Coarse kMC-based replica exchange algorithms for the accelerated simulation of protein folding in explicit solvent

Emanuel K. Peter, a, b Joan-Emma Shea and Igor V. Pivkin a, c

In this paper, we present a coarse replica exchange molecular dynamics (REMD) approach, based on kinetic Monte Carlo (kMC). The new development significantly can reduce the amount of replicas and the computational cost needed to enhance sampling in protein simulations. We introduce 2 different methods which primarily differ in the exchange scheme between the parallel ensembles. We apply this approach on folding of 2 different β-stranded peptides: the C-terminal β-hairpin fragment of GB1 and TrpZip4. Additionally, we use the new simulation technique to study the folding of TrpCage, a small fast folding α-helical peptide. Subsequently, we apply the new methodology on conformation changes in signaling of the light-oxygen voltage (LOV) sensitive domain from Avena sativa (AsLOV2). Our results agree well with data reported in the literature. In simulations of dialanine, we compare the statistical sampling of the 2 techniques with conventional REMD and analyze their performance. The new techniques can reduce the computational cost of REMD significantly and can be used in enhanced sampling simulations of biomolecules.

1 Introduction

Small peptides fold into their native structure within short timescales of few microseconds. 1 Molecular dynamics (MD) simulations serve as a tool to elucidate pathways of folding and protein structure prediction.2–4 It has been found that proteins are rare event systems, with free energy barriers between the different minima along the folding landscape, leading to the time-scale problem of standard molecular dynamics.5 To enhance sampling in folding simulations, a number of developments have been made, including meta-dynamics,6 the weighted ensemble technique,7 milestoning8 and transition path sampling.9

In contrast to MD, the Monte Carlo (MC) technique propagates the peptide system along an un-correlated trajectory and thereby enables enhanced sampling of events in folding simulations.10 Recently, we developed kMC-MD technique suitable for folding of α-helical and β-stranded peptides in explicit solvent11,12 and signaling of photoactivatable proteins.13,14 We found that the pathways of folding are strongly dependent on the nature of the MC move.11 In another work, we developed a specific move-set which enables folding of β-stranded and z-helical peptides.12

REMD is a method, which is extensively used in simulations of protein folding and aggregation.15–18 This method runs different trajectories in parallel and exchanges the configurations at different temperatures, taking the finite difference in the potential energies in each replica as criterion.19 This approach leads to enhanced sampling of protein folding through an acceleration by the temperature gradient.

Despite its wide use in enhanced sampling simulations, and although a vast number of developments have been made to improve REMD,20–40 its computational cost is high and its sampling efficiency suffers from a decrease in its acceptance rate with increasing size of the system, especially in simulations where explicit solvent is used.

In this paper, we present 2 novel REMD-based algorithms in which we use a coarse-grained exchange scheme, leading to a significant reduction of replicas needed in a wide temperature range, which can reduce the computational cost significantly in comparison with conventional REMD. The exchange probabilities and the selection of exchanges between replicas are based on kinetic Monte Carlo (kMC).41,42 Each exchange step consists of a collection of exchanges between the different replicas. We use our new techniques for folding simulations of TrpCage, GB1 and TrpZip4 and find that the 3 peptides fold to their native states at low computational cost. We additionally apply the new methodology on the signaling dynamics of the AsLOV2 domain and find the correct conformation changes of this protein in agreement with previous studies. We present a detailed
comparison between REMD and the coarse-graining techniques in simulations of dialanine, where we compare efficiencies of our coarse-graining algorithms and conventional REMD.

2 Methods

The algorithms we present in this paper simulate the protein system in a set of equilibrium ensembles at different temperatures. For the exchange of configurations between the ensembles, we perform a coarse-graining in replica space and use probability-based kMC criteria for the exchange. In contrast to original kMC, the algorithm does not produce kinetic information due to the applied parallel integration and exchanges. We make approximations for the coarse-graining in replica space, which allow a reduction of replicas to simulate folding processes and conformation changes. In recent work, we have implemented a hybrid KMC/MD method in which the transition rates were sampled by discrete moves and thermodynamic integration. Using this method, we have shown that the pathway of protein folding strongly depends on the chosen move-set. In another work, we used a generalized move set along cartesian coordinates for folding of α-helical and β-stranded peptides and introduced a new scheme for the time-update. In contrast to these prior works, we introduce a new approach based on a coarsening in replica space. We introduce 2 different methods which primarily differ in the exchange scheme between the parallel ensembles. The Φ-ensemble technique uses a fluctuation-based criterion Φ for a collective exchange, while kinetic exchange MD (KEMD) uses the distribution of the replicas into different sub-states which are selected for an exchange within one group. The Φ-ensemble propagates the system in parallel trajectories and in each exchange step one configuration is chosen for the subsequent parallel propagation. In contrast, KEMD is a method which clusters the replicas into groups and performs collective exchanges.

2.1 Theory

The general approach of the methods presented in this paper is based on coarse-graining of the exchange criteria derived for conventional REMD. Conventional REMD depends on the potential energy overlap between the different ensembles and a Metropolis criterion for the exchange between 2 neighboring replicas. In this method, the conformation space of the system is explored while the conformations swap between different temperature levels. Our approaches aim on a reduction of replicas needed in a wide temperature space, which is achieved through a coarse-graining of the replica space. We developed 2 different methods, which express the probabilities for collective exchanges that the number of replicas needed is reduced significantly, while the swapping efficiency is enhanced. We note that the approach presented here is based on approximations, which we will explain in the following.

2.1.1 Φ-ensemble. In the Φ-ensemble technique, each replica has an individual probability to be selected and exchanged with all other replicas, that we can write the state of the extended ensemble as one collective state \( \{X_i\} \). For a set of temperatures \( \beta_i \), the probability distribution of finding \( \{X\} \) is written as:

\[
P(\{X, \beta_i\}) = P_c(X_c, \beta_c).
\]

This collective state can pass into another collective state, which is expressed by:

\[
P_c(X_c, \beta_c)W(X_c, \beta_c|X_c', \beta_c') = P_c(X_c', \beta_c')W(X_c, \beta_c|X_c', \beta_c); \quad (2)
\]

and

\[
\frac{W(X_c, \beta_c|X_c', \beta_c')}{W(X_c, \beta_c|X_c, \beta_c)} = \exp(-\Delta), \quad (3)
\]

where \( \exp(-\Delta) = \exp(-\Delta \Phi_i) = p_i \), where \( \Phi_i \) stands for a fluctuation dependent term in one single replica. Finally, we select one event using the Bortz–Kalos Lebowitz algorithm by solving:

\[
P_{j-1} \leq \xi p_i < p_j, \quad (4)
\]

where \( \xi \) is a uniformly distributed random number \( \xi \in [0,1] \).

The fluctuation dependent term is determined for each MD-phase and is calculated as:

\[
\Delta \Phi_i = A(\{U_i\}) + V_i \Delta p_i - T_i \Delta S_i, \quad (5)
\]

in the NPT-case and \( \Delta \Phi_i = A(\{U_i\}) - T_i \Delta S_i \) in the NVT ensemble. We calculate the variance in the internal energy of the peptide using:

\[
A(\{U_i\}) = \sqrt{\langle (U_i)^2 \rangle - \langle U_i \rangle^2}. \quad (6)
\]

We determine the changes in the conformation entropy of the peptide, using:

\[
\Delta S_i = k_B \sum_j \left( \frac{h_0}{e^{h_0/(k_B T)}} - \ln \left( 1 - e^{-h_0/(k_B T)} \right) \right), \quad (7)
\]

where \( k_B \) stands for Boltzmann’s constant, \( h \) for the reduced Planck’s constant, \( T \) for the temperature in the replica and \( \omega_j \) for the root mean square fluctuation in the system

\[
\omega_j = \frac{1}{\tau} \Delta \text{RMSF}_j / \langle \text{RMSF}_j \rangle, \quad (8)
\]

according to the definition of the harmonic approximation to the conformational entropy change and \( \tau \) stands for the simulation time of one single MD-phase. The term \( \Phi \) expresses the general magnitude of the change in the potential energy and the conformation entropy of the peptide as well as the volume fluctuation. This term is proportional to the change in free energy of the system, while it remains an approximation since it is connected to a short sampling period. This exchange probability is proportional to the variance in the potential energy, the volume fluctuation and the conformation entropy, i.e. the variance in the peptide structure. That means that the replica with a larger conformation change within the last sampling
period also have a higher probability for an exchange and the following propagation in the next MD-phase. In that way, the conformation sampling is accelerated and the transition rate from one state to another is enhanced within a fluctuation dependent ensemble.

In this algorithm, we make an estimate of the fluctuation based term $\Delta \Phi$ as an approximation of the free energy change $\Delta G$ in the system. The absolute value of $\Phi$ is different from $\Delta G$ by a constant factor, i.e., the exchange probabilities are approximately proportional to the change in free energy. In other words, we approximate that other terms, e.g., the solvation-free energy or internal binding free energies remain constant. We tested this approximation in a quantitative comparison, and found that the difference of the instantaneous average solvation free energies is not larger than 1 kJ mol$^{-1}$, which would negligibly affect the exchange probabilities (see Fig. 2d). We estimate the error of this algorithm to be proportional to this constant factor. In contrast to the expression of the system potential energy of the protein and the change in the conformation entropy of the peptide in the system. We choose these terms as an approximation of the free energy change $F$.

We estimate the error of this algorithm to be proportional to their instantaneous potential energy $U$.

$\frac{\Delta \Phi}{\Delta G}$, where $\Delta G = \langle \beta_n - \beta_m \rangle (H(X) - H(X')) + \langle p_n \beta_n - p_m \beta_m \rangle (V_n - V_m)$, for the NPT ensemble, where $p_n$ stands for the pressure and $V$ is the volume of each replica$^{16} (\Delta = \langle \beta_n - \beta_m \rangle (H(X) - H(X'))$ for the NVT ensemble).

In KEMD, we distribute the replicas with indexes 1, 2, ..., $m$, $m + 1$, ... $M$ into $U$ groups with $U < M$. The total number of groups $U$ is given as input parameter, while the individual group size varies for each exchange attempt. The minimal size of a group is a pair of 2 neighboring replicas, which also determines the largest possible group size in dependency of a given number of groups $U$. Thus, the largest possible number of groups $U$ is $M/2$, where $M$ stands for the total number of replicas. Each group consists of pairs of interacting (neighboring) replicas $(m, m + 1, ..., k, k + 1, ... )$. The distribution of pairs of neighboring replica-indices into groups for each exchange is determined using a random assignment, which is performed before each exchange attempt. The cumulative pairwise differences in potential energy, volume and pressure within one group determines its transition rate and the probability to be selected in the collective (pairwise) exchange process. Using the Bortz–Kals Lebowitz algorithm, one group is selected for the pair-wise exchange within this group.$^{41}$ In the framework of this algorithm, we reformulate eqn (10) into:

$$P(\{X, \beta\}) = \prod_{m=1}^{M} P_{eq}(X_m, \beta_m) = \prod_{u=1}^{U} P(X_u, \beta_u);$$

where $P(X_u, \beta_u) = \sum_{m=1}^{u} P_{eq}(X_m, \beta_m)$, for $X_u$ which represents a subset of states. The members in this subset of states must fulfill the condition that:

$$\exp(-\langle [\beta_i - \beta_j] (U_i - U_j) + \langle p_i \beta_i - p_j \beta_j \rangle (V_i - V_j) \rangle) \geq \epsilon$$

with $i < j$ for all pairs in one specific group and $\epsilon$ is a constant. In our simulations, we chose a value for $\epsilon$ of 0.15. Thus, we can write that:

$$P(\ldots; X_u, \beta_{u}; \ldots; X_u', \beta_{u'}; \ldots) \frac{W(X_u, \beta_u, X_u', \beta_{u'})}{W(X_u', \beta_{u'}, X_u, \beta_u)} = P(\ldots; X_u', \beta_{u'}; \ldots; X_u, \beta_{u}; \ldots) \frac{W(X_u', \beta_{u'}, X_u, \beta_u)}{W(X_u, \beta_u, X_u', \beta_{u'})};$$

where

$$\frac{W(X_u, \beta_u, X_u', \beta_{u'})}{W(X_u', \beta_{u'}, X_u, \beta_u)} = \exp(-\Delta),$$

and $\Delta = \langle \beta_n - \beta_m \rangle (H(X) - H(X'))$, where

$$H(X_u) = \sum_{m=1}^{u} H_m(X) + \sum_{m=1}^{u} p_m \beta_m \sum_{m=1}^{u} V_m,$$
and

\[ \beta_u = \sum_{m=1}^{N} \beta_m. \]  

\[ (H(X_u) = \sum_{m=1}^{u} H_m(X) \text{ in the NVT-case}). \]

We then apply a kMC-related formalism on all N subsets of possible exchanges \( i \) between the subsets of states, by assigning a probability \( p_i \) as follows:

\[ p_i = \exp(-\Delta_t) = \left( \frac{W(X_u, \beta_u|X_u, \beta_u)}{W(X_u, \beta_u|X_u, \beta_u)} \right), \]

and finally can select one subset of exchanges by the Bortz–Kalos Lebowitz algorithm,\(^{41,42} \) by solving:

\[ P_j - 1 \leq \xi P_N < P_j, \]

where \( \xi \) is a uniformly distributed random number \( \xi \in [0,1] \), \( P_j = \sum_{i=1}^{j} p_i \) and \( P_N = \sum_{i=1}^{N} p_i \).

Through the re-assignment of replicas to groups, it is possible to enhance the exchange rate and to increase the effective number of swaps between the replicas. Another advantage is the larger independence from the temperature-range for the replicas due to the distribution into groups.

For the pairwise clustering of several replicas into specific groups, it is necessary that there is sufficient overlap of the potential energies of neighboring replicas in ascending order. However, the temperature gradient between the replicas can be larger by one order of magnitude than in conventional REMD, while it still fulfills the condition in eqn (14). At the same time, the exchange efficiency is at least 2 times larger (see section: Performance and convergence analysis), since the exchange criterion is not dependent on the Boltzmann factor between 2 replicas, but the cumulative probability of one specific group.

### 2.2 Algorithms

#### 2.2.1 \( \Phi \)-ensemble. (1) MD-phase in \( M \) parallel replicas.

(2) Analysis of free energy change using eqn (5)–(7).

(3) Global exchange of one replica with all other replicas using eqn (20).

(4) Continue with MD-phase.

#### 2.2.2 KEMD. (1) MD-phase in \( M \) parallel replicas.

(2) Selection of \( U \) sub-states out of \( M \) replicas, i.e. distribution of \( M \) replicas into \( U \) groups (\( U < M \)) consisting of pairs of neighboring replicas (\( \ldots, m, m + 1, \ldots, k, k + 1, \ldots \)).

(3) Determination of exchange probabilities using eqn (17) and (18).

(4) Selection of one group according to its exchange probability using eqn (20).

(5) All pairs of replicas of one group (\( \{ \ldots, m, m + 1, \ldots, k, k + 1, \ldots \} \)) are exchanged with each other.

(6) Continue with MD-phase.

### 2.3 Technical details

Implementations of KEMD and the \( \Phi \)-ensemble technique have been performed within the molecular dynamics program GROMACS-4.5.5. For the \( \Phi \)-ensemble we applied changes in the kernal (/src/kernel/md.c) to collect and average global volume, potential energy and conformation changes and passed them to the replica-exchange module (/src/kernel/repl_ex.c), where we used MPI_Allgather to the global master rank (MULTIMASTER{cr}), where the fluctuation based terms \( \Phi \) are computed for each replica. The exchange flag is communicated to all other master ranks. One master rank of one replica performs a global exchange (MPI_Bcast) with the individual master ranks of each replica (MASTER{cr}). For KEMD, all changes have been applied in the replica-exchange (/src/kernel/repl_ex.c) module. The exchange flags are bundled for each selected group. The exchanges themselves are performed pairwise, as implemented in the replica exchange module.

#### 2.4 Simulation parameters and system preparation

All extended polypeptide conformations were built using the Ribosome program.\(^{49} \) For the simulation of the GB1 peptide, we centered the extended protein in a box with dimensions \( 7.3 \times 7.3 \times 7.3 \) nm\(^3 \) and added 12672 SPC/E waters. We used the SPC/E model, which has self-diffusion coefficients comparable to the experiment.\(^{50} \) We added 3 sodium ions to neutralize the system. In the case of TrpZip4, we centered the extended peptide chain in a box with dimensions \( 7.4 \times 7.4 \times 7.4 \) nm\(^3 \) and solvated the protein with 13242 SPC/E waters. We added 2 sodium ions to neutralize the system. For the simulations of TrpCage, we centered the peptide into a box with dimensions \( 7.022 \times 7.022 \times 7.022 \) nm\(^3 \), filled the box with 11424 SPC/E waters and neutralized the system through addition of one chloride ion. For the simulations of the LOV2 domain from \( \textit{Avena sativa} \) (AsLOV2) we centered the X-ray structure (PDB: 2V01\(^{43} \)) in a box with dimensions \( 6.42 \times 6.42 \times 6.42 \) nm\(^3 \) and filled the box with 8054 SPC/E waters. The system was neutralized with 8 sodium ions. For the generation of the light state structure, we applied the cysteinyl-FMN forcefield parameters\(^{51} \) on the dark state structure of the LOV2 domain and performed a subsequent energy minimization until the bond between Cys450 and FMN-C4a formed. For the simulations of both dark and light state of AsLOV2, we used the GROMOS96-43a1 forcefield.\(^{52} \) In simulations of dialanine, we placed the peptide in a cubic box with dimensions \( 2.27 \times 2.27 \times 2.27 \) nm\(^3 \) and solvated the system with 371 SPC/E water molecules.

In our simulations using the \( \Phi \)-ensemble or KEMD, we simulated the proteins for 10 ps in each MD-phase with an exchange every 10 000 steps using a time-step of 1 fs. In the REMD simulations, we applied an exchange attempt every 1000 steps with a timestep of 1 fs. The number of groups \( U \) in KEMD used in our simulations was 8 (32 replicas), 4 (16 replicas), 4 (8 replicas) and 2 (4 replicas). In all simulations, Particle Mesh Ewald electrostatics (PME) electrostatics with a cutoff of 1.0 nm were used to calculate the electrostatic interactions. The Lennard-Jones 12-6 (LJ) interactions were calculated with the same cutoff using a shift function. Except for AsLOV2, we employed the AMBER99SB forcefield for the description of the interactions of each protein.\(^{53} \) The folding simulations (including the simulation with conventional \( NPT \)-REMD\(^{16} \)) were performed using 32 replicas.
in the temperature range from 300 to 362 K. The simulations with a lower number of 6 and 12 replicas were performed in the temperature range from 300 to 410 K. We performed the $\Phi$-ensemble test simulation of TrpZip4 in explicit solvent over 10 000 integration steps using 6 replicas in a temperature range from 300 to 360 K (see Fig. 2c and d). For the determination of the solvation free energies from these simulations, we used g_sas.\cite{57,58} For the REMD simulation of dialanine, we used 24 replicas in the temperature range from 300 to 400 K over 80 ns. We simulated the same system in the same temperature range using 8 and 4 replicas with our newly implemented techniques over 80 ns. For the simulations of AsLOV2 we used 6 replicas in the temperature range from 300 to 305 K. For the simulation in the NPT ensemble, we maintained the pressure using the Parrinello–Rahman barostat using a coupling constant of $\tau_p$ of 1.0 ps and a Nosé–Hoover thermostat using $\tau_t = 1.0$ ps. In the main text the abbreviation RMSD$_{C_a-C_a}$ stands for the root-mean square deviation of peptide backbone conformations from the backbone of the corresponding native structure (see Fig. 1). The probability distributions were determined through averaging over a 2-dimensional mesh with the radius of gyration ($R_g$) and the RMSD$_{C_a-C_a}$ as measures for this distribution. For the determination of RMSD$_{C_a-C_a}$ we used g_rms, while we used g_gyrate for the calculation of the radius of gyration.\cite{54} We determined the swapping rate using in-house code. The demultiplexed plots have been calculated using demux.pl.\cite{54} We determined the free energies using the relation

$$\Delta F = -k_B T \ln P / P_{\text{min}}$$

where $P$ stands for the probability and $P_{\text{min}}$ for the corresponding minimum probability in the distribution. For the measurement of the statistical inefficiency $s$ of property $A$, we used the relation for an averaging of $n_b$ blocks $b$:

$$\sigma^2(\langle A \rangle_b) = \frac{1}{n_b} \sum_{i=1}^{n_b} (\langle A \rangle_b - \langle A \rangle_{\text{run}})^2,$$

and

$$s = \lim_{\tau_b \to \infty} \frac{\tau_b \sigma^2(\langle A \rangle_b)}{\sigma^2(A)}.$$  

where $\langle A \rangle_b$ stands for the block average, $\tau_b$ stands for the simulation time block size and $\sigma^2(A)$ for the variance of property $A$.\cite{55}

3 Results

We start with the description of our results based on simulations of folding of TrpCage, TrpZip4 and GB1. Then, we present our results on the signaling behavior of AsLOV2. Finally, we show our results in simulations of dialanine and discuss the performance and convergence behavior of the 2 algorithms.

3.1 Folding simulations of TrpCage

We start with the description of folding simulations of TrpCage in Fig. 3. In this figure, we show the time-dependent RMSD-plots to the native structure of TrpCage and the probability surfaces as function of the RMSD$_{C_a-C_a}$ and the radius of gyration $R_g$. Simulations have been performed starting from the fully extended peptide, and we observe formation of the native structure (RMSD$_{C_a-C_a} < 0.35$ nm) within simulation times in the range below 20 ns (see Fig. 3a and c). In the case of the simulation using the $\Phi$-ensemble technique, we observe a fast collapse of the extended peptide chain within the first few ps of the simulation, with a simultaneous bending in the region of Ser13. This event is followed by the slow formation of the $\alpha$-helical part between Asn1 and Gly10 in the time-range from 1 ns to approximately 9 ns, while the 3–10 helix near Ser13 forms prior to the $\alpha$-helix between 4 and 8 ns (see Fig. 3a). The folding process is finalized by a closure of the hydrophobic core, when the Pro18 and Pro19 approach Trp6 (see Fig. 3a). The folding behavior is reflected by the probability surface as function of RMSD$_{C_a-C_a}$ and $R_g$. Here, we observe that the conformer with the $\alpha$-helical and the 3–10 helical parts dominates the free energy landscape (see Fig. 3b). In contrast, extended structures with RMSD$_{C_a-C_a} > 0.7$ nm are not observed.

In the folding simulation using KEMD, we do not find a fast collapse as in the folding simulation with the $\Phi$-ensemble technique. In contrast, we see that the peptide resides in the extended region for approximately 3.5 ns till it forms its complete tertiary fold at 4 ns (see Fig. 3c). We mention that the 3–10 helix in the region around Ser13 has already formed at this stage. Subsequently, the peptide resides in this state for approximately 10 ns till it partially forms its secondary structure in the form of the $\alpha$-helix and is close to its native form. In the
probability distribution of the same simulation, we find that the peptide resides in a broader near-native region than in the ϕ-ensemble simulation. Additionally, the extended ensemble is populated with a higher probability in this case (see Fig. 3d).

Both simulations show a folding profile along a diagonal in the space of the probability landscapes, i.e. the collapse of the polypeptide chain occurs with the direct formation of the tertiary structure of TrpCage. This indicates a strong preference for a nucleation condensation mechanism in both simulations. Using the 2 different methodologies, a fast formation of the tertiary fold is followed by a slow formation of the α-helix and the poly-proline helix. Additionally, Trp6 is oriented in the direction to the poly-proline helix already at the early stages of the simulations, which facilitates fast formation of the hydrophobic core. Juraszek et al. have studied the folding mechanism of TrpCage using REMD and transition path-sampling. They found that nucleation-condensation and diffusion-collision folding pathways are possible for this protein, while transitions between the main-intermediates are possible. Their findings agree with our results, while we only observe the nucleation condensation pathway due to the fast collapse in the beginning of the simulation. In a bias-exchange metadynamics study, Marinelli et al. investigated the kinetic pathways of TrpCage-folding, and found that condensed misfolded intermediates might lead to a slow-down of the folding reaction of TrpCage.

Although we did not find misfolded states in our simulations, our intermediates with a pre-formed α-helix and 3–10 helix agree well with their observations. In a recent study, we developed a hybrid kMC-MD algorithm for protein folding in explicit solvent and applied it on folding of TrpCage.11,12 We found that TrpCage folds along different pathways, with either mon-exponential or multi-exponential decay-behavior. In detail, we found that formation of the 3–10 helix occurred prior to folding of the α-helical part and formation of the hydrophobic core, which is in agreement with our results.

3.2 Folding simulations of TrpZip4

We continue to show our results on TrpZip4 in Fig. 4. In this figure, we show the RMSD plots and the probability surfaces from the folding simulations. We find that TrpZip4 remains in the unfolded region till approximately 16 ns MD-time until a first collapse occurs down to a RMSD of 0.5 nm (see Fig. 4a). In this collapsed state, the protein has a misaligned hydrophobic core in which Trp5 and Trp14 form a contact. Subsequently, the loop reopens and the protein remains at a RMSD of 0.3 nm and 0.37 nm, corresponding to different alignments of the hydrophobic core region. While the first maximum in the probability represents an intermediate state where Trp5 and Trp12 are in contact and the hydrophobic core is not compact, the second intermediate state represents a configuration with a compact hydrophobic core.  

Juraszek et al. have studied the folding mechanism of TrpCage using REMD and transition path-sampling. They found that nucleation-condensation and diffusion-collision folding pathways are possible for this protein, while transitions between the main-intermediates are possible. Their findings agree with our results, while we only observe the nucleation condensation pathway due to the fast collapse in the beginning of the simulation. In a bias-exchange metadynamics study, Marinelli et al. investigated the kinetic pathways of TrpCage-folding, and found that condensed misfolded intermediates might lead to a slow-down of the folding reaction of TrpCage. Although we did not find misfolded states in our simulations, our intermediates with a pre-formed α-helix and 3–10 helix agree well with their observations. In a recent study, we developed a hybrid kMC-MD algorithm for protein folding in explicit solvent and applied it on folding of TrpCage. We found that TrpCage folds along different pathways, with either mon-exponential or multi-exponential decay-behavior. In detail, we found that formation of the 3–10 helix occurred prior to folding of the α-helical part and formation of the hydrophobic core, which is in agreement with our results.

3.2 Folding simulations of TrpZip4

We continue to show our results on TrpZip4 in Fig. 4. In this figure, we show the RMSD plots and the probability surfaces from the folding simulations. We find that TrpZip4 remains in the unfolded region till approximately 16 ns MD-time until a first collapse occurs down to a RMSD of 0.5 nm (see Fig. 4a). In this collapsed state, the protein has a misaligned hydrophobic core in which Trp5 and Trp14 form a contact. Subsequently, the loop reopens and the protein remains at a RMSD of 0.3 nm and 0.37 nm, corresponding to different alignments of the hydrophobic core region. While the first maximum in the probability represents an intermediate state where Trp5 and Trp12 are in contact and the hydrophobic core is not compact, the second intermediate state represents a configuration with a compact hydrophobic core.
core region. We conclude that although the overall mechanism is 2 state like, we observe 2 different intermediate states in the folded state of TrpZip4 with different alignments of the hydrophobic core region.

We continue with our results on folding simulations of TrpZip4 using KEMD. In the plot of the RMSD $C_\alpha-C_\alpha$ as function of simulation time, we see that TrpZip4 folds along a monoexponential decay into its native structure (see Fig. 4c). We find that the protein forms a contact between Trp5 and Trp12 prior to the final formation of the hydrophobic core which is indicative of a zipper folding mechanism. As for the simulation with the $\Phi$-ensemble technique, we observe a monoexponential decay indicative of a 2-state folding mechanism. The overall relative MD-folding time is in good quantitative agreement with our prior result using the $\Phi$-ensemble technique. In the probability distribution we find that only one maximum exists along the folding landscape of TrpZip4 with one well defined hydrophobic core (see Fig. 4d). This indicates that only near native intermediates exist along the 2-state folding pathway of this protein.

In simulations where we used the recently developed kMC-MD method which employs move sets, we observed that TrpZip4 remains in the collapsed state (at an RMSD $C_\alpha-C_\alpha$ lower than 0.5 nm to the native structure) for comparably long periods ($3-5 \times 10^{-6}$ s) until it collapses to its folded state.\textsuperscript{12} The folding mechanism, which we observed in 5 independent simulations, involves formation of a hydrophobic contact between Trp3–Trp14, Trp3–Trp12 and Trp5–Trp14. Here, we observe formation of a contact between Trp5 and Trp12 as a first event upon folding in both simulations indicative of a zipper mechanism. Interestingly, we also observe fast formation of a near native collapsed state (using KEMD, see Fig. 4c) which is in agreement with our prior observation of a direct and fast folding pathway. In contrast to the prior simulations, we do not observe a heterogeneous intermediate region between the folded and the unfolded ensemble, but only one native maximum in the probability distribution, which indicates a 2-state folding pathway. We mention that a heterogeneous landscape could appear with a larger number of simulations.

Using both novel exchange methods in the NPT ensemble, we observe good agreement with experiments where an overall 2-state folding mechanism has been observed.\textsuperscript{1,60,61}

![Fig. 3 Results from simulations of folding of TrpCage ($\Phi$-ensemble). (a) RMSD $C_\alpha-C_\alpha$ as function of MD-time. (b) Probability surface as function of RMSD $C_\alpha-C_\alpha$ and the radius of gyration of the same simulation. Results from simulation of folding of TrpCage (KEMD). (c) RMSD $C_\alpha-C_\alpha$ as function of time-steps. (d) Probability surface as function of RMSD $C_\alpha-C_\alpha$ and the radius of gyration of the same simulation.](image-url)
Folding simulations of GB1

We continue with the description of our results on the folding simulations of GB1 using the \( \Phi \)-ensemble technique (see Fig. 5). Here, we show the RMSD \( C_\alpha-C_\alpha \) as function of simulation time and the probability distributions as function of RMSD \( C_\alpha-C_\alpha \) and \( R_g \). In the RMSD plot of the folding simulation of GB1, we see that GB1 collapses very fast within the first few ns MD-time to RMSD \( C_\alpha-C_\alpha \) of 0.32 nm but then reopens the loop within the next 5 ns MD-time (see Fig. 5a). Subsequently, the protein remains in the unfolded ensemble till it collapses a second time at a simulation time of 40 ns MD-time. We observe that the first contact for GB1 during loop closure occurs between Tyr5 and Phe12 which indicates a zipper collapse mechanism in this simulation. We do not observe a 2-state mechanism, because GB1 remains in the unfolded ensemble for a comparably long period. In the probability distribution, we find that GB1 has a heterogeneous folding landscape with 4 maxima close to the native state. We find a major maximum in the probability distribution at RMSD \( C_\alpha-C_\alpha \) of 0.5 nm and \( R_g \) of 0.75 nm, in which the loop is formed but no hydrophobic core exists. We find the folded maximum at RMSD \( C_\alpha-C_\alpha \) of 0.25 nm and \( R_g \) of 0.75 nm. Following the time-trace of folding of GB1, we observe folding along a heterogeneous folding landscape which cannot be fit to a mono-exponential function (see Fig. 5b), which indicates that folding occurs on a pathway with multiple minima along the free energy landscape.

We continue with the folding simulations on GB1 using KEMD. In this case, we see that the protein folds within 48 ns of MD-time. As in the simulation on GB1 using the \( \Phi \)-ensemble technique, GB1 stays for a comparably long period in the unfolded ensemble (43 ns MD-time) after a fast collapse which occurs within the first 2 ns MD-time (see Fig. 5c). The aggregated MD-time of folding is similar for both methods. In the probability distribution, we see that the major probability corresponds to the folded state, in which the hydrophobic core is formed (see Fig. 5d). We mention that we also observed partial helix formation in the case of the \( \Phi \)-ensemble technique in the region of Lys10 and Phe12.

In prior simulations based on kMC-MD with move-sets, we observed that comparably short transition times occur between the unfolded and the folded state of GB1.\(^{12}\) This finding agrees well with our simulations using both methods where the transition times are also in the range of few ns MD-time. In our prior simulation study on GB1, we also have observed that the folding landscapes of GB1 are broadened compared to that of TrpZip4, which agrees well with our present result. As in the case of the simulations using move-set based kMC-MD, we observe a higher probability in the region at RMSD \( C_\alpha-C_\alpha \) of 0.4 to 0.7 nm which agrees with our result. In contrast, we did not observe formation of a misfolded conformer with a partial helical fold in the region between Lys10 and Phe12.
3.4 Simulations of the early stages of signaling of the LOV2 domain from *Avena sativa*

We tested the behavior of folded protein systems in our development and simulated the structure of the light-oxygen voltage sensitive domain 2 (LOV2) domain from *Avena sativa* (AsLOV2) in its light and dark state using the Φ-ensemble technique. It is well accepted, that the flavin-mono-nucleotide (FMN) undergoes a photocycle after the excitation with blue light and binds to an adjacent cysteine residue 450 to form the light activated adduct state (in the electronic ground state). The system remains in this state for a time-range of many seconds until it converts back to its dark form.65 In simulations and experiments it has been shown that the protein undergoes major conformation changes in the light activated form leading to changes in the behavior of the organism, e.g. phototropism.66 In Fig. 6, we show the time dependent behavior of this protein as function of MD-time. In the dark state, we observe that Gln513 remains stable and remains H-bonded towards FMN-O4 throughout the simulation over more than 10 ns of MD-time (see Fig. 6a and b). In the light state, the sidechain of Gln513 switches and binds to Asn492 while the H-bond between Gln513-NE2 and CFN-O4 is cleaved (see Fig. 6a and c). For the distance between Gln513-NE2 and CFN-O4, we observe that the cleavage takes place within the first few ns of the simulation. Throughout the simulation, Gln513 remains at a distance ranging from 0.5 nm to 0.7 nm, which indicates that the H-bond between Gln513 and FMN-O4 is broken permanently (see Fig. 6a). We mention that an identical time-dependent behavior of Gln513 has been observed in many simulations before14,62–64 and is in agreement with experiments.67 In detail, this triggering event of signaling involving Gln513 has been observed in the first 20 ns of MD simulation for AsLOV214,62–64 and also other signaling proteins.68 We mention that there is a consensus about the function of this conserved glutamine and its role in tightening the β-strands H- and I- which leads to cleavage of the J-α helix and subsequent signaling of the adjacent kinase.69 We conclude that the assumptions on the behavior of our algorithm has the potential to describe complex signaling dynamics of folded proteins.

3.5 Performance and convergence analysis

For the comparison with conventional REMD, we simulated dialanine using REMD (24 replicas, 48 Intel® Xeon® E5-2670 cores), the Φ-ensemble (8 and 4 replicas, using 16 and 8 Intel® Xeon® E5-2670 cores) and KEMD (8 and 4 replicas, using 16 and 8 Intel® Xeon® E5-2670 cores) for a total integration time of 80 ns within the same temperature range from 300 to 400 K. In Fig. 7, we show the free energy landscapes at 300 K as function of the backbone dihedral angles Φ and Ψ for different simulation times. At a simulation time of 4000 ps, we find a
minimum of $-2.5k_BT$ at $\Phi = -80^\circ$, $\Psi = 150^\circ$ in all simulations. Additionally, we find that the region at $\Phi = -80^\circ$, $\Psi = 20^\circ$ is populated also in all simulations, while the population densities are different for the $\Phi$-ensemble simulation using 4 replicas. In this case, a population with $-0.45k_BT$ appears at $\Phi = 50^\circ$, $\Psi = 10^\circ$ (see Fig. 7a). At 12 ns, we find free energy minima in the range from $-2$ to $-3.5k_BT$ at $\Phi = -80^\circ$, $\Psi = 150^\circ - \Phi = -150^\circ$, $\Psi = 150^\circ$ and $\Phi = -80^\circ$, $\Psi = -20^\circ$ in the case of REMD and the 2 $\Phi$-ensemble simulations. For the KEMD simulations at the same simulation time, we find a stronger population of $-2.5k_BT$ at $\Phi = 50^\circ$, $\Psi = 20^\circ$, while the population at $\Phi = -80^\circ$, $\Psi = 20^\circ$ is higher in free energy at approximately $-1k_BT$ (see Fig. 7b). In the time range from 28 to 48 ns, we observe that the population-density at $\Phi = -80^\circ$, $\Psi = 150^\circ$ increases to a free energy value of $-4.5k_BT$ in all simulations, which is also the case for the minima at $-150^\circ$, $\Psi = 150^\circ$ and $\Phi = -80^\circ$, $\Psi = -20^\circ$. The free energy for the minimum at $\Phi = 50^\circ$, $\Psi = 20^\circ$ is approximately $-1.5k_BT$ for REMD and the 2 $\Phi$-ensemble simulations, while the KEMD simulations reach a value of $-2.7k_BT$ (see Fig. 7c and d). At 80 ns, we observe that all simulations converge to a free energy value of $-5k_BT$ for the minimum at $\Phi = -80^\circ$, $\Psi = 150^\circ$ and $-3.5k_BT$ for the minima at $\Phi = -150^\circ$, $\Psi = 150^\circ$ and $\Phi = -80^\circ$, $\Psi = -20^\circ$ (see Fig. 7e). We only find slight differences in the population at $\Phi = 50^\circ$, $\Psi = 20^\circ$, where the free energy varies between $-2.7k_BT$ and $-3k_BT$. In comparison with the REMD result, we conclude that both methods converge to the same free energy landscape. This peptide has been investigated in a number of computational studies, and free energies and the related populations reported here at 80 ns are in agreement with these results. The ratio of computational cost of REMD to the novel techniques is 48:16 and 48:8 in simulations, while the same statistical average for dialanine has been obtained.

We continue with a comparison of the swapping efficiency in the simulation of dialanine in Fig. 8. We define this efficiency as the average number of exchange attempts in which the lowest replica with index #0 has visited the highest temperature and returns back to its original index #0. All simulations have been performed with dialanine within the same temperature range. We find an average number of 816.81 exchange attempts (817 ps for one attempt every 1000 integration steps) for conventional REMD for one complete cycle. For KEMD, where we used 8 and 4 replicas, we find an average number of 210.23 and 78.63 attempts for one complete cycle. In the case of the $\Phi$-ensemble we measured a swapping efficiency of 13.43 for the simulation with 8 replicas and 5.3 for the simulation using 4 replicas. Using the measure of the efficiency, with which frequency the replicas can travel within temperature space, KEMD is 3.8 to 10 times more efficient than REMD, while the $\Phi$-ensemble achieves a swapping rate which is 60 to 154 times larger than the exchange rate in REMD (see Fig. 8a and f–j).

Next, we consider the statistical efficiency of the different algorithms to test the relative rate of the 2 methods to reach statistically independent configurations in comparison with REMD (see Fig. 8b). We used the $\Phi$-angle at 300 K from the 80 ns simulations of dialanine as function of simulation time and determined the statistical inefficiency $s$ (see section Methods). We find that the number of steps needed to reach new statistically independent configurations is approximately 2 times larger for the $\Phi$-ensemble (2000 steps) and 3 to 4.5 (4500 steps) times larger in KEMD than the statistical inefficiency in the case of the REMD technique. We find that REMD needs approximately 1000 steps ($s = 1$) to reach a new statistically independent configuration. That means that the propagation within the $\Phi$-ensemble and in KEMD contains more correlated information than it is the case for REMD, which is caused by the different exchange schemes implemented for the $\Phi$-ensemble and KEMD. When we compare the $\Phi$-ensemble technique with KEMD, we find that the statistical inefficiency for the $\Phi$-ensemble technique remains approximately constant at a value of 2, independent from the number of replicas. Clearly, this is a
consequence from the global exchange applied in this method. In contrast, the statistical inefficiency of KEMD depends on the selected number of groups and the number of replicas (see Fig. 8b).

As a measure for the rate of formation of the native state in our simulations, we used the RMSD to the backbone of the native structure. Using TrpZip4 as our test system, we initiated simulations from extended conformations using 6, 12 and 32 replicas over the same temperature range of 300 to 410 K (see section Methods/Simulation parameters and system preparation). For both, the $\Phi$-ensemble and KEMD, we find that TrpZip4 folds within different relative folding MD-times, as we see in the RMSD$_{C_{\alpha}-C_{\alpha}}$ as function of simulation time (see Fig. 8c). We measured the efficiency of each replica exchange implementation using the product of the total walltime, the number of replicas and the number of processors used until RMSD$_{C_{\alpha}-C_{\alpha}} < 0.3$ nm was reached. In terms of the conventional REMD sampling, we made the estimate that (using a mono-exponential least square fit) it converges to the native state after approx. 150 to 250 ns, while the distribution has not converged within the actual simulation time of 50 ns using 32 replicas (see Fig. 8d). We note, that the scaling between both methods is approximately equivalent at an equal number of replicas. In this case, if compared to conventional REMD, the computational expense to form the native state of TrpZip4 (the product of walltime, number of replicas and number of CPUs used – Intel® Xeon® E5-2670) can be lower by a factor of 168 for both methods (see Fig. 8e).

4 Conclusions
In this work, we developed 2 new coarse REMD algorithms, which coarse-grain the replica space based on approximations.
These algorithms enhance the sampling in simulations of peptide folding and protein signaling through an elevated swapping efficiency and a reduction of the number of needed replicas in a wide temperature range. We use a kMC-type approach for the coarse exchange of replicas, in contrast to the standard REMD method.\textsuperscript{15,16} While the $\Phi$-ensemble technique is based on a fluctuation criterion $\Phi$ in each replica followed by an exchange over all other replicas, kinetic exchange MD (KEMD) is based on an indexing of groups in which a whole set of replicas within one group becomes exchanged if the event is selected for execution. Through these implementations we obtain results which are in agreement with results obtained with the kMC-MD method which uses discrete moves.\textsuperscript{11,12} In an analysis of the performance and convergence behavior of new algorithms, we find that the algorithms converge to the same statistical average for dialanine as conventional REMD, while both methods are computationally 6 times more efficient. Considering the number of exchange attempts needed for the replicas to swap in temperature space, we find that both methods significantly enhance the exchange within low and high replica indexes. Clearly, the $\Phi$-ensemble technique has a higher swapping rate in comparison with KEMD. We note that the statistical efficiency of both methods is lower than for conventional REMD, due to an increase in statistical correlation between the replicas resulting from the use of the novel exchange scheme. While the statistical inefficiency $s$ of the $\Phi$-ensemble is approximately independent from the selected number of replicas, we find a dependency of $s$ on the selected number of groups and replicas in the case of KEMD. Additionally, we considered the computational cost to simulate the formation of the native state of TrpZip4, and found that the product of the computational time to reach the native state, the number of replicas and the number of cores used for the simulation can be 168 times lower than for conventional REMD due to an enhancement of the swapping rate in temperature space. The new methods can be used in simulations, where an enhanced exploration of free energy landscapes in explicit solvent is needed at low computational cost. We emphasize the applicability of our new approach in folding simulations and fast conformation sampling of large protein systems. At the same time, the number of replicas can be reduced significantly. A kinetic interpretation of our results based on this development might be possible on the basis of histogram techniques.\textsuperscript{72}

In our simulations of folding of TrpCage, GB1 and TrpZip4, we observe that both proteins fold into their native structure. For TrpCage, we observed prior formation of the tertiary structure and subsequent slow formation of the $\alpha$-helix, while the 3–10 helix has formed prior. This pathway is consistent with experiments\textsuperscript{58,73} and simulations.\textsuperscript{11,12,56–59,74}

In the case of GB1, we see that the protein shows a broadened folding landscape. Juraszek et al. investigated the differences in folding of GB1 and TrpZip4, and they demonstrated that the transition state energy is higher for TrpZip4, leading to a longer folding time of this peptide compared to GB1.\textsuperscript{75} This is in agreement with our observation of longer transition times for TrpZip4 from the unfolded ensemble to the near native state, while we notice that the overall folding MD-time of TrpZip4 is lower than for GB1. For both proteins, we observe a zipper pathway of folding, while for GB1, we observe misfolded states in 2 cases, which is in agreement with Bonomi et al. who found a similar misfolded state during folding of GB1.\textsuperscript{76} It is much debated in computational and experimental studies, whether GB1 folds along a zipper like mechanism or a hydrophobic collapse pathway.\textsuperscript{10,77–84} In another work, Sancho et al. analyzed the folding kinetics of GB1 and found that depending on the initial conditions, the protein folds either along a 2-state or a multistate pathway, the latter connected to slow dynamics in the unfolded region and a heterogeneous intermediate state.\textsuperscript{85} This is in agreement with our simulations,
and we note that the folding pathways (and the misfolded intermediates) might not occur with different starting conditions. In simulations on the signaling behavior of AsLOV2, we find that we accurately can describe the conformation changes related to signaling in the chromophore binding pocket, when we compare our results with prior studies on the same system.\textsuperscript{14,62–64,67,69}

Finally, we emphasize that our new method is free of any choice of move sets and enhances the formation of the native structures of larger proteins.

Acknowledgements

EKP and IVP acknowledge support by a grant from the Swiss National Supercomputer Center (CSCS) under the Platform for Advanced Scientific Computing. Simulations were carried out at Swiss National Supercomputer Center (CSCS) under the projects u4 and s583. JES acknowledges support of the NSF grant no. MCB-1158577 and the David and Lucile Packard Foundation.

References