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

Back-mapping based sampling: Coarse grained free energy landscapes as a guideline for atomistic exploration

Cite as: J. Chem. Phys. 151, 154102 (2019); doi: 10.1063/1.5115398
Submitted: 18 June 2019 • Accepted: 29 September 2019 •
Published Online: 15 October 2019

View Online Export Citation CrossMark

Simon Hunkler, Tobias Lemke, Christine Peter, and Oleksandra Kukharenko[a)]

AFFILIATIONS

University of Konstanz, Konstanz, Germany

[a)]oleksandra.kukharenko@uni-konstanz.de

ABSTRACT

One ongoing topic of research in MD simulations is how to enable sampling to chemically and biologically relevant time scales. We address this question by introducing a back-mapping based sampling (BMBS) that combines multiple aspects of different sampling techniques. BMBS uses coarse grained (CG) free energy surfaces (FESs) and dimensionality reduction to initiate new atomistic simulations. These new simulations are started from atomistic conformations that were back-mapped from CG points all over the FES in order to sample the entire accessible phase space as fast as possible. In the context of BMBS, we address relevant back-mapping related questions like where to start the back-mapping from and how to judge the atomistic ensemble that results from the BMBS. The latter is done with the use of the earth mover's distance, which allows us to quantitatively compare distributions of CG and atomistic ensembles. By using this metric, we can also show that the BMBS is able to correct inaccuracies of the CG model. In this paper, BMBS is applied to a just recently introduced neural network (NN) based approach for a radical coarse graining to predict free energy surfaces for oligopeptides. The BMBS scheme back-maps these FESs to the atomistic scale, justifying and complementing the proposed NN based CG approach. The efficiency benefit of the algorithm scales with the length of the oligomer. Already for the heptamers, the algorithm is about one order of magnitude faster in sampling compared to a standard MD simulation.

I. INTRODUCTION

Nowadays, molecular dynamics (MD) simulations are extensively used to study the conformational ensemble of proteins[1,2] and their complexes.[3] However, to this day, one big shortcoming of MD simulations is the discrepancy between computationally accessible and chemically or biologically relevant time and length scales. A major reason for this drawback is a slow sampling of phase space of most molecules, since the MD simulation is repeatedly trapped in stable or metastable states. The energy landscapes, especially of biomolecules, are often quite rough, where different states are separated by relatively high energy barriers. These barriers have to be overcome in order for the MD simulation to sample new conformations.[4]

There are different types of methods that intend to improve the conformational sampling: biasing methods (e.g., umbrella sampling,[5] metadynamics,[6] or variationally enhanced sampling[7]), expansion schemes,[8,9] and coarse graining.[10–13] Biasing methods generally use additional external potentials in order to drive the system toward unsampled areas of phase space. Biasing potentials are usually defined in the space of a few (usually one, two, or three) collective variables (CVs), and the success of the sampling depends on a proper choice of the low-dimensional descriptors. Expansion schemes, on the other hand, typically rely on a smart choice of starting conformations for new simulations to sample more efficiently. Here, all the simulations are run without introducing any biasing terms to the potential, but are biased by the choice of initial structures and have to be simulated long enough to sample the correct statistical ensemble. Finally, coarse graining enables much faster sampling compared to atomistic simulations by reducing the number of degrees of freedom in a system. A CG model is much faster but it is always partially inaccurate, since in the process of coarse graining, transferability and representability limitations are necessarily introduced.

In this work, we combined advantages of all three methods into a scheme which is called back-mapping based sampling (BMBS). This scheme identifies structures that represent the CG sampling of conformational phase space according to their probability in CG space, starting structures for atomistic runs—by providing a quantitative comparison of the obtained landscapes at the different resolution levels—and allows for an assessment of the obtained sampling and its convergence. The combination of methods helps us to overcome drawbacks of the individual techniques and simulation models. The method takes advantage of three resolution levels: atomistic, CG, and a 2-dimensional (2D) projection of the conformational phase space (see Fig. 1) and is able to significantly accelerate the conformational sampling of oligopeptides. BMBS contributes to the family of multiscale methods that exploit a natural idea to use coarse graining to enhance the atomistic sampling of a system. This can be done, for example, by seeding atomistic simulations with CG structures,[14,15] by simulating simultaneously at two resolution levels,[16,17] or by exchanging directly between low-resolution and high-resolution replicas of a system.[18,19] Another approach is to combine an atomistic model with the accelerated dynamics of the CG degrees of freedom via an extra coupling potential, together with a Hamiltonian replica exchange to remove the biasing potential from the coupling term by exchanging the coupling parameters.[20] BMBS adds to these approaches by mapping between CG conformations of complex systems—even for radical coarse graining methods—and a higher resolution model via a third resolution level (2D), which allows us to bridge possible discontinuities in free energy landscapes. The efficiency of the proposed algorithm increases with the length of the oligomer chain, but already for a heptapeptide, the BMBS scheme is about one order of magnitude faster in sampling at least the same amount of phase space as a brute-force atomistic MD simulation.

BMBS relies on a CG ensemble. The way this ensemble was produced and the used CG model is, in principle, irrelevant as long as it reproduces the general atomistic behavior reasonably well. In recent years, more and more machine learning methods have been introduced in the MD community to analyze or accelerate coarse grained simulations.[21,22] In this paper, we rely on CG data produced by Lemke and Peter.[23] They applied a quite radical coarse graining procedure together with a neural networks (NNs) based approach to predict free energy surfaces (FESs) of aspartic (Asp) and glutamic (Glu) acid oligomers corresponding to the conformational sampling of a model with a very coarse representation. In their approach, a convolutional NN was trained on atomistic simulations of hexamers, heptamers, and octamers of aspartic or glutamic acid, respectively. High dimensional FESs of longer oligomers, which were not part of the training, were predicted. These high-dimensional FESs were then used as the Hamiltonian in Metropolis MC simulations, which generated CG ensembles, where the peptide conformations were represented by only one bead for each carboxylic acid group in the side chains. To compare atomistic and CG ensembles, CVs suitable for both resolution levels need to be chosen. The only CVs that can be used are those that derive from the CG positions. Therefore, it is not *a priori* obvious how to judge whether the CG datasets actually correctly represent the accessible backbone conformations of these oligomers. Also, at this point, one cannot be sure if the FESs that are predicted for the longer chains, which were not part of the training data for the NN, are actually correct and if there is a way to correct for a bias due to deficiencies of the CG model (here, the NN potential). In the following, we will show that BMBS is able to address these issues.

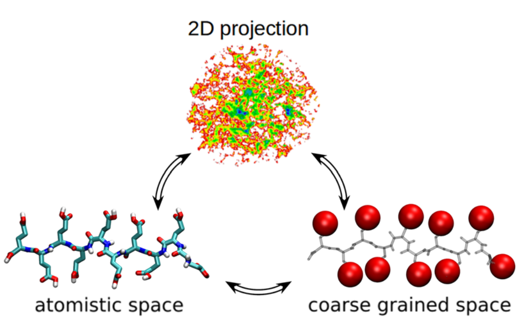

FIG. 1. Illustrative scheme of the three resolution levels used by BMBS: atomistic space, where the positions of all atoms are known; coarse grained space, where only the positions of the carboxylic acid groups in the side chains are present; and 2D projections.

An essential part in the BMBS is the back-mapping step where an atomistic resolution is reintroduced to the selected CG conformations. The challenge of how to do the back-mapping has already been addressed for quite some time.[24–26] One of the issues that still remains is how to use back-mapping quantitatively. A possible approach to this was presented by Chen and Roux.[19] They map atomistic structures to a CG model, propagate the CG model for a short time, and then do a back-mapping step via a nonequilibrium MD simulation, which drives the initial atomistic structure toward the newly sampled CG conformation. The resulting atomistic conformation is then accepted or rejected based on the Metropolis criterion. We present another approach on how to use back-mapping quantitatively: the atomistic exploration of a system is guided based on a 2D free energy landscape and the subsequent development of the atomistic sampling is monitored. Since the CVs describing the full CG FES are still very high-dimensional (although not as high dimensional as the atomistic representation), BMBS starts with the projection of the CG CVs to a low-dimensional (usually 2D) landscape. This is then used as a template to initiate new atomistic simulations. Points from all over the 2D landscapes are selected and back-mapped to an atomistic resolution by restrained MD simulations where additional external potentials drive an atomistic simulation toward the CG data points. These back-mapped conformations are then used to initiate new free atomistic simulations and thereby explore new areas of phase space. In the context of the BMBS, we will show how to decide where to initiate the back-mapping and how to judge the resulting atomistic ensemble using a metric called earth mover's distance (EMD) (which was already successfully applied for the comparison of FESs of CG and atomistic simulations[15]). Finally, we will use these EMDs to show that the BMBS-constructed atomistic ensemble is able to repair inaccuracies in the CG template.

II. METHODS

A. Test systems

We use aspartic (Asp) and glutamic (Glu) acid oligomers of different lengths as first test systems for the BMBS scheme. Even though the only difference between the two systems is an additional methylene group in the side chains, they show very different behavior: while the Asp oligomers are more intrinsically disordered and switch between different conformations much more frequently, the Glu oligomers form very stable alpha-helical structures and reside in these states for extended periods of time. Therefore, especially for the Glu oligomers, long standard MD simulations are required to overcome the free energy barriers to enter and leave the helical states and sample new areas of phase space.

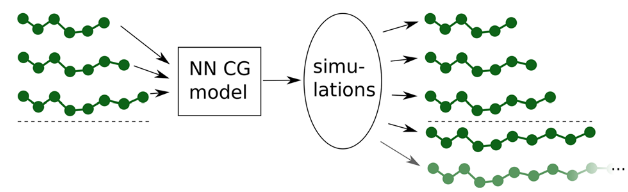

FIG. 2. Overview of the machine learning procedure used in Ref. 23. The model is parameterized on the basis of hexamer, heptamer, and octamer data. Simulations with this model can also be performed for longer chains.

B. Three resolution levels

1. Atomistic MD simulations

All simulations were performed using the 2016.4 version of GROMACS[27] package with the GROMOS 54A7[28] force field and the simple point charge (SPC) water model.[29] If not stated otherwise, the following settings were used: the time step was set to 2 fs, the energy minimization was done using the steepest decent algorithm, the temperature was set to 300 K using the velocity rescale thermostat,[30] and the pressure was set to 1 bar with the Berendsen barostat.[31] As an integrator algorithm, the leap-frog algorithm was used.[32] Long-range interactions were computed with the particle mash Ewald method,[33] where a Fourier grid spacing of 0.12 nm and a pme-order of 4 were used. For Coulomb and van der Waals interactions, a cut-off of 1.4 nm was used. In order to constrain all bonds, the LINCS algorithm was applied.[34] Box parameters and the number of water molecules for each oligomer are provided in the supplementary material.

2. NN CG model

For the CG representations, a model was used where only the carboxylic acid C-atoms in the side chains are represented as CG centers. All other atoms, e.g., the backbone, are not explicitly represented. To obtain an accurate description even with such a drastically simplified representation, sophisticated high-dimensional interactions between these centers have to be obtained. Lemke and Peter[23] used a neural network based approach to obtain such high-dimensional FESs from molecular dynamics trajectories. In this work, we use their CG models and simulation data for Asp and Glu oligomers.

As depicted in Fig. 2, the CG model was parameterized on the basis of data from different chain lengths (atomistic simulations of hexamers, heptamers, and octamers in water). With the obtained model, Monte Carlo simulations of different chain lengths were performed. The CG datasets consist of 256 000 frames from 512 parallel MC simulations for each oligomer. All 512 simulations were run in parallel on a graphics processing unit (GPU) (Nvidia Geforce GTX 1060), and it took 6–8 h to complete the computations. For a more thorough description of the coarse-graining procedure, we refer to the original paper by Lemke and Peter.[23] The CG simulations of the hexamer, heptamer, and octamer allow us to judge how accurately the model reproduces the behavior of chains that were part of the reference data. The CG simulation of the decamer allows us to judge how well the model predicts conformations for a longer chain which was not part of the data used for the parameterization. In Ref. 23, the validation of the NN CG model had been done by comparing distributions of different internal distances as well as the end to end distances to atomistic data. It was found that the overall features of these distributions are very well reproduced by the NN CG model. Yet, it is difficult to judge the sampling of a high dimensional conformational phase space based on distributions of individual collective variables (internal distances). In the present work, we will present a more detailed assessment in the course of the BMBS scheme.

3. 2D projection

To link the CG and the atomistic resolutions and to navigate in the high-dimensional CG FESs of the side-chain (carboxylic-acid) positions, we applied a dimensionality reduction to 2D space. This can be any projection method, e.g. PCA,[35,36] TICA,[37–39] diffusion-maps,[40,41] a multidimensional scaling method such as a sketch-map,[42] or an autoencoder-based scheme.[43,44] An advantage of linear transformations such as PCA and TICA is the knowledge of transformation matrices, but the projection quality can be limited in 2D compared to nonlinear methods. This may be crucial for systems with a high degree of disorder and a large number of (relatively low-populated) states. Diffusion maps is a good choice of a nonlinear dimensionality reduction, but the dynamics information needed for a good projection is lost during the MC steps in the present case; moreover, the algorithm complexity is $\mathcal{O}(n^3)$, where n is a number of frames in the simulation, so the calculations can be very expensive for long trajectories.

In this work, we used the sketch-map algorithm for dimensionality reduction, which is a nonlinear, multidimensional scaling approach that aims for preserving relative distances during the projection from high to low dimensional space with a focus on an intermediate range of distances between data points (i.e., conformations). Importantly, the sketch-map can be very efficiently used for very large sets of data points: the generation of the map is done based on a subset of representative data points (landmarks). All other points are projected based on their relative positions to those landmarks. One advantage of this algorithm is that it was found to produce meaningful 2D projections for disordered systems.[8] In the following, we will call a once obtained 2D projection for a given set of landmarks a "sketch-map space." Most importantly for the present application is the fact that we can project CG and the atomistic data to the same sketch-map space and thus compare the resulting probability distributions (2D histograms) or FESs, respectively (the latter generated by Boltzmann inverting the probability distributions).

As CVs we used throughout all of this work all pairwise distances between the CG centers which were calculated for each frame of the CG dataset. Table I gives the number of CVs for the different oligomers.

We should also underline that the proposed scheme is not limited to 2D projection, especially when one uses other projection techniques, which are not so efficient in producing informative enough 2D representation. However, increasing the number of dimensions may increase the respective computational cost and can complicate assessments of the results.

C. Back-mapping based enhanced sampling scheme

The idea behind the BMBS is to use two-dimensional free energy landscapes of very efficiently produced CG ensembles as a template on where to initiate new atomistic simulations in order to sample all of the accessible phase space as fast as possible. In Fig. 3, a short summary of the individual steps of the BMBS is given.

As we already mentioned in Sec. II B 2, the CG datasets were produced by metropolis MC simulations. From there, the very first step in the BMBS scheme is to project the CG data to 2D. Based on the 2D projection of the CG datasets, we determine the first target structure as the structure with the highest probability in CG space (the highest density peak of the obtained projection). Here, we used the binning approach described in Ref. 8 to define the bin with the highest population. One random point in it is chosen as the first target structure.

In the next step, an atomistic resolution is reintroduced to the CG target structures. To do this, a short 1 ns atomistic MD simulation is run. It starts from an energy minimized structure of the respective oligomer. This simulation is called the "initial simulation" and includes the following steps: solvation, equilibration, energy minimization, temperature tuning, pressure tuning, and production. In principle, the initial simulation is only needed for choosing an atomistic conformation to start the back-mapping from and can be generated in any other way. For more complex systems, one can, in principle, also use an ensemble of atomistic starting structures from different regions of phase space. This initial simulation is projected to the 2D sketch-map space of the CG CVs: the same landmarks are used for both projections, which implies that both sets of CVs are projected to the same sketch-map space and therefore can be easily compared.

In the consecutive step, a point from the 2D projection of the initial atomistic simulation with the minimal Euclidean distance to the target structure in the 2D sketch-map space is identified. This atomistic conformation is used as a starting structure for a distance restrained simulation, where the distances between the C-gammas/C-deltas are restrained to those of the CG target

TABLE I. Dimensionality of the descriptor of oligomer conformations, given by the number of pairwise distances between the terminal C-atoms in the side chains.

Oligomer	Dimensionality
Heptamer	21
Octamer	28
Decamer	45

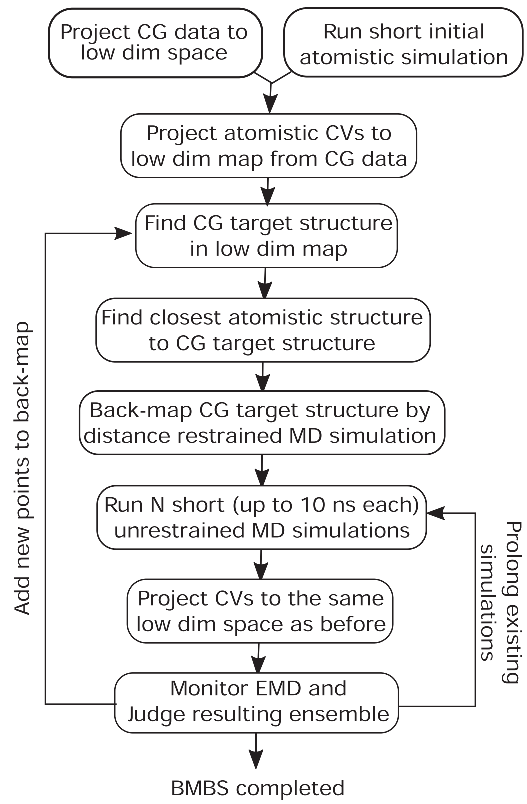

FIG. 3. Flowchart giving an overview of the individual steps of the BMBS scheme.

configuration that corresponds to the target point in the 2D sketch-map space. Distance restraint simulations were performed for 50 ps with a 0.5 fs time step. In these simulations, all pairwise distances between terminal C-atoms of the side chains are restrained to the values that correspond to the target structure. Practically, this is a way to back-map a CG configuration to atomistic resolution. If the atomistic structure from which the restraint simulation is started differs significantly from the target structure, the simulation might potentially crash due to the forces from the restraining potentials. In such cases, one possible way to prevent this is to use gradually increasing force constants in order to give the atomistic conformations more time to assume the enforced distances between the terminal C-atoms in the side chains.

Following this back-mapping procedure, a short (few nanoseconds) unrestrained atomistic simulation is started, in order to explore an area where the coarse grained projection is highly populated. After the simulation, the CVs of the structures visited in that unrestrained trajectory are again projected to the sketch-map space using the same landmarks as before. These structures are then also added to the pool of atomistic structures from which the nearest point in the sketch-map space to the next CG target configuration can be found. The intention behind this increasing pool of structures is to make sure that the back-mapping step is always started from the most similar atomistic conformation to the CG target configuration that is available to this point.

For all oligomers, we have performed 100 back-mappings, followed by the corresponding unrestrained 1 ns–8 ns simulation.

We have included information about the statistical weights of the selected target structures corresponding to the probability density of the CG sketch-map projection: the number of target structures randomly selected from a specific bin and used to start simulations is proportional to that bin's probability. Once the amount of simulations that correspond to the probability of the most populated bin are performed, the same procedure is applied to the second most populated bin and so on. All CG target structures are in each case found by randomly choosing a data point from the corresponding bin. The collective CVs of all simulations are projected to the sketch-map space.

For all oligomers, the 100 simulations were parallelized so that always 4 simulations were run at once and the calculations for each simulation were done on 10 central processing unit (CPU) cores and a GPU unit (GeForce GTX 1080). For the heptamers, this gave a simulation speed of around 200 ns/day and for the decamers around 180 ns/day, resulting in ~24 h for the heptamers and ~27 h for the decamers. However, this simulation time can easily be further reduced by using more cores, computing on multiple GPUs, and by increasing the number of simulations that are run parallelly.

In the following, more details regarding the restrained simulations are given. After selecting a target structure from the CG CVs and finding the atomistic data point in the sketch-map space with the smallest Euclidean distance to it, an atomistic simulation starting from this point is directed toward the target structure with the help of restraining potentials. These restraining potentials act on all pairwise distances between terminal C-atoms of the side chains. The functional form of the restraining potential is given in Eq. (1) and illustrated in Fig. 4,

$$V_{dr}(r_{ij}) = \begin{cases} \frac{1}{2}k_{dr}(r_{ij}-r_0)^2, \ r_{ij} < r_0, \\ 0, \ r_0 \le r_{ij} < r_1, \\ \frac{1}{2}k_{dr}(r_{ij}-r_1)^2, \ r_1 \le r_{ij} < r_2, \\ \frac{1}{2}k_{dr}(r_2-r_1)(2r_{ij}-r_2-r_1), \ r_2 \le r_{ij}. \end{cases} \quad (1)$$

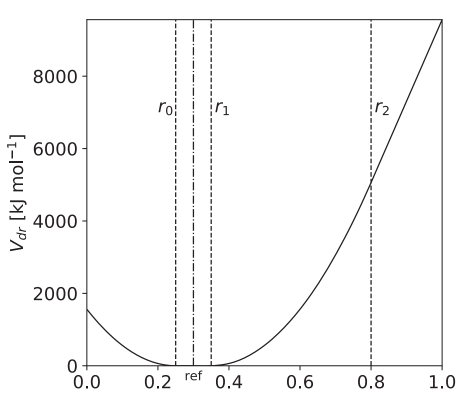

FIG. 4. The restraining potential applied during the back-mapping [see Eq. (1)]. r_{ref} indicates the reference distance of the target structure which will be back-mapped.

Here, V_{dr} is the distance restraining potential, r_{ij} is the pairwise distance between (terminal C-)atoms i and j, and k_{dr} is the force constant. The potential has a minimum well with the respective reference distance r_{ref} from the CG target structure at its center (see Fig. 4). The potential is harmonic below a user specified lower bound r_0 and between two upper bounds r_1 and r_2, zero if the distance is between the lower bound and the first upper bound, and linear beyond the second upper bound. The bounds are chosen such that the lower and the first upper bounds are 0.05 nm below and above the respective reference distance from the CG target structure and the second upper bound was set to 0.5 nm above the CG target distance. The force constant k_{dr} was set to a high value of 50 000 kJ mol^{-1} nm^{-2} to guarantee a fast and strict back-mapping.

It should be noted that the chosen parameters of the restraining potential were found to be suitable for the oligomers considered here. Further refinement of the parameters might be needed for the back-mapping of longer oligomers. For example, it might be necessary to adapt the range of distances in which the potential is harmonic or linear. To be more specific, the range of distances in which the potential is linear might need to be changed in order to appropriately cap the force if the pairwise distances deviate very much from those of the target structure. Moreover, it may be necessary to alter the potential slightly by adding another linear term for very short distances, equivalently to the linear term for very large ones. The latter would result in a force capping if the distances are far smaller than the ones of the target structure.

D. Earth mover's distance

To assess the quality of the enhanced sampling and to be able to qualitatively compare different FESs, we needed to introduce some measure of the difference (also termed distance) between two probability distributions. Important criteria for such a measure are as follows: it should be symmetric, should work for multidimensional distributions, and should not rely on any reference distribution. There are numerous different probability metrics/distances;[45] here, we used the so called the earth mover's distance (EMD), also known as Wasserstein's metric or the Mallows distance, which is mostly used in the image recognition community.[46–50] Unlike the Kullback-Leibler divergence and the statistical χ^2 distance, which are widely used in the simulation community, EMD is a cross-bin distance and is not affected by binning differences.

The idea behind the EMD is to envision a distribution as earth that is spread unevenly over a certain space, while a second distribution can be seen as holes spread in the same space. Then, the EMD describes the minimal amount of work that has to be done in order to fill the holes with earth or in other words to transform one distribution into the other. For a formal definition, see, e.g., Ref. 50. The disadvantage of the EMD is the $\mathcal{O}(n^3 \log n)$ complexity of the calculations, where n is the number of bins. Here, we used an accelerated version as proposed in Refs. 48 and 49, which is an order of magnitude faster than the original algorithm.

For an easier interpretation of the results, we used unity-based normalized EMDs to bring all values into the range [0,1],

$$EMD' = \frac{EMD - \min(EMD)}{\max(EMD) - \min(EMD)}, \quad (2)$$

with $\min(EMD) = 0$ and $\max(EMD) = 18.36$.

The normalization coefficient max(*EMD*) is defined as the EMD for the comparison of the histogram in the 2D sketch-map space of the 5 μs atomistic simulation of Glu 10 and a 2D uniform distribution with the same number of data points. The data points for this uniform distribution are spread evenly over a rectangular area whose dimensions are given by the minimum and maximum x and y values of the projection of the 5 μs atomistic simulation to the 2D sketch-map space. In this implementation, an EMD of 0 means that two histograms are identical and an EMD of 1 means that the two considered histograms are as different as a uniformly distributed dataset compared to the projection of the 5 μs atomistic simulation of Glu 10.

III. RESULTS AND DISCUSSION

We applied the proposed back-mapping-based sampling to Glu and Asp oligomers with 6, 7, 8, and 10 residues. In Subsections III A–III C, we illustrate results mainly for the heptamers and decamers, and additional results are given in the supplementary material.

A. Back-mapping based sampling of Glu and Asp heptamers and decamers

In order to illustrate the back-mapping procedure, Glu 7 is used as an example. The first backmapping step after the initial atomistic simulation of the BMBS is described in detail. Figure 5 shows in grayscale the 2D projection of the CG MC simulation of Glu 7 in the sketch-map space. The location of the initial (1 ns) atomistic simulation is marked in purple, the first CG target structure is highlighted with the yellow triangle, and the closest data point from the initial atomistic simulation to the CG target structure is marked with the blue square. Since the blue point has the lowest Euclidean distance in the sketch-map space to the CG target structure, its atomistic conformation is used to start the distance restrained MD simulation. The projection of a final structure of the distance restrained simulation is shown as the green data point in Fig. 5. The target structure (yellow) and the backmapped atomistic structure (green) lie almost on top of each other in the sketch-map space, which indicates that the backmapping step was successful and the proposed approach is capable to reintroduce an atomistic resolution to the selected CG conformations. From this atomistic data point, free atomistic simulations were performed and added to the (growing) pool of atomistic structures. In a similar manner, all the other starting points for the free exploring simulations are produced.

We used the 2D sketch-map projections to assess the degree of sampling obtained with the BMBS scheme and compared it to the 5 μs atomistic simulations and the CG data. In Fig. 6, the respective FESs are depicted for Asp and Glu heptamers and decamers. For the BMBS scheme, 100 simulations with 8 ns each were performed (800 ns in total) for every oligomer. Figure 6 shows an excellent agreement for the heptamers between the FESs from all three sampling schemes, the (long) free atomistic simulation, the MC simulation with the NN CG model, and the atomistic data from BMBS. The same areas of phase space are sampled with very similar distributions. Especially, the deepest free energy minima match very well in shape and location. The same holds true for the hexamers and octomers as shown in Fig. SI 1. The agreement of the oligomer conformations between the CG FES and the two atomistic ones should, in particular, be noted. While in Ref. 23, end-to-end and internal distance distributions within the oligomers had been employed to assess the quality of the CG model, the 2D FESs used here are not only sensitive to differences in individual distances but also to correlations. Thus, they are better suited to point out possible differences in the sampling of overall conformations of the oligomers and provide further validation of the NN CG model. The FESs of the BMBS simulations show that the scheme efficiently produces an atomistic conformational ensemble from the CG data. The agreement with the long atomistic simulation is very good. A comparison of the sampling obtained with 100 ns total simulation time of BMBS with an equivalent length of free atomistic MD simulations is presented in Fig. SI 2. It demonstrates the computational efficiency of the scheme which rapidly explores the conformational phase space with weights on the different phase space regions according to the CG model.

So far, the discussed data refer to the three—comparatively short—oligomers that had been part of the training data for the NN CG model. The cases of the decamers are more interesting, as the decamer data were not used during the parameterization of the CG model. In Ref. 23, the decamers had served as a test system to get an impression of the ability of the NN to *predict* oligomer

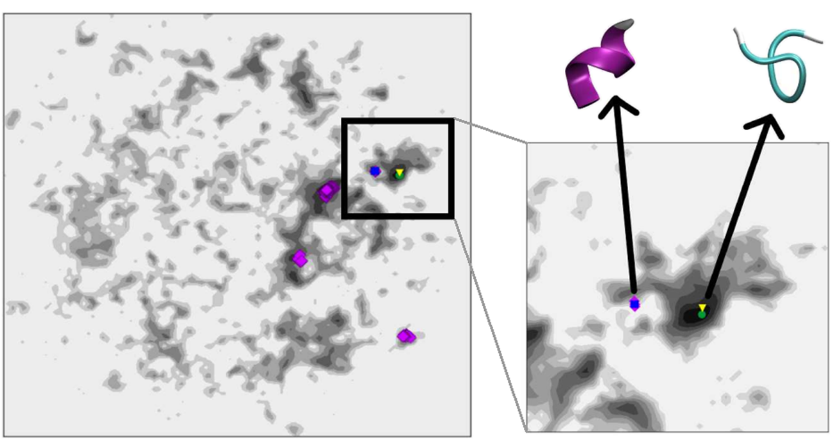

FIG. 5. 2D histogram of CG MC simulation of Glu 7 in the sketch-map space (gray). Purple data points belong to the initial atomistic simulation, the yellow point is the coarse grained target structure, the blue data point is the nearest point of the initial atomistic simulation to the target structure, and the green point corresponds to the projected result of the back-mapping scheme. On the right-hand side, an enlarged excerpt is shown.

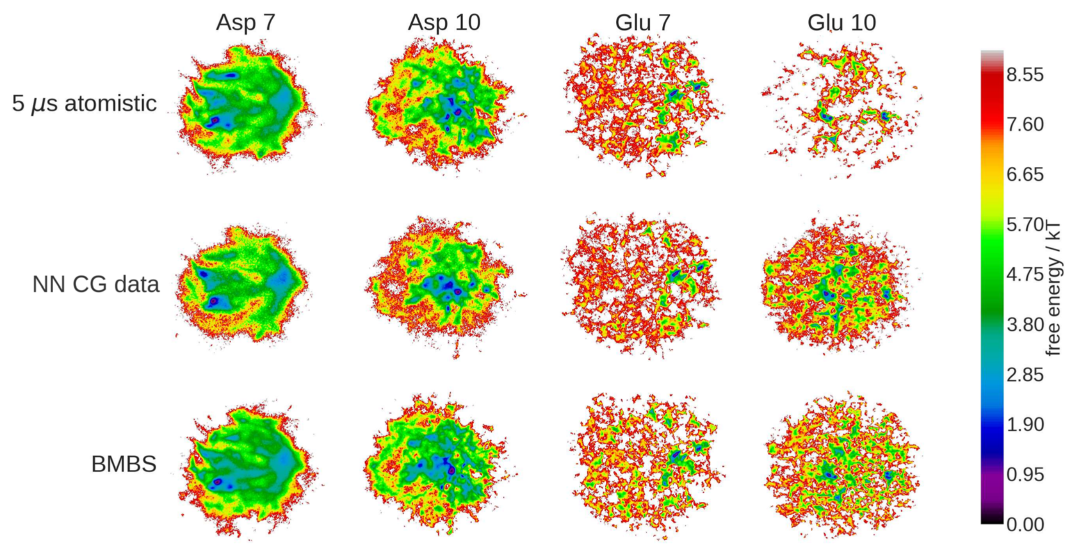

FIG. 6. FESs (generated by Boltzmann inversion, i.e., the negative logarithm of the probability distribution) of the 5 μs atomistic simulations (the first row), 256 000 conformations from the CG MC simulations (the second row), and 800 ns of BMBS (the third row) for the heptamers and decamers in the sketch-map space.

conformations. For both decamers, one sees some differences between the three FESs. For Asp 10, all three landscapes occupy the same area in the sketch-map, but the distributions in the 2D space vary slightly. For Glu 10, however, the differences are more pronounced. Despite the fact that the deepest free energy minima of the 5 μs atomistic simulation seem to coincide very well with the ones of both the CG MC simulation and the BMBS, the overall sampled area as displayed in the sketch-map space is considerably smaller in the case of the atomistic simulation. There are two possible explanations for such a discrepancy, namely, a lack of sampling of the—in principle, accessible—conformational phase space by the atomistic simulation and deficiencies in the CG model that translate into the backmapped ensemble. In order to probe the convergence of the 5 μs atomistic simulation, we continued the simulation for another 5 μs. No previously unvisited areas were sampled. While analyzing this additional trajectory, we found that the system was staying trapped in a helical conformation during the entire length of the simulation. This nicely illustrates both the great stability of the helical state and the amount of time that it can take for a trajectory to be able to leave this free energy minimum. We compare the conformational ensembles of Glu 10 from the 5 μs atomistic simulation and from BMBS in more detail in Sec. III C. In the following, we will first provide a more quantitative comparison between the FESs from atomistic, CG, and BMBS simulations and monitor the behavior of BMBS depending on the amount of sampling coming from short individual simulations. This will allow us to better understand the origin of the discrepancies (limited atomistic sampling vs CG errors) and show how the BMBS scheme allows us to address both issues.

B. Quantitative comparison of FES

In order to give a more quantitative comparison between the FESs produced by the BMBS compared to the CG and atomistic landscapes, we calculated the earth mover's distance between the respective (normalized) probability distributions (2D histograms) in the sketch-map space from which the FESs are derived. Figure 7 shows the normalized EMDs [Eq. (2)] between the three sampling schemes for all Asp and Glu oligomers. For the Asp oligomers, the EMDs confirm the visual observation that the FESs match really well. Especially, for Asp 6 and Asp 7, the landscapes are almost identical with EMDs of 0.02 and lower. However, even for the decamer,

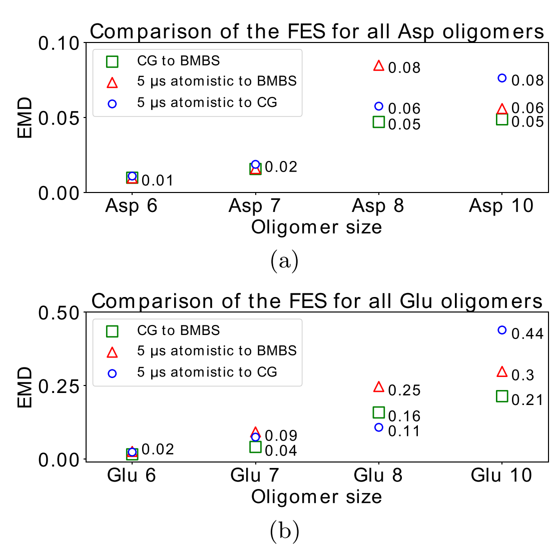

FIG. 7. Earth mover's distances between the FESs of the 5 μs atomistic, the CG MC, and BMBS simulations data for (a) Asp and (b) Glu oligomers.

the EMD shows a good agreement between the landscapes from the BMBS and both the atomistic and the CG simulations with values of 0.06 and 0.05, respectively.

For Glu 6 and Glu 7, the EMDs are again very low, with values of 0.09 or lower, but for Glu 10, all three EMDs are significantly larger. The by far largest EMD is found between the CG landscape and the one from 5 μs atomistic simulation (blue circle), again matching the observation made by visual inspection. In contrast, the EMD between the atomistic simulation and the BMBS (red triangle) is considerably lower and lies in the same range as the EMD between BMBS and CG data (green square). This is remarkable since structures from the CG landscape (with weights corresponding to the FES) had served as the seeding template for the BMBS simulations. These two low EMDs can thus be interpreted in such a way that the atomistic simulations after the backmapping step in the BMBS scheme do, on the one hand, reflect the broader conformational phase space of the CG model (i.e., overcome sampling limitations of the free atomistic simulation) while at the same time correcting for flaws which inevitably exist in the purely predictive CG FES for Glu 10. To further confirm this interpretation, we calculated the time evolution of the EMDs for Glu 10 [Fig. 8(a)] and Asp 10 [Fig. 8(b)] (i.e., as a function of the length of the free atomistic simulation after the restrained simulation pulled the atomistic system to the CG data point). To visually aid in the interpretation of the EMD data, this time evolution is also illustrated in terms of the expansion in the 2D sketch-map space in Fig. 9 for Glu 10 and in Fig. SI 3 for Asp 10.

Figure 8 shows that—not unexpectedly—the BMBS FESs become more and more similar to the atomistic reference with increasing simulation time (blue circles). On the other hand, a comparison of the landscapes from the BMBS and the NN CG dataset (red triangles) shows that the free energy landscape produced by the enhanced sampling scheme first approaches the NN predicted template but then slightly moves away. This suggests that—after

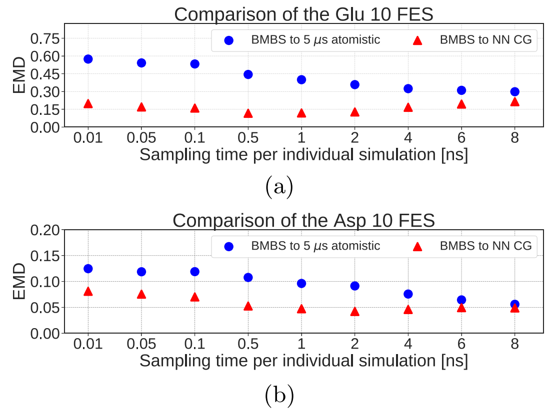

FIG. 8. EMDs between the 2D histograms in the sketch-map space obtained from the BMBS scheme and the 5 μs atomistic simulation (blue spheres) or the NN CG data (red triangles), respectively. The EMDs have been computed as a function of the length of the sampling time in the BMBS scheme (i.e., the length of the unrestrained atomistic sampling after the restrained simulation pulled the atomistic system to the CG data point). The data for Glu 10 are shown in (a), and the data for Asp 10 are shown in (b).

initially simply filling up the sampling of the landscape that had been spanned by the 100 reference data points drawn from the CG data—the free atomistic simulations of the BMBS scheme correct for the inevitable flaws of the landscapes produced by the CG model. Such a correction for flaws in the CG model is a remarkable and valuable observation, since it shows that the BMBS scheme allows us to assess the validity of the CG simulations and that its performance in terms of correctly sampling phase space is not strictly limited by the quality of the CG model. Note that the effect is much larger for Glu 10 [panel (a)] compared to Asp 10 [panel (b)], which is not unexpected

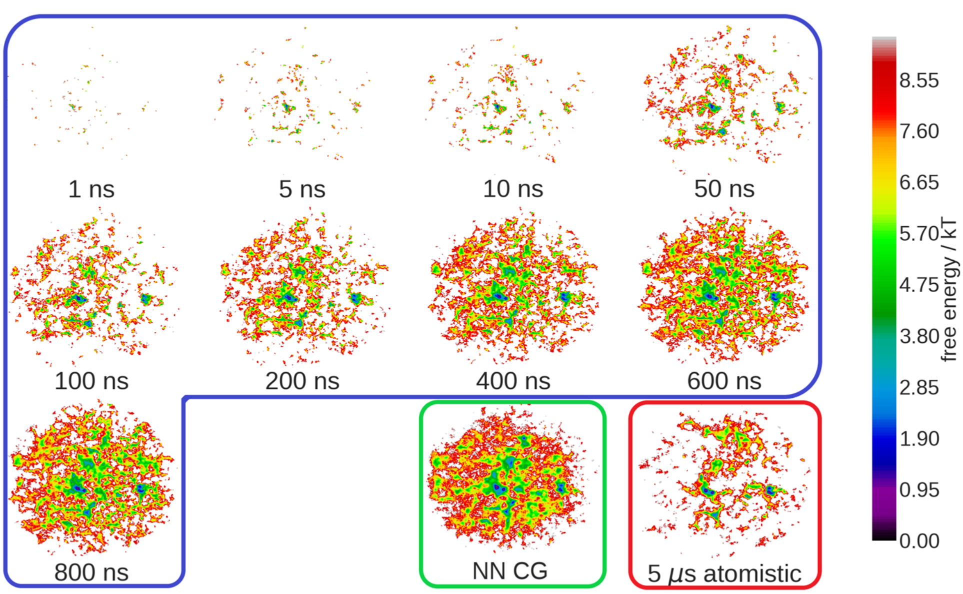

FIG. 9. Time evolution of the sampling of the conformational phase space of Glu 10 with the BMBS scheme is shown by the Boltzmann inverted histograms within the blue frame, with the number below the FES indicating the total simulation time accumulated by the 100 individual simulations. For comparison, the projections within the green and red frames show the NN CG and the 5 μs atomistic landscapes, respectively.

since for Asp 10 the agreement between CG and atomistic FES is already very good (see Fig. 7).

For an 800 ns expansion simulation of Glu 7 (see Fig. SI 4), on the other hand, this trend is not observable because Glu 7 was part of the training data for the neural network and therefore the FES was not purely predicted, like in the case of Asp 10 and Glu 10. Hence, the FES produced by the NN is very accurate and there is no need for correction. That is why for Glu 7, the EMDs for the comparison of the expansion scheme with the NN CG data gradually decrease with increasing simulation time.

Summarizing, we can conclude that for all oligomers, which were used for the training of the NN, the FESs of the 5 μs atomistic simulation and the NN CG data are very similar, which is confirmed by a visual comparison of the 2D landscapes as well as by the calculated EMDs. This is also the case for Asp 10, which was not used for training, and therefore, the NN purely predicted the side-chain distance distributions. Remarkably, for Glu 10, the EMD between the CG and the atomistic landscape is much larger than for every other oligomer. From the BMBS-sampling time-dependent comparison with both reference datasets, we conclude that both the reference 5 μs atomistic Glu 10 simulation is not converged and the CG model samples the FES somewhat inaccurately. The BMBS scheme seems to be able to compensate for both drawbacks. In Sec. III C, we will more closely investigate the atomistic ensemble of conformations obtained from this scheme.

Motivated by the observation that the BMBS may be able to compensate for flaws of the underlying CG model, we have applied the scheme to two further sets of CG simulation data for Glu 10. These were obtained from CG models which were intentionally chosen because they are flawed: first, we used a NN CG model that was trained only on atomistic reference data for a heptamer. Because the NN model was trained on only one oligomer instead on a series of oligomers (Glu 6–8), this model is less well transferable to Glu 10 and gives a representation of the Glu 10 conformational FES. Additionally, we used the MARTINI CG force field[51,52]—with the secondary structure assigned to coil formation, thus not representing the strong tendency of Glu 10 to form α helices and missing relevant conformational states of the peptide. We found that in both cases, BMBS was indeed able to improve the sampling of the conformational phase space of the CG models. Naturally, if the models have major flaws, BMBS requires more simulation time for the individual simulations, as well as a different strategy to check the convergence. More information about these test cases is provided in the supplementary material.

C. Clustering of backbone conformations

So far, we compared the FESs only with sketch-map projections based on the pairwise distances between side-chains as CVs, as this is the only information available from the CG representation. To verify that the new parts of the phase space explored by the BMBS scheme actually correspond to new backbone conformations, we reinvestigated the atomistic ensembles of Glu 10 with the help of a new set of CVs, namely, the 18 backbone dihedral angles (ϕ and ψ angles). We computed the sketch-map projections based on the dihedral angles of the 800 ns expansions from the BMBS scheme. The landmarks of this projection were then subsequently used to project the dihedral angles of the 5 μs atomistic simulation to the same space

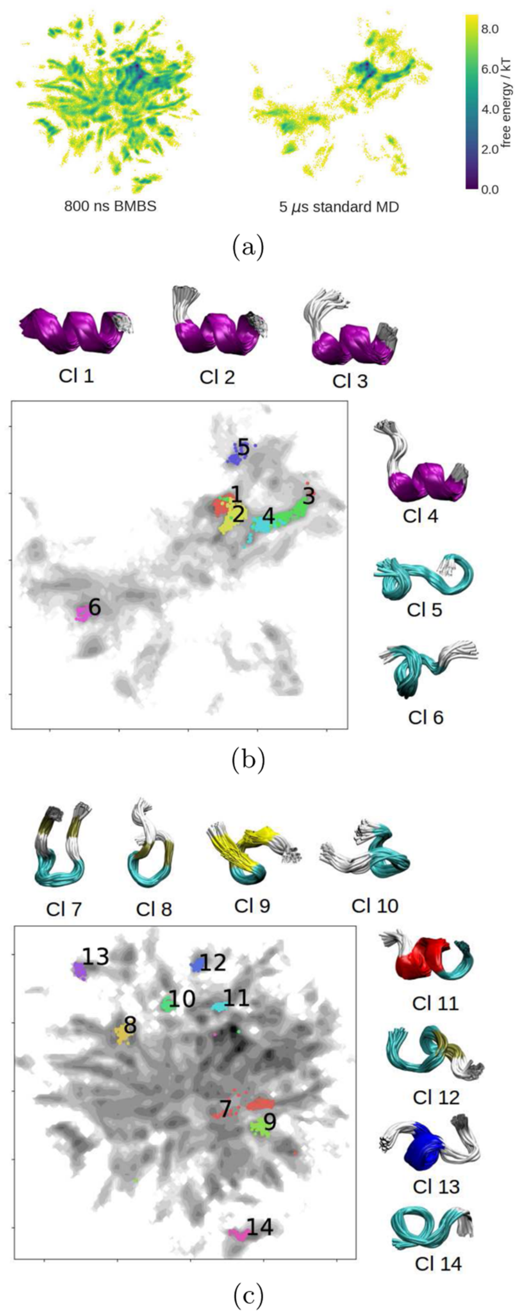

FIG. 10. Projections of the conformations visited by Glu 10 to a sketch-map space derived from the backbone dihedral angles. (a) FES (i.e., Boltzmann inverted probability densities) of the 800 ns BMBS data (left panel) and the data from the 5 μs atomistic simulations (right panel). (b) Projection of the data points from the 5 μs atomistic simulations colored according to the membership clusters obtained from RMSD-based clustering (only some selected clusters and representative structure bundles are shown). (c) Projection of the data points from the 800 ns BMBS scheme; colored points correspond to the selected clusters found in RSMD-based clustering, which were sampled by BMBS and are not present in the 5 μs atomistic simulation (representative structure bundles of these clusters are indicated).

[see Fig. 10(a)]. The projections show that the trajectories of the BMBS explore many areas in the dihedral space that had not been visited by the 5 μs atomistic simulation.

In order to guarantee that these newly sampled areas in the FES indeed correspond to new backbone conformations, we used a clustering algorithm to analyze the conformational states visited in the simulations.[53] To verify that the high probability density regions in the sketch-maps represent distinct conformational states, we used a clustering method based on root-mean-square-deviations (RMSDs) between pairs of conformations, which is independent from the sketch map projections. The clustering was done separately for the BMBS and the free atomistic trajectory, but for both clusterings, we used a RMSD cutoff of 0.2 nm. After obtaining the characteristic clusters, we colored the respective data points in the projections accordingly. In Fig. 10(b), the sketch-map projection of the 5 μs atomistic simulation is colored according to the membership in the resulting 6 largest clusters. In contrast, Fig. 10(c) shows various additional clusters in the sketch-map space of BMBS not visited by the 5 μs atomistic simulation. All of the clusters that are shown in Fig. 10(b) were also sampled by the BMBS; however, they are not shown again in Fig. 10(c). A more complete account of where the 60 largest clusters obtained from RMSD-based clustering are found in the sketch-map space is presented in Fig. SI 7.

By comparing Figs. 10(b) and 10(c), it is easy to see that the newly sampled areas in the dihedral angle sketch-map space indeed also correspond to newly sampled backbone conformations. This also illustrates very well the power of BMBS, where the combination of 100 simulations of 8 ns each manages to sample much more of the accessible conformational space than a 5 μs simulation. As BMBS relies on the weights of CG sampling, we performed temperature replica exchange (REMD) simulations[54] for Glu 10 to better assess how probable the newly found regions are in the atomistic model (see Fig. SI 8 of the supplementary material for projections of the lowest replica and all replicas). In general, BMBS explores almost the same phase space as 350 ns of 40 replicas, but does so with significantly smaller computational effort (800 ns compared to 14 μs). However, the exact statistical weights of many new regions sampled with BMBS differ from those obtained from REMD and some regions of BMBS are not present in the lowest replica. (Note that we did not apply reweighting techniques to the REMD ensemble.) These results show that BMBS does indeed allow us to overcome sampling limitations of the atomistic simulations and "sees" most of the structures sampled by REMD with much less computational effort. However, the quality of the obtained landscape and thus its interpretability in terms of free energies does rely on the correctness of the CG weights with which the BMBS is "seeded," as well as the number and length of the atomistic runs after back-mapping—especially if the transitions in the atomistic simulations are slow.

IV. CONCLUSION

We tackled the sampling problem of classical molecular dynamics simulations by introducing a new back-mapping based sampling approach, which combines the advantages of three resolution levels (atomistic, coarse grained, and low-dimensional projection). On the coarse grained level, a fast exploration of wide areas of phase space takes place. BMBS then allows us to reintroduce the full atomistic resolution to the CG ensemble. This is possible even if the coarse graining is quite drastic and, for example, only consists of one bead per amino acid. An important advantage of BMBS is that it is not strictly bound to the quality of an underlying CG sampling. Instead, it can compensate for possible inaccuracies in the CG simulation. The behavior of BMBS with incorrect CG models is an interesting aspect for future research. If no deviations between the CG and atomistic ensemble are observed, this is a good verification for the validity of the CG model. Such deviations can easily be monitored in the low-dimensional projections with the help of the earth mover's distance as a metric to allow for a quantitative comparison of the resulting free energy surfaces. To guide the atomistic sampling, BMBS uses probability distributions of 2D projections of a coarse grained ensemble. This guidance leads to an efficiency benefit of the proposed scheme compared to brute-force atomistic MD simulations, which increases with the size of the system. Already for a heptapeptide, a sampling is about one order of magnitude faster. For longer peptides like Glu-10 where convergence is out of reach even with 10 μs of simulation time, BMBS can show its full potential. Only 800 ns of simulations with BMBS resulted in sampling of a significantly larger portion of the phase space compared to 10 μs of classical MD simulations.

SUPPLEMENTARY MATERIAL

See the supplementary material for additional plots, sketch-map parameters, and software[55] used for extraction of CVs and plotting.

ACKNOWLEDGMENTS

This work was supported by the DFG through CRC 969, the Zukunftskolleg of the University of Konstanz, and the Carl Zeiss Foundation. Computational resources of the bwHPC project were used (DFG Grant No. INST 35/1134-1 FUGG and the state of Baden-Württemberg).

REFERENCES

1. M. Karplus and J. A. McCammon, Nat. Struct. Biol. **9**, 646 (2002).
2. R. O. Dror, R. M. Dirks, J. Grossman, H. Xu, and D. E. Shaw, Annu. Rev. Biophys. **41**, 429 (2012).
3. J. R. Perilla, B. C. Goh, C. K. Cassidy, B. Liu, R. C. Bernardi, T. Rudack, H. Yu, Z. Wu, and K. Schulten, Curr. Opin. Struct. Biol. **31**, 64 (2015).
4. R. C. Bernardi, M. C. Melo, and K. Schulten, Biochim. Biophys. Acta, Gen. Subj. **1850**, 872 (2015), special issue on Recent Developments of Molecular Dynamics.
5. G. Torrie and J. Valleau, J. Comput. Phys. **23**, 187 (1977).
6. A. Barducci, M. Bonomi, and M. Parrinello, Wiley Interdiscip. Rev.: Comput. Mol. Sci. **1**, 826 (2011).
7. O. Valsson and M. Parrinello, "Variationally enhanced sampling," in *Handbook of Materials Modeling: Methods, Theory and Modeling*, edited by W. Andreoni and S. Yip (Springer International Publishing, Cham, 2018), pp. 1–14.
8. O. Kukharenko, K. Sawade, J. Steuer, and C. Peter, J. Chem. Theory Comput. **12**, 4726 (2016).
9. E. Chiavazzo, R. Covino, R. R. Coifman, C. W. Gear, A. S. Georgiou, G. Hummer, and I. G. Kevrekidis, Proc. Natl. Acad. Sci. U. S. A. **114**, E5494 (2017).
10. S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries, J. Phys. Chem. B **111**, 7812 (2007).
11. R. Potestio, C. Peter, and K. Kremer, Entropy **16**, 4199 (2014).
12. W. G. Noid, J. Chem. Phys. **139**, 090901 (2013).

[13] E. Brini, E. A. Algaer, P. Ganguly, C. Li, F. Rodríguez-Ropero, and N. F. A. van der Vegt, Soft Matter **9**, 2108 (2013).
[14] R. Menichetti, K. Kremer, and T. Bereau, Biochem. Biophys. Res. Commun. **498**, 282 (2018), special issue on Multiscale Modeling.
[15] A. Berg, O. Kukharenko, M. Scheffner, and C. Peter, PLoS Comput. Biol. **14**, e1006589 (2018).
[16] P. Kar and M. Feig, J. Chem. Theory Comput. **13**, 5753 (2017).
[17] M. Heidari, K. Kremer, R. Cortes-Huerto, and R. Potestio, J. Chem. Theory Comput. **14**, 3409 (2018).
[18] E. Lyman, F. M. Ytreberg, and D. M. Zuckerman, Phys. Rev. Lett. **96**, 028105 (2006).
[19] Y. Chen and B. Roux, J. Chem. Theory Comput. **11**, 3572 (2015).
[20] K. Moritsugu, T. Terada, H. Kokubo, S. Endo, T. Tanaka, and A. Kidera, J. Chem. Phys. **149**, 072314 (2018).
[21] N. E. Jackson, M. A. Webb, and J. J. de Pablo, Curr. Opin. Chem. Eng. **23**, 106 (2019), special issue on Frontiers of Chemical Engineering: Molecular Modeling.
[22] M. Haghighatlari and J. Hachmann, Curr. Opin. Chem. Eng. **23**, 51 (2019).
[23] T. Lemke and C. Peter, J. Chem. Theory Comput. **13**, 6213 (2017).
[24] T. Bereau, C. Globisch, M. Deserno, and C. Peter, J. Chem. Theory Comput. **8**, 3750 (2012).
[25] A. J. Rzepiela, L. V. Schäfer, N. Goga, H. J. Risselada, A. H. De Vries, and S. J. Marrink, J. Comput. Chem. **31**, 1333 (2010).
[26] A. Villa, C. Peter, and N. F. A. van der Vegt, Phys. Chem. Chem. Phys. **11**, 2077 (2009).
[27] H. Berendsen, D. van der Spoel, and R. van Drunen, Comput. Phys. Commun. **91**, 43 (1995).
[28] N. Schmid, A. P. Eichenberger, A. Choutko, S. Riniker, M. Winger, A. E. Mark, and W. F. van Gunsteren, Eur. Biophys. J. **40**, 843 (2011).
[29] B. Pullman, *Intermolecular Forces in Proceedings of the Fourteenth Jerusalem Symposium on Quantum Chemistry and Biochemistry Held in Jerusalem, Israel, April 13-16, 1981*, Jerusalem Symposium on Quantum Chemistry and Biochemistry (D. Reidel, 1981).
[30] G. Bussi, D. Donadio, and M. Parrinello, J. Chem. Phys. **126**, 014101 (2007).
[31] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, J. Chem. Phys. **81**, 3684 (1984).
[32] R. Hockney, S. Goel, and J. Eastwood, J. Comput. Phys. **14**, 148 (1974).
[33] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, J. Chem. Phys. **103**, 8577 (1995).
[34] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, J. Comput. Chem. **18**, 1463 (1997).
[35] K. Pearson F.R.S., London, Edinburgh, Dublin Philos. Mag. J. Sci. **2**, 559 (1901).
[36] H. Hotelling, J. Educ. Psychol. **24**, 417 (1933).
[37] L. Molgedey and H. G. Schuster, Phys. Rev. Lett. **72**, 3634 (1994).
[38] A. Hyvärinen, J. Karhunen, and E. Oja, "Methods using time structure," in *Independent Component Analysis* (John Wiley & Sons, 2002), Chap. 18, p. 344.
[39] G. Pérez-Hernández, F. Paul, T. Giorgino, G. D. Fabritiis, and F. Noé, J. Chem. Phys. **139**, 015102 (2013).
[40] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, Proc. Natl. Acad. Sci. U. S. A. **102**, 7426 (2005).
[41] R. R. Coifman and S. Lafon, Appl. Comput. Harmon. Anal. **21**, 5 (2006).
[42] M. Ceriotti, G. A. Tribello, and M. Parrinello, Proc. Natl. Acad. Sci. U. S. A. **108**, 13023 (2011).
[43] W. Chen, A. R. Tan, and A. L. Ferguson, J. Chem. Phys. **149**, 072312 (2018).
[44] T. Lemke and C. Peter, J. Chem. Theory Comput. **15**, 1209 (2019).
[45] A. L. Gibbs and F. E. Su, Int. Stat. Rev. **70**, 419 (2002).
[46] Y. Rubner, C. Tomasi, and L. J. Guibas, in *Sixth International Conference on Computer Vision*, IEEE Cat. No. 98CH36271 (IEEE, 1998), pp. 59–66.
[47] Y. Rubner, C. Tomasi, and L. J. Guibas, Int. J. Comput. Vision **40**, 99 (2000).
[48] O. Pele and M. Werman, *Computer Vision–ECCV 2008* (Springer, 2008), pp. 495–508.
[49] O. Pele and M. Werman, in *2009 IEEE 12th International Conference on Computer Vision* (IEEE, 2009), pp. 460–467.
[50] D. Applegate, T. Dasu, S. Krishnan, and S. Urbanek, in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '11* (ACM, New York, NY, USA, 2011), pp. 636–644.
[51] L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S.-J. Marrink, J. Chem. Theory Comput. **4**, 819 (2008).
[52] D. H. de Jong, G. Singh, W. F. D. Bennett, C. Arnarez, T. A. Wassenaar, L. V. Schäfer, X. Periole, D. P. Tieleman, and S. J. Marrink, J. Chem. Theory Comput. **9**, 687 (2013).
[53] X. Daura, K. Gademann, B. Jaun, D. Seebach, W. Gunsteren, and A. Mark, Angew. Chem., Int. Ed. **38**, 236 (1999).
[54] Y. Sugita and Y. Okamoto, Chem. Phys. Lett. **314**, 141 (1999).
[55] M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz, and F. Noé, J. Chem. Theory Comput. **11**, 5525 (2015).