at Trp6, or second, the closure of the hydrophobic contact leading to the subsequent formation of the N-terminal $\alpha$-helix of TrpCage. The existence of these two major folding pathways was observed in transition path-sampling simulations[53,54] and a theoretical study using extensive biases along each potential minimum.[55] Highly parallel simulations of TrpCage with the folding@home project yielded accurate results on the folding transition times.[56] Another very accurate folding time has been observed with simulations of folding on a special purpose machine ANTON.[3] The observation of a predominant folding pathway is,

in fact, also strongly force field dependent, and quantities as the propensity for defined secondary structure elements can be affected by the choice of the underlying energy functional.[73–75] The presence of surfaces and their chemical nature define the affinity of TrpCage for adsorption, which can change the folding landscape with a stabilized native structure.[76,77] In hybrid kMC/MD simulations, we observed the formation of the 3–10 helix, which is followed by the closure of the Poly-Pro helical segment of the peptide on timescales ranging from 800 ns to 4 $\mu$s. The overall folding pathway was dependent on the selected moves. In general, we observed three different folding pathways.[51,52] In the path-dependent biased MD simulations, we observed the same patterns of folding pathways, while the folded minimum also contained a free energy of $-7$ to $-9k_BT$.[58,63,78,79] These values also agree with results from replica exchange MD simulations.[60] Very generally speaking, our observed folding pathways are in agreement with these previous results, while we find that the number of folding events is much higher than in the previous simulations. The qualitative shape of the free energy landscape changes slightly, when we compare our result with our previous simulation studies. That observation can be attributed to the implicit solvent environment used in that study. In explicit solvent, the protein collapses within a larger heterogeneity of timescales already in the unfolded state.[51] Related to the total convergence of the folded state, the CORE-MD algorithm shows a higher efficiency than our previously developed algorithms applied on the same protein.[51,52,58,63] Compared with a recent path-sampling MD implementation, the algorithm is ~1500 times faster in terms of the sampling of folding events of the same peptide.[58] Our results are in agreement with our previous findings and other theoretical studies.[47–60]

## IV. CONCLUSIONS

In this paper, we presented a new enhanced sampling MD method, which is independent from the requirement of any *a priori* structural information about the system and needs only one energy parameter as an additional input. We call this method the correlation dependent MD method (CORE-MD). The method uses path-dependent correlation functions as a collective variable and constructs a history dependent correlation density in an adaptive way. The resulting gradient is calculated within the dimensions of the correlation function and evolves along a unit vector in contrast to conventional metadynamics implementations. We apply an additional correlation dependent formalism to the resulting gradient, which leads to an acceleration of the decay kinetics of the auto-correlations in the system. We validate the new method on the conformational landscape of dialanine and the folding of TrpCage. In both cases, we find good agreement with the data reported in the literature. In the sampling of TrpCage folding, the method reaches an acceleration of folding events with a factor of 1500 compared with a recent path-sampling implementation. The method is broadly applicable to MD simulations, in general, such as the sampling of polymer systems.

## SUPPLEMENTARY MATERIAL

The supplementary material contains the derivation of the effect of the factor $r_i(t)$ on the propagation (see Sec. II) and the free energy landscapes of dialanine, where we varied the energy factor $\alpha$.

## ACKNOWLEDGMENTS

The authors gratefully acknowledge the Gauss Centre for Supercomputing eV (www.gauss-centre.eu) for funding this project by providing computing time through the John von Neumann Institute for Computing (NIC) on the GCS Supercomputer JUWELS at the Jülich Supercomputing Centre (JSC). J.-E. S. acknowledges the support from the National Science Foundation (NSF grant MCB-1716956).

## DATA AVAILABILITY

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

## REFERENCES

[1] J. N. Onuchic and P. G. Wolynes, Curr. Opin. Struct. Biol. **14**, 70–75 (2004).
[2] T. P. J. Knowles, C. A. Waudby, G. L. Devlin, S. I. A. Cohen, A. Aguzzi, M. Vendruscolo, E. M. Terentjev, M. E. Welland, and C. M. Dobson, Science **326**, 1533–1537 (2009).
[3] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, Science **334**, 517–520 (2011).
[4] L. L. Chavez, J. N. Onuchic, and C. Clementi, J. Am. Chem. Soc. **126**, 8426–8432 (2004).
[5] A. Schug, P. C. Whitford, Y. Levy, and J. Ń. Onuchic, Proc. Natl. Acad. Sci. U. S. A. **104**, 17674–17679 (2007).
[6] S. Gosavi, P. C. Whitford, P. A. Jennings, and J. N. Onuchic, Proc. Natl. Acad. Sci. U. S. A. **105**, 10384–10389 (2008).
[7] P. C. Whitford, J. K. Noel, S. Gosavi, A. Schug, K. Y. Sanbonmatsu, and J. N. Onuchic, Proteins **75**, 430–441 (2009).
[8] A. Schug, T. Herges, and W. Wenzel, Phys. Rev. Lett. **91**, 158102 (2003).
[9] A. Schug, W. Wenzel, and U. H. E. Hansmann, J. Chem. Phys. **122**, 194711 (2005).
[10] G. M. Torrie and J. P. Valleau, J. Comput. Phys. **23**, 187–199 (1977).
[11] D. P. Landau, S.-H. Tsai, and M. Exler, Am. J. Phys. **72**, 1294 (2004).
[12] A. Laio and M. Parrinello, Proc. Natl. Acad. Sci. U. S. A. **99**, 12562–12566 (2002).
[13] A. F. Voter, Phys. Rev. Lett. **78**, 3908 (1997).
[14] H. Grubmüller, Phys. Rev. E **52**, 2893 (1995).
[15] T. Huber, A. E. Torda, and W. F. van Gunsteren, J. Comput.-Aided Mol. Des. **8**, 695–708 (1994).
[16] U. H. E. Hansmann and L. T. Wille, Phys. Rev. Lett. **88**, 068105 (2002).
[17] M. M. Sultan and V. S. Pande, J. Chem. Phys. **149**, 094106 (2018).
[18] J. M. L. Ribeiro and P. Tiwary, J. Chem. Theory Comput. **15**, 708–719 (2018).
[19] J. M. Jumper, N. F. Faruk, K. F. Freed, and T. R. Sosnick, PLoS Comput. Biol. **14**, e1006578 (2018).
[20] O. Valsson and M. Parrinello, Phys. Rev. Lett. **113**, 090601 (2014).
[21] L. Donati and B. G. Keller, J. Chem. Phys. **149**, 072335 (2018).
[22] Y. Wang, O. Valsson, P. Tiwary, M. Parrinello, and K. Lindorff-Larsen, J. Chem. Phys. **149**, 072309 (2018).
[23] A. Perez, J. L. MacCallum, and K. A. Dill, Proc. Natl. Acad. Sci. U. S. A. **112**, 11846–11851 (2015).
[24] A. Perez, J. A. Morrone, E. Brini, J. L. MacCallum, and K. A. Dill, Sci. Adv. **2**, e1601274 (2016).
[25] A. Perez, F. Sittel, G. Stock, and K. Dill, J. Chem. Theory Comput. **14**, 2109–2116 (2018).
[26] V. Lindahl, J. Lidmar, and B. Hess, Phys. Rev. E **98**, 023312 (2018).
[27] F. Palazzesi, O. Valsson, and M. Parrinello, J. Phys. Chem. Lett. **8**, 4752–4756 (2017).

[28]D. Branduardi, F. L. Gervasio, and M. Parrinello, J. Chem. Phys. **126**, 054103 (2007).
[29]M. Allen and D. Tildesley, *Computer Simulation of Liquids* (Clarendon Press, Oxford, 1987).
[30]H. C. Andersen, J. Comput. Chem. **52**, 24–34 (1983).
[31]S. Miyamoto and P. A. Kollman, J. Comput. Chem. **13**, 952–962 (1992).
[32]B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, J. Comput. Chem. **18**, 1463–1472 (1997).
[33]V. Kräutler, W. F. van Gunsteren, and P. H. Hünenberger, J. Comput. Chem. **22**, 501–508 (2001).
[34]Q. Ma and J. A. Izaguirre, Multiscale Model. Simul. **2**, 1–21 (2003).
[35]R. Olender and R. Elber, J. Chem. Phys. **105**, 9299–9315 (1996).
[36]B. Leimkuhler, D. T. Margul, and M. E. Tuckerman, Mol. Phys. **111**, 3579–3594 (2013).
[37]M. Tuckerman, B. J. Berne, and G. J. Martyna, J. Chem. Phys. **97**, 1990 (1992).
[38]P. G. Bolhuis, C. Dellago, and D. Chandler, Proc. Natl. Acad. Sci. U. S. A. **97**, 5877 (2000).
[39]W.-N. Du and P. G. Bolhuis, J. Chem. Phys. **139**, 044105 (2013).
[40]R. Elber, Biophys. J. **92**, L85–L87 (2007).
[41]S. a Beccara, T. Skrbic, R. Covino, and P. Faccioli, Proc. Natl. Acad. Sci. U. S. A. **109**, 2330–2335 (2012).
[42]E. Suárez, J. L. Adelman, and D. M. Zuckerman, J. Chem. Theory Comput. **12**, 3473–3481 (2016).
[43]B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, J. Chem. Theory Comput. **4**, 435–447 (2008).
[44]D. J. Tobias and C. L. Brooks III, J. Phys. Chem. **96**, 3864–3870 (1992).
[45]W. C. Swope, J. W. Pitera, F. Suits, M. Pitman, M. Eleftheriou, B. G. Fitch, R. S. Germain, A. Rayshubski, Y. Zhestkov, and R. Zhou, J. Chem. Phys. B **108**, 6582–6594 (2004).
[46]P. Tiwary and M. Parrinello, Phys. Rev. Lett. **111**, 230602 (2013).
[47]L. Qiu, S. A. Pabit, A. E. Roitberg, and S. J. Hagen, J. Am. Chem. Soc. **124**, 12952–12953 (2002).
[48]H. Neuweiler, S. Doose, and M. Sauer, Proc. Natl. Acad. Sci. U. S. A. **102**, 16650–16655 (2005).
[49]R. M. Culik, A. L. Serrano, M. R. Bunagan, and F. Gai, Angew. Chem. **123**, 11076–11079 (2011).
[50]H. Meuzelaar, K. A. Marino, A. Huerta-Viga, M. R. Panman, L. E. J. Smeenk, A. J. Kettelarij, P. T. J. H. van Maarseveen, P. G. Bolhuis, and S. Woutersen, J. Phys. Chem. B **117**, 11490–11501 (2013).
[51]E. K. Peter and J.-E. Shea, Phys. Chem. Chem. Phys. **16**, 6430–6440 (2014).
[52]E. K. Peter, I. V. Pivkin, and J.-E. Shea, J. Chem. Phys. **142**, 144903 (2015).
[53]J. Juraszek and P. G. Bolhuis, Proc. Natl. Acad. Sci. U. S. A. **103**, 15859–15864 (2006).
[54]J. Juraszek and P. G. Bolhuis, Biophys. J. **95**, 4246–4257 (2008).
[55]F. Marinelli, F. Pietrucci, A. Laio, and S. Piana, PLoS Comput. Biol. **5**, e1000452 (2009).
[56]C. D. Snow, B. Zagrovic, and V. S. Pande, J. Am. Chem. Soc. **124**, 14548–14549 (2002).
[57]H. Ren, Z. Lai, J. D. Biggs, J. Wang, and S. Mukamel, Phys. Chem. Chem. Phys. **15**, 19457–19464 (2013).
[58]E. K. Peter and J.-E. Shea, Phys. Chem. Chem. Phys. **19**, 17373–17382 (2017).
[59]E. K. Peter, J. Chem. Phys. **147**, 214902 (2017).
[60]R. Zhou, Proc. Natl. Acad. Sci. U. S. A. **100**, 13280–13285 (2003).
[61]H. Kleinert, *Path Integrals In Quantum Mechanics, Statistics, Polymer Physics, And Financial Markets*, (5th ed) (World Scientific, 2009), pp. 1–1547.
[62]R. Feynman and A. R. Hibbs, *Quantum Mechanics and Path Integrals* (McGraw Hill Companies, 1965).
[63]E. K. Peter and J. Černý, Int. J. Mol. Sci. **20**, 370 (2019).
[64]L. R. Mead and N. Papanicolaou, J. Math. Phys. **25**, 2404 (1984).
[65]K. Y. Sanbonmatsu, Curr. Opin. Struct. Biol. **22**, 168–174 (2012).
[66]L. S. Bigman and Y. Levy, Curr. Opin. Struct. Biol. **60**, 50–56 (2020).
[67]P. A. Kollman, Acc. Chem. Res. **29**, 461–469 (1996).
[68]V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, Proteins **65**, 712–725 (2006).
[69]G. Bussi, D. Donadio, and M. Parrinello, J. Chem. Phys. **126**, 014101 (2007).
[70]J. W. Neidigh and R. M. Fesinmeyer, Nat. Struct. Biol. **9**, 425–430 (2002).
[71]A. Shehu, L. E. Kavraki, and C. Clementi, Biophys. J. **92**, 1503–1511 (2007).
[72]J. I. Sulkowska, P. Sulkowski, and J. Onuchic, Proc. Natl. Acad. Sci. U. S. A. **106**, 3119–3124 (2009).
[73]R. Day, D. Paschek, and A. E. Garcia, Proteins: Struct., Funct., Bioinf. **78**, 1889–1899 (2010).
[74]D. Paschek, S. Hempel, and A. E. Garcia, Proc. Natl. Acad. Sci. U. S. A. **105**, 17754–17759 (2008).
[75]D. Paschek, R. Day, and A. E. García, Phys. Chem. Chem. Phys. **13**, 19840–19847 (2011).
[76]G. H. Zerze, R. G. Mullen, Z. A. Levine, J.-E. Shea, and J. Mittal, Langmuir **31**, 12223–12230 (2015).
[77]Z. A. Levine, S. A. Fischer, J.-E. Shea, and J. Pfaendtner, J. Phys. Chem. B **119**, 10417–10425 (2015).
[78]E. K. Peter, J.-E. Shea, and I. V. Pivkin, Phys. Chem. Chem. Phys. **18**, 13052–13065 (2016).
[79]E. K. Peter and J. Černý, Int. J. Mol. Sci. **19**, E3405 (2018).