ORIGINAL PAPER

ZIBgridfree: efficient conformational analysis by partition-of-unity coupling

Alexander Bujotzek · Ole Schütt · Adam Nielsen · Konstantin Fackeldey · Marcus Weber

Abstract Obtaining a sufficient sampling of conformational space is a common problem in molecular simulation. We present the implementation of an umbrella-like adaptive sampling approach based on function-based meshless discretization of conformational space that is compatible with state of the art molecular dynamics code and that integrates an eigenvector-based clustering approach for conformational analysis and the computation of inter-conformational transition rates. The approach is applied to three example systems, namely n-pentane, alanine dipeptide, and a small synthetic host-guest system, the latter two including explicitly modeled solvent.

Keywords Conformational analysis · Meshfree methods · Molecular simulation

1 Introduction

The dynamics of molecular systems exhibits a distinct metastable character: Molecular systems tend to remain within an almost invariant subset of conformational space for a long time—long in relation to the step size of the numerical integration, which for atomistic simulations is in the order of one or two femtoseconds—while transitions between different almost invariant subsets (i.e. conformational changes) are rarely observed events. This characteristic is due to the rough potential energy landscape inherent to most molecular systems. Basins of low potential energy, grouped around local minima, are separated by high energy barriers, corresponding to conformational

A. Bujotzek · O. Schütt · A. Nielsen · K. Fackeldey (✉) · M. Weber
Computational Molecular Design Group, Zuse Institute Berlin, Berlin, Germany
e-mail: fackeldey@zib.de

A. Bujotzek
e-mail: alexander.bujotzek@gmail.com

M. Weber
e-mail: weber@zib.de

changes , or changes from unbound to bound state. This complicates the sampling of conformational space, as molecular dynamics (MD) trajectories tend to generate states from within the basin of one local minimum for a long time, while transitions between different local minima are achieved only very seldom, or not at all. This effect, often denoted as *trapping*, can lead to incomplete coverage of conformational space, and thus to insufficient statistics. It is particularly severe with regard to the sampling of transient regions of conformational space, e.g. in the study of ligand-receptor binding processes, as the dynamics of the system will try to avoid the energetically unfavorable (but most interesting) transition states.

While, as of yet, thermostated long-time MD remains the predominant tool in the molecular simulation community, several successful strategies for overcoming (or rather lessening) the sampling problem have been developed, including umbrella sampling [1], essential dynamics [2] and replica exchange [3]. An excellent implementation of various enhanced sampling schemes is available in terms of the PLUMED plug-in [4] that is compatible with various popular MD packages.

In this article, we present an enhanced version of the ZIBgridfree sampling algorithm [5], which is inspired by the umbrella sampling approach. ZIBgridfree uses an adaptive refinement strategy in order to enable efficient and thorough sampling even in transient regions of conformational space. The main feature of ZIBgridfree as presented here is that it combines an efficient importance sampling scheme with a comprehensive and visual framework for conformational analysis w.r.t. both single molecules and binding processes.

In the initial step of the algorithm, conformational space is partitioned into subsets. Each subset is sampled independently toward convergence of the correct local distribution. More precisely, instead of computing only one trajectory for exploring the potential energy landscape, we compute short trajectories which are confined to a subset of the conformation space by restraints. These subsets then are defined by a partition of unity on the conformation space. If convergence fails (e.g. when the sampling keeps on "jumping" between two local minima), a refinement of the partitioning is triggered, followed by additional sampling. In the subsequent step, each local sampling will be weighted such that the overall histogram yields the *global* Boltzmann distribution, so that the identification of conformations is reduced to a clustering problem based on the eigenstates of the overlap matrix of the partitioning. Finally, conformational weights and inter-conformational transition probabilities can be determined. The extended version of ZIBgridfree presented here broadens the scope of this sampling scheme by combining it with a standard MD software package so as to give access to the most up-to-date molecular force fields and solvent models.

2 Theory and implementation

2.1 Conformation dynamics

As partitioning methods based on meshes or grids suffer from the "curse of dimensionality", ZIBgridfree implements a meshless, function-based partitioning approach. This is motivated by the concept of conformation dynamics [6,7], where conforma-

tions of a molecular system are defined in terms of soft-characteristic membership functions, rather than classical sets in position space (below denoted as Ω). We are interested in a soft partitioning of the position space, i.e. we want to have a set of functions $\chi_1, \ldots, \chi_{n_c} : \Omega \to [0, 1]$ such that

$$\sum_{i=1}^{n_C} \chi_i(q) = 1, \tag{1}$$

holds for all $q \in \Omega$. One can regard χ_i as a probability distribution. For a set of position states we say that they are distributed according χ_i when for each collection of conformations A we find $\int_A \frac{\chi_i(q)}{\tilde{w}_i} \rho(q)\,dq$ percent of position states from the set in a conformation from A, with the corresponding thermodynamical weights

$$\tilde{w}_i := \int_\Omega \chi_i(q)\,\rho(q)\,dq. \tag{2}$$

This means the position states are distributed according to the partial density function $\tilde{\rho}_i$:

$$\tilde{\rho}_i = \frac{\chi_i(q)\rho(q)}{\tilde{w}_i}. \tag{3}$$

Note that for the special case $\chi_1, \ldots, \chi_{n_c} : \Omega \to \{0, 1\}$ our approach reduces to the well known Markov State Model [8–10]. In this case $\tilde{w}_i$ is the probability to be in set $A_i := \{q \in \Omega \mid \chi_i(q) = 1\}$ and the transition matrix T for some fixed time step τ is defined such that T_{ij} denotes the probability to move from set A_i to set A_j in time τ. In general $\tilde{w}_i$ denotes the probability that the molecule will be found in the conformation represented by χ_i and the transition matrix T for some time step τ is given in the following way: If we have a set of position states distributed according χ_i then after a time step τ they will be distributed according $\sum_{k=1}^n \chi_k T_{ik}$. One new property of T is that the entries do not need to be positive. A partition into metastable conformation is given if we find a soft partitioning such that each distribution χ_i represents a metastable conformation, i.e. $T_{ii} \approx 1$ for $i = 1, \ldots, n_c$. In the following we show how one can obtain such a soft partitioning in metastable conformations and conclude with three examples where we have approximated $\tilde{w}_i$ for each. For one example we have also approximated the transitions matrix.

To find $\chi_1, \ldots, \chi_{n_C}$ we start off with a function basis $\phi_1, \ldots, \phi_s : \Omega \to [0, 1]$, where the initial number of basis functions s should be chosen larger than the anticipated number of conformations n_c. The function basis is chosen such that is has the same properties as the membership functions $\chi_1, \ldots, \chi_{n_C}$, i.e. partition of unity (cp. Eq. 1). Therefore, each conformation membership function χ_j can be constructed from a convex combination of the basis functions ϕ_i [11]:

$$\chi_j = \sum_{i=1}^s \chi_{disc}(i, j)\phi_i, \quad j = 1, \ldots, n_C, \tag{4}$$

where χ_{disc} is a row-stochastic matrix containing the linear combination factors. Analogous to $\tilde{\rho}_i$ and $\tilde{w}_i$ in Eqs. 3 and 2, each of the basis function is associated with a partial density ρ_i and a thermodynamic weight w_i. In order to calculate a set of points distributed according ϕ_i one can simulate a trajectory according to the modified potential energy function $\tilde{U}_i$ as [11]

$$\tilde{U}_i(q) = U(q) + \hat{U}_i(q) = U(q) - \frac{1}{\beta} \ln (\phi_i(q)). \tag{5}$$

This fact will come in handy for calculating the corresponding w_i and the subsequent cluster analysis which aims at identifying both the correct number of clusters n_C, as well as the matrix χ_{disc} of linear combination factors, from which we obtain the set of membership functions χ_j by applying Eq. 4.

As a precondition for the partitioning discussed above, a rough scheme of the relevant position space has to be given. This can be delivered in terms of a long-time MD trajectory (possibly using elevated temperature for improved coverage of position space), a targeted MD or pulling trajectory, the output of certain tools for exploring conformational space (e.g. CONCOORD [12] for protein structures) or even by manually preparing a sequence of geometries. From this *presampling* is selected a set of nodes $\{n_1, \ldots, n_s\} \in \Omega$ to each of which is attached a radial basis function W_i given by

$$W_i(q) = \exp(-\alpha \, \delta^2(q, n_i)), \; i = 1, \ldots, s, \tag{6}$$

where α is a shape parameter, and δ^2 a distance measure to be specified in the next section. As the basis functions W_i do not satisfy Equation 1, we construct a partition of unity with basis functions ϕ_i by following Shepard's approach [13]:

$$\phi_i := \frac{W_i}{\sum\limits_{j=1}^{s} W_j}, \; i = 1, \ldots, s. \tag{7}$$

The basis functions ϕ_i take on their maximum at the defining node n_i, and decrease exponentially as the distance δ^2 of a state q to n_i increases. As a consequence, the difference between $\tilde{U}_i$ (Eq. 5) and U is minimal within the state n_i, and increases exponentially with the distance to n_i. This ensures thorough sampling in the area belonging to basis function ϕ_i, as the sampling process is restrained from wandering off into a lower energy basin. The shape parameter α is chosen in dependence on the number of nodes s and the mean node distance θ, and defines the degree of separation of the meshless discretization. For $\alpha \to \infty$, the discretization converges to a Voronoi tessellation, i.e. the soft partitioning degenerates into a hard partitioning without overlaps between the basis functions.

In practice, the sampling of the basis functions ϕ_i is run in parallel, as each $\tilde{U}_i$ can be evaluated at every position $q \in \Omega$ independently of all $\tilde{U}_j$ with $j \neq i$. Depending on the available resources, one can either sample several basis functions in parallel,

evaluate the potential $\tilde{U}_i$ in parallel (which in turn accelerates the sampling of the associated basis function), or combine both approaches.

2.2 Internal coordinates

ZIBgridfree uses internal coordinates (either torsion angles and/or distances) as collective variables in order to define the conformation of the system under observation. Prior to picking a set of nodes for discretization, a set of n_K internal coordinates has to be specified by the user. The distance $\delta^2(q, n_i)$ between state q and node n_i (Eq. 6) is measured in the space of internal coordinates. Therefore, the outcome of the discretization is directly related to the choice of internal coordinates. Deciding on a meaningful set of internal coordinates is not always trivial. For conformational analysis of small molecules, picking all rotatable torsion angles is an obvious choice, whereas for peptides or proteins, picking only backbone torsion angles is practical. For complexes of multiple molecules, the set of torsion angles has to be complemented by a set of distances in order to describe the molecules' relative positioning to each other.

Whereas angular internal coordinates can only take on values between $-\pi$ and $+\pi$, distance (or linear) coordinates can in principle take on any positive value. This leads to problems whenever linear coordinates with a large spread or a large absolute value are overly dominant, as other internal coordinates with more subtle changes are rendered irrelevant when the distance function δ^2 is evaluated. In order to tackle this problem, linear coordinates can be weighted and normalized automatically by calling zgf_create_pool with option '–balance-linears'.

Let k be a linear coordinate that corresponds to the Euclidean distance between two particles in the system under observation. The weight of this coordinate is then determined as follows:

$$\text{coord\_weight}(k) = \frac{\text{coord\_weight}(k)_{initial}}{\sqrt{2 * var(k)}}, \tag{8}$$

where $\text{coord\_weight}(k)_{initial}$ is one, unless specified differently by the user. This means that coordinates with a high spread are downgraded by dividing the initial weight by the full width at half maximum. Furthermore, an offset for k is applied by subtracting its mean value in order to compensate for high absolute values. This leads to the following weighting formula:

$$\begin{aligned} k_{balanced} &= \text{coord\_weight}(k) \cdot (k - \text{offset}(k)) \\ &= \text{coord\_weight}(k) \cdot (k - (\text{offset}(k)_{initial} + mean(k))), \end{aligned} \tag{9}$$

where $\text{offset}(k)_{initial}$ is zero, unless specified differently by the user. This approach realizes an equal weighting of all internal coordinates involved. Nonetheless, certain applications might call for biased weighting of the internal coordinates, e.g. when the distance between ligand and receptor (defined by linear internal coordinates) is to be

stressed in comparison to more subtle conformational changes in the ligand molecule (defined by torsion angle internal coordinates).

2.3 Implementing the potential modification

Sampling the ZIBgridfree basis function ϕ_i requires a modification of the potential function $U(q)$ (Eq. 5). Our aim was to change the algorithm such that it can be run with standard force fields and unmodified molecular dynamics (MD) packages such as GROMACS[14]. Treating the MD code as a black box has several advantages: The user can use readily available software (pre-compiled for many Linux distributions and pre-installed on most computing clusters), and plug in new versions as they are released. Full flexibility regarding the choice of force field and other simulation parameters is sustained. Furthermore, internal changes to the highly optimized MD code, possibly having a negative impact on the simulation performance, are evaded.

Adapting ZIBgridfree to a standard MD package is a two-step procedure. First, for each selected node n_i, the n_K-dimensional ϕ_i function is projected on a single dimension by coordinate-wise evaluation: Instead of considering the joint distance $\delta^2(q, n_i)$ (involving all internal coordinates) we now exclusively consider the distance regarding coordinate k:

$$\phi_{i_k}(q) := \frac{\exp\left(-\alpha\,\delta_k^2(q, n_i)\right)}{\sum_{j=1}^{s}\exp\left(-\alpha\,\delta_k^2(q, n_j)\right)}, \ k = 1, \ldots, n_K. \tag{10}$$

The above expression yields the membership of state q with respect to coordinate k regarding basis function ϕ_i. The one-dimensional penalty potential acting on coordinate k of state q can simply be obtained as:

$$\hat{U}_{i_k}(q) = -\frac{1}{\beta}\ln\left(\phi_{i_k}(q)\right). \tag{11}$$

Finally, in order to approximate $\hat{U}_i$, for every internal coordinate k, a generic cubic restraint potential (as available in many common MD packages) is fitted to the penalty potential $\hat{U}_{i_k}$ and added to the force field representing the unmodified potential U. We implemented this approach for the GROMACS MD package, where restraint potentials of the form

$$U_{res}(\Phi') = \begin{cases} \frac{1}{2}k_{res}\left(\Phi' - \Delta\Phi\right)^2, & \text{for } \Phi' > \Delta\Phi \\ 0, & \text{for } \Phi' \leq \Delta\Phi \end{cases} \tag{12}$$

are readily available (given here for a torsion angle restraint on torsion angle $\Phi' = (\Phi_0 - \Phi) \mod 2\pi$, with rest position Φ_0 and unrestrained region $\Delta\Phi$, analogous for distance restraints). The concept of fitting restraint potentials to the coordinate-wise projected basis function penalty potentials of ZIBgridfree is depicted in Fig. 1.

The imperfect approximation of multi-dimensional basis functions by harmonic restraints introduces a certain error, as sampling points may be generated from areas
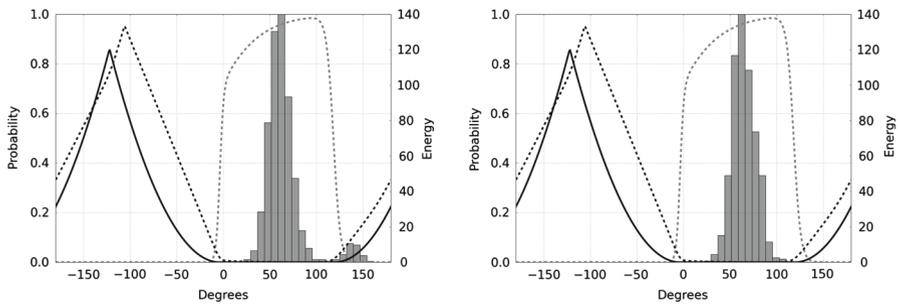
Fig. 1 Sampling of a torsion angle distribution (*gray histogram*) with ZIBgridfree. The sampling is forced to stay within the area of an exemplary basis function (*dashed gray line*) by its penalty potential (*dashed black line*). For use with GROMACS, the penalty potential is approximated by a harmonic restraint potential (*solid black line*). Due to the approximation error, the sampling is not sufficiently limited to the area covered by its basis function (*left*). After reweighting the sampling points with regard to the their basis function (*right*), the approximation error is removed

of Ω that are not covered by the basis function in question. This is especially true for boundary regions, where several basis functions are overlapping. This approximation error can be removed by giving each sampling point q a weight $\texttt{frame\_weight}_i(q)$ with respect to basis function ϕ_i:

$$\texttt{frame\_weight}_i(q) = \frac{\phi_i(q)}{\exp(-\beta \cdot U_{res}(q))}. \tag{13}$$

The effect of reweighting on the sampling distribution is depicted in Fig. 1. Calculating the sampling point weights is inexpensive in terms of computation time. Subsequently, when checking for convergence of the sampling, or when calculating observables of any kind, only the reweighted distribution is considered.

2.4 Adaptive refinement of the partitioning

In order to ascertain a sufficient sampling of the partial densities ρ_i, ZIBgridfree pursues an adaptive refinement approach. After a certain number of simulation steps, convergence of the sampling is tested by evaluating the variance-based Gelman-Rubin convergence criterion [15]. If the convergence test fails, the sampling will be extended by n simulation steps (followed by another convergence test) for a maximum of m times (where n and m are user-defined settings). If convergence has not been achieved after m extensions of the original sampling length, a refinement of the partitioning in the area of the affected basis function is triggered. By default, two children nodes n_{i_1} and n_{i_2} are introduced, whereas the original parent n_i is removed from the partitioning, along with its basis function ϕ_i. This principle is illustrated in Fig. 2.

Removal and addition of nodes have an impact on the overall partitioning, as with the number of nodes s, the mean node distance θ is bound to change. Hence, the shape parameter α (Eq. 6) is recalculated following each refinement step. With proceeding refinement and increasing s, α will become larger, which in turn leads to a higher
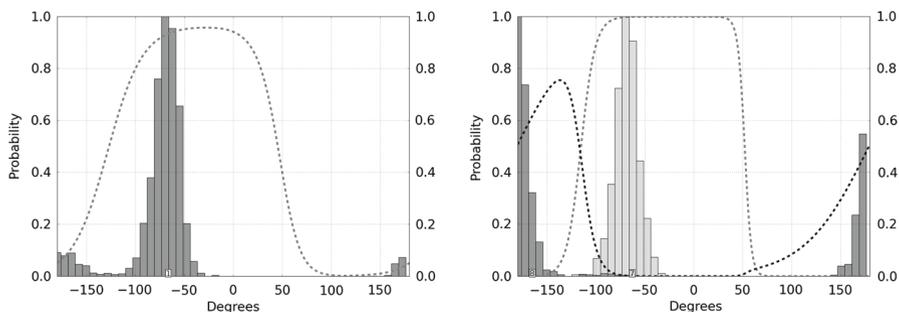
Fig. 2 The sampling of basis function '1' (associated with a node at $-68°$) has come upon a second minimum in the region around $-180°$ (*left*). In this case, convergence of the sampling is not achieved in the allocated number of sampling steps. A failed convergence test triggers an automatic refinement of the partitioning (*right*). The parent node '1' is removed and replaced by two children named '7' ($-65°$) and '8' ($-167°$). The samplings of the associated basis functions converge quickly, as they are now confined to a single energy minimum each

degree of separation between basis functions. This mechanism leads to increased convergence rates over the course of the refinement.

Despite several cycles of refinement, the sampling of transition regions (e.g. when a node is situated on the steep flank of a potential energy barrier) may not lead to convergence according to the Gelman-Rubin criterion. In these cases, the sampling has to be discontinued as soon as a sufficient number of data points from the transition region has been collected.

2.5 Reweighting and cluster analysis

2.5.1 Direct free energy reweighting

The local confined samplings are distributed according to

$$\rho_i = \frac{\phi_i \cdot \rho}{w_i}.$$

If we can calculate the terms $w_1, \ldots, w_s$ we can approximate the correct Boltzmann distribution by weighting the local histogram of ρ_i. The correct weighting is given through

$$\sum_{i=1}^{s} w_i \rho_i = \sum_{i=1}^{s} \phi_i \rho = \rho$$

since the ϕ_i's sum up to one. The partition of unity assures that the passage between the overlapping subsets is described correctly. We remark that this partition of the conformation space is for the purpose of efficiency only and has thus no real physical or chemical meaning. In order to get the "true" global distribution we thus have to account for these local restraints, since otherwise spurious effects might occur which
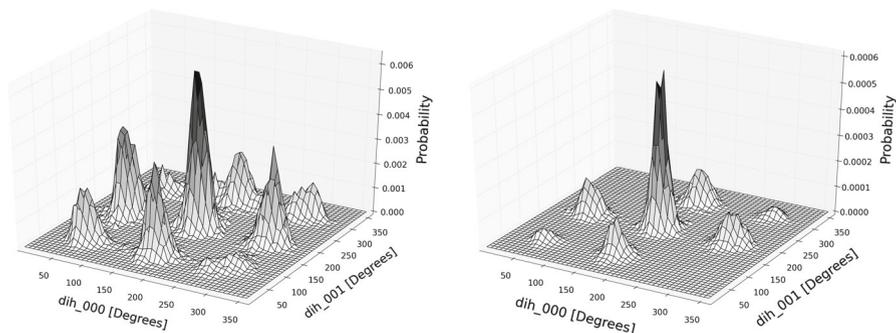
Fig. 3 Torsion angle distribution of the two torsion angles of *n*-pentane at 300 K, assembled from 25 individual node samplings. Before reweighting, each partial density contributes equally to the joint distribution (*left*). This leads to disproportionately high weights of the gauche/trans, trans/gauche and gauche/gauche conformations. After thermodynamic reweighting, the correct relative weights of the partial densities are restored, which leads to an improved joint distribution (*right*)

is illustrated in Fig. 3 for the torsion angle distribution of *n*-pentane. In order to arrive at a balanced joint Boltzmann distribution, we need to find the correct w_i. This is done with the free energy difference estimate implemented in the tool zgf_reweight, based on the approach of Klimm et al. [16]. This approach, which is not dependent on explicit overlap between the partial densities, is outlined shortly in the following. In principle, other methods for thermodynamic reweighting, such as the popular weighted histogram analysis method (WHAM) [17,18], can be employed as well.

1. From each set of states $\{q_n^{(i)}\}_{n=1,...,N^{(i)}} \in \Omega$ representing the partial density $\rho_i, i = 1, \ldots, s$, choose a set of reference points $\{q_r^{(i)}\}_{r=1,...,R^{(i)}}$. A reference point is characterized by having a potential energy value within the energy standard deviation of ρ_i. More precisely, with $\langle U^{(i)} \rangle$ being the mean potential energy of set $q^{(i)}$,

$$\left\| U(q_r^{(i)}) - \langle U^{(i)} \rangle \right\| \le \sqrt{\frac{1}{N^{(i)}} \sum_n^{N^{(i)}} \left(U(q_n^{(i)}) - \langle U^{(i)} \rangle \right)^2}.$$

2. Approximate the local density of sampling points by evaluating expression D_{vol_i}, which counts the number $N_{near}^{(i)}$ of sampling points that are *near*, i.e. within a certain distance vol_i around each reference point $q_r^{(i)}$, and compute its inverse

$$\left(D_{vol_i}(q_r^{(i)}) \right)^{-1} \approx \frac{N^{(i)}}{N_{near}^{(i)} + 1}.$$

For our purpose, vol_i is chosen as large as the mean variance of the internal coordinates regarding all sets of states $q^{(i)}$, which is precomputed in a first iteration over the sampling data. The variance for each set is computed in terms of the distance function δ^2, dependent on the type of the internal coordinates that are involved in the discretization.

3. Compute the entropy estimate

$$S_i = k_B \ln \left(\frac{1}{R^{(i)}} \sum_{l=1}^{R^{(i)}} \left(D_{vol_i}(q_r^{(i)}) \right)^{-1} \right),$$

the free energy

$$G_i = \langle U^{(i)} \rangle - T \cdot S_i,$$

and the statistical weights

$$w_i = w_{i-1} \cdot \exp\left(-\beta\left(G_i - G_{i-1}\right)\right),$$

with $w_1 = 1$. The free energy values have to be ordered by size before calculating the statistical weights. Finally, the statistical weights have to be normalized so that $\sum_{i=1}^{s} w_i = 1$.

2.5.2 Overlap weight correction

The reweighting method introduced in the previous section works best for well-separated basis functions. Depending on the given discretization and the nature of the system under observation, the basis functions in ZIBgridfree can have a more or less pronounced overlap. We perform a correction of the statistical weights w_i in order to take basis function overlap into account. The degree of overlap between each pair of basis functions ϕ_i and ϕ_j is quantified in terms of the overlap integral matrix $S \in \mathbb{R}^{s \times s}$:

$$S_{ij} = \int_{\Omega} \phi_i(q) \rho_j(q) \, dq, \tag{14}$$

which for large numbers is approximated as

$$S_{ij} = \frac{1}{N^{(i)}} \sum_{n=1}^{N^{(i)}} \phi_j(q_n^{(i)}) \cdot \texttt{frame\_weight}_i(q_n^{(i)}) \tag{15}$$

from the states $\{q_n^{(i)}\}_{n=1,\dots,N^{(i)}}$ that represent the partial density ρ_i. Note that the shape of S is influenced by the chosen discretization, in particular by the number of discretization nodes s. For fine discretizations (large α, cp. Eqs. 6 and 7), S will resemble a diagonal matrix. For very coarse discretizations and small α, it will degenerate into a full matrix.

The statistical weights w of the basis functions can be derived by solving the eigenvalue problem $w^\top S = w^\top$, which means that w corresponds to the unique, positive and normalized left eigenvector of S with regard to its eigenvalue $\lambda_1 = 1$ [11]. This eigenvector-based approach is not well-conditioned and highly dependent on

sufficient sampling in the overlap regions between the basis functions [19]. In order to benefit from the advantages of both direct free energy reweighting and the eigenvector-based approach, we start a number of power iteration steps from the original weights w with the stochastic matrix, until the corrected weights (again denoted as w) are convergent.

The row sums of the matrix S do not correspond to the corrected weights w. According to the method of Sinkhorn[20], an iterative rescaling of the row sums to meet w, followed by a symmetrization of S, leads to a corrected overlap integral matrix that is consistent with the precomputed statistical weights.

2.5.3 Metastability analysis with PCCA+

From the chemical perspective, metastable subsets correspond to the main conformations of the underlying molecular system. In the presence of metastable states, any matrix describing the transition behavior of the system (including the matrix S) exhibits a virtual block-diagonal structure, i.e. there exists a permutation of indices so that the metastable subsets of the system are represented by (more or less) quadratic blocks along the diagonal of the matrix (see Fig. 4).

Every block in this matrix is associated with an eigenvector of the matrix whose eigenvalue is almost one. The set of the eigenvalues in the vicinity of one is denoted as the Perron cluster, and the size of this set corresponds to the number of chemical conformations n_C. The linear combinations of the eigenvectors associated with the eigenvalues of the Perron cluster contain, for each basis function ϕ_i, the degree of membership with regard to each of the n_C conformations. Robust Perron cluster analysis (PCCA+) [21,22] is used to find the permutation yielding the block-diagonal structure, and hence the matrix of linear combination factors χ_{disc} (cp. Eq. 4). The result is the matrix $\chi \in \mathbb{R}^{s \times n_C}$, where the entry $\chi(i, j) \in [0, 1]$ denotes the degree of membership of basis function ϕ_i with regard to the jth metastable subset.

Using the weight vector w containing the thermodynamic weights of the basis functions ϕ_i, it is then possible to calculate the weights $\tilde{w}$ of the conformations as $\tilde{w} = \chi^\top w$.

3 Molecular simulation details

All molecular simulations were performed with GROMACS, versions 4.5.4 and 4.5.5 (single precision, unless stated differently). All molecules were parametrized for the

Fig. 4 Schematic of a (permutated) transition matrix in the presence of metastable subsets. Within the three conformations c_1 to c_3, states are mixing quickly. By contrast, transitions from conformation to conformation (*light gray* off-diagonal area) are rare events
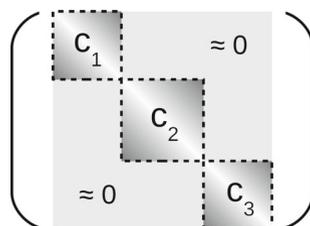
Amber-99SB force field [23]. Residues not already included in the standard force field were prepared using the software ACPYPE [24] and Antechamber [25,26] from AmberTools [27], with charges calculated by the AM1-BCC method [28,29].

For the vacuum simulations (n-pentane), van der Waals and Coulomb interactions were computed without cut-off (all vs. all). For the explicit solvent alanine dipeptide simulations, the TIP4P-Ew water model [30,31] was used. The solute was placed in a rhombic dodecahedron periodic box of 4.0 nm side length. The host-guest system structure in non-complexed form (with the guest molecule displaced by 1.5 nm) was placed in a cubic periodic box of 6.5 nm side length and solvated in a 10:1 mixture of chloroform and methanol. The force field parameters for chloroform and methanol were obtained from the GROMACS Molecule & Liquid Database at URL http://virtualchemistry.org/gmld.php [32,33]. To neutralize the overall charge, a single counter ion was added to the simulation box. In both cases, a twin range cut-off of 1.0/1.4 nm for van der Waals interactions was applied and the smooth particle mesh Ewald algorithm [34] was used for Coulomb interactions, with a switching distance of 1.0 nm.

In order to generate the NVT ensemble of states for the desired temperature of 298/300 K, either the velocity-rescaling thermostat [35] in combination with an MD leap-frog integrator, or a Langevin-type stochastic dynamics [36] integrator was used. For the explicit solvent NpT simulations (alanine dipeptide), the velocity-rescaling thermostat/stochastic dynamics integrator was supplemented by the Parrinello-Rahman barostat [37,38], with a reference pressure of 1 bar. For the host-guest system transition node samplings, neither thermostat nor barostat were applied in order to realize an NVE ensemble setup. The integration step was set to 1 fs for all simulations. The error threshold for the symmetrization of the S matrices was set to 10^{-2} for n-pentane, to 10^{-4} for alanine dipeptide, and to 10^{-3} for the host-guest system.

4 Results and discussion

4.1 Pentane in vacuo

In order to evaluate basic properties of the algorithm, vacuum simulations of n-pentane, a small alkane with five carbon atoms (see Fig. 5), were conducted. The two backbone

Fig. 5 Three-dimensional representation of n-pentane. The two backbone torsion angles chosen as internal coordinates are *highlighted*
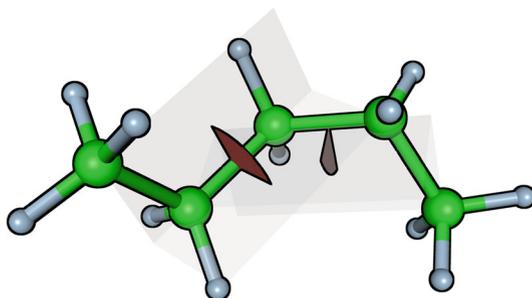
Table 1 Conformational weights of n-pentane at 300 K, derived from a hybrid Monte Carlo (HMC) simulation using the Merck molecular force field [40]

c	tr/tr	g^-/tr	g^+/tr	tr/g^-	tr/g^+	g^+/g^+	g^-/g^-	g^+/g^-	g^-/g^+
$\tilde{w}_c$	0.473	0.120	0.132	0.117	0.132	0.013	0.012	<0.005	<0.005

tr(ans): $\approx \pm 180°$, g(auche)$^+$: $\approx +60°$, g(auche)$^-$: $\approx -60°$. Torsion angles are given on the scale $[-180, \ldots, 180]$

torsion angles of n-pentane were chosen as internal coordinates for the discretization. With regard to these internal coordinates, n-pentane has nine main conformations, separated by distinct energy barriers. The presampling of conformational space was obtained in terms of a 100 ns MD simulation at a very high (and physically unrealistic) temperature of 1,000 K. Reference weights for the conformations of n-pentane were taken from the literature [39] (see Table 1).

4.1.1 Stability regarding randomness of impulse and discretization

In order to monitor the impact of choosing a different discretization (placing of nodes in conformational space) on the sampling outcome, three experiments with ten runs of ZIBgridfree each were conducted: a) Equally placed nodes, but random MD starting impulse, b) randomly placed nodes, but equal MD starting impulse, and c) randomly placed nodes and random MD starting impulse. All runs were conducted with 20 discretization nodes and a minimum sampling time of 100 ps per node, leading to a mean overall sampling time per run of 2.8 ns. The results are shown in Fig. 6, left.

Randomizing the MD starting impulse leads to a maximum standard deviation of 0.025 regarding the weight of the most dominant conformation, tr/tr. Randomizing the node placement by picking different initial seeds for the k-means algorithm leads to a maximum standard deviation of 0.031 for conformation tr/tr. When both MD
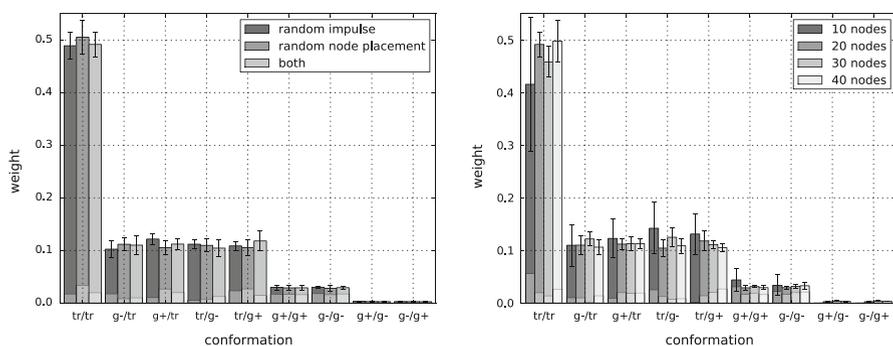
Fig. 6 Conformational weights of n-pentane. *Error bars* indicate the standard deviation w.r.t. 10 runs. Deviation from the literature values is indicated as intra-bar plot. *Left* 20 nodes, 100 ps minimum sampling time per node, with equally placed nodes, random MD starting impulse (*dark gray*), randomly placed nodes, equal MD starting impulse (*gray*), and randomly placed nodes and random MD starting impulse (*light gray*). *Right* 100 ps minimum sampling time per node, 10, 20, 30 and 40 nodes (*dark gray* to *light gray*), random MD starting impulse, random node placement

starting impulse and node placement are randomized at the same time (mimicking a standard sampling setup), the maximum standard deviation is slightly smaller (0.23 for conformation tr/tr), which indicates that the uncertainty regarding both choices is not additive.

4.1.2 Stability regarding fineness of discretization

Similar simulations (random MD starting impulse, random node placement, 100 ps minimum sampling time per node) were performed with varying number of sampling nodes in order to evaluate the impact of the fineness of the discretization. For this experiment, automatic refinement of the discretization was switched off. The results are shown in Fig. 6, right. When only ten discretization nodes are used (only one more than the expected number of conformations), the error becomes very large (0.128 for conformation tr/tr), and, despite a relatively large mean overall sampling time of 3.2 ns per run, the rare conformations g+/g- and g-/g+ are not identified at all. For 20, 30 and 40 discretization nodes (mean overall sampling times 2.79, 4.45 and 5.5 ns per run), the results are comparable, but do not improve visibly with increasing fineness of the discretization.

4.1.3 Stability regarding sampling time

Finally, it was looked into how the sampling time per node determines the quality of the results. The outcome is shown in Fig. 7. A very short minimum sampling time of 10
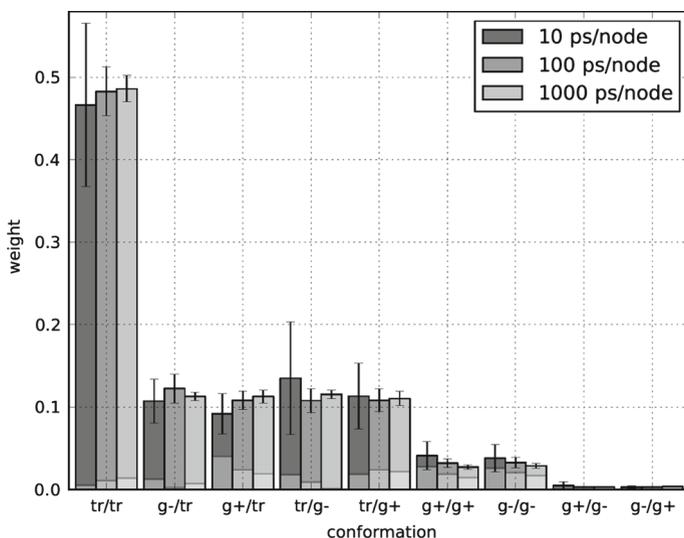
Fig. 7 Conformational weights of *n*-pentane. *Error bars* indicate the standard deviation w.r.t. 10 runs. Deviation from the literature values is indicated as sub-bar plot. 25 nodes, with 10, 100 and 1,000 ps minimum sampling time per node (*dark gray* to *light gray*), random MD starting impulse, random node placement

Table 2 Averaged conformational weights of n-pentane at 300 K, derived from ten runs of ZIBgridfree using the Amber-99SB force field

c	tr/tr	g^-/tr	g^+/tr	tr/g^-	tr/g^+	g^+/g^+	g^-/g^-	g^+/g^-	g^-/g^+
$\tilde{w}_c$	0.486	0.113	0.113	0.116	0.110	0.027	0.029	0.003	0.004

Fig. 8 Three-dimensional representation of alanine dipeptide (ACE-ALA-NME, i.e. terminally blocked alanine). The two backbone torsion angles Φ and Ψ chosen as internal coordinates are *highlighted*
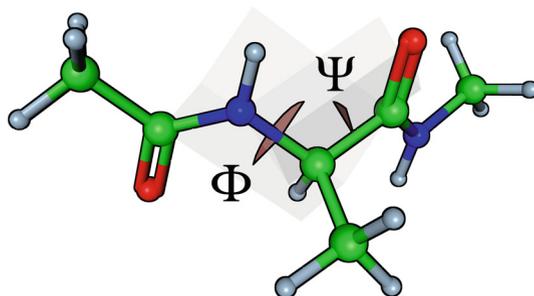
ps per node produces a large error (0.099 for conformation tr/tr), but, given the mean overall sampling time of only 365 ps per run, the averaged conformational weights are acceptable. With increasing sampling time per node, the error can be significantly reduced. For a minimum sampling time of 1,000 ps per node (mean overall sampling time 26.7 ns), the maximum standard deviation (conformation tr/tr) is reduced to 0.016, and below one percent for all other conformations. One can conclude that a rough estimate of the conformational weights can be obtained at a very low cost, whereas precise results have to be paid for with thorough sampling of the partial densities.

The results show a perceivable deviation w.r.t. to the conformational weights found in the literature (cp. Table 1), which most likely can be attributed to the use of a different force field and (possibly) the different dynamics for propagating the system. For comparison, the conformational weights obtained from ZIBgridfree with 25 nodes and 1,000 ps minimum sampling time per node, averaged over ten runs, are given in Table 2.

4.2 Alanine dipeptide in water

As a second example, the conformations of alanine dipeptide in explicit TIP4P-Ew water were studied. Alanine dipeptide is the most basic (or "minimal") polypeptide and serves as a popular test case for evaluating biological force fields. The two backbone torsion angles Φ and Ψ span the relevant conformational space of alanine dipeptide, and were hence chosen as internal coordinates for the discretization. With regard to these internal coordinates, alanine dipeptide has six main conformations, which however are not as well-separated as in the previous example, n-pentane. Obtaining correct conformational weights from explicit solvent simulations is more difficult compared to vacuum or implicit solvent settings, as the dynamics of a solvated system is decelerated, while the computational cost of producing sufficient sampling data multiplies (Fig. 8).

Table 3 Conformational weights of alanine dipeptide at 300 K in the NVT and in the NpT ensemble, derived from two 200 ns MD simulations using the Amber-99SB force field

		C_5	P_{II}	α_R	α_P	α_L	C_7^α
NVT	$\tilde{w}_c$	0.2696	0.4043	0.1745	0.1369	0.0136	0.0010
NpT	$\tilde{w}_c$	0.2794	0.4363	0.1563	0.1190	0.0070	0.0020

$C_5 \approx 143°/ -158°$, $P_{II} \approx 70°/ -158°$, $\alpha_R \approx 70°/11°$, $\alpha_P \approx 136°/ -11°$, $\alpha_L \approx 55°/ -40°$, and $C_7^\alpha \approx -60°/ \pm 180°$. Torsion angles are given on the scale $[-180, \ldots, 180]$. Conformation labels taken from Chodera et al. [41]

Reference weights for the conformations of alanine dipetide at 300 K in the NVT and in the NpT ensemble were obtained from two 200 ns MD simulations (see Table 3).

Explicitly modeled water also complicates the presampling of conformational space: High (or elevated) temperature presampling is possibly only to a certain extent, and requires a re-equilibration of the simulation boxes before the partial densities can be sampled at the target temperature. In principle, discretization nodes can also be picked from a vacuum or implicit solvent trajectory of the molecule of interest, to be put in explicit solvent only before the sampling of partial densities with ZIBgridfree is commenced (implemented in the tools `zgf_solvate_nodes` and `zgf_genion`). Again, another cycle of energy minimization and simulation box equilibration is needed before usable sampling data can be collected. For this example, the presampling consisted of a 100 ns MD trajectory at the target temperature of 300 K, which means that re-equilibration after node selection was not necessary.

4.2.1 Stability regarding sampling time

First, it was looked into how the sampling time per node determines the quality of the results using random MD starting impulse and random node placement in an NVT ensemble. The outcome is shown in Fig. 9. In comparison to the (vacuum) n-pentane example, a longer minimum sampling time per node is required in order to yield acceptable results. For a very short minimum sampling time of 10 ps per node, the results were not interpretable due to the large error (data not shown). A minimum sampling time of 100 ps per node (mean overall sampling time 2.4 ns) produces large errors of around 15 % in terms of standard deviation for the three largest conformations P_{II}, C_5 and α_R. When the minimum sampling time per node is increased to 500 ps (mean overall sampling time 7.7 ns), the error can be reduced below 6 % for all conformations (largest error is 0.0581 for conformation P_{II}). Finally, with a minimum sampling time of 1,000 ps per node (mean overall sampling time 15 ns), the error is in the range of 5 %, and mainly below (largest error is 0.0533 for conformation P_{II}).

An auxiliary trial with a minimum sampling time of 1000 ps per node (mean overall sampling time 15.56 ns) using a double precision version of GROMACS did not lead to a further decrease in standard deviation, contrary to what might have been expected from an increase in precision of coordinates and observables.
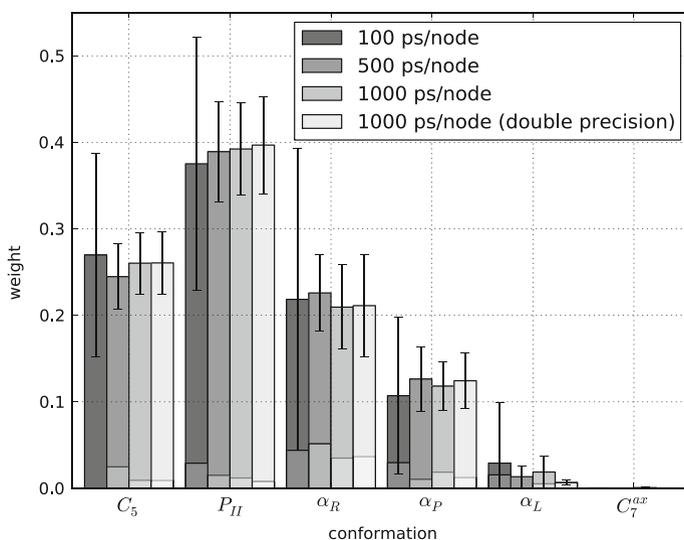
Fig. 9 Conformational weights of alanine dipeptide. *Error bars* indicate the standard deviation w.r.t. 10 runs. Deviation from the reference values is indicated as sub-bar plot. 15 nodes, with 100, 500 and 1,000 ps minimum sampling time per node (*dark gray* to *light gray*), including an auxiliary 1,000 ps double precision trial, random MD starting impulse, random node placement

4.2.2 Stability regarding choice of dynamics

Second, similar simulations (random MD starting impulse, 15 randomly placed nodes, 1000 ps minimum sampling time per node) were performed while exchanging the common MD integrator with a stochastic dynamics (SD) integrator. Both integrators were compared in the context of an NVT and an NpT ensemble, the latter realized by using a Parrinello-Rahman barostat. All trial runs were conducted with a double precision version of GROMACS. The results are shown in Fig. 10. In both NVT and NpT ensemble, the SD integrator delivers better results with regard to the standard deviation over ten runs. In the NVT ensemble, the largest error obtained with the SD integrator is 3.618 % (conformation P_{II}), compared to 5.86 % when the MD integrator is used (conformation α_R). This gap becomes somewhat closer in the NpT ensemble, where the largest error obtained with the SD integrator is 5.35 %, compared to 6.3 % when the MD integrator is used (both w.r.t. conformation α_R).

The chosen dynamics also has an impact on the mean conformational weights. When the SD integrator is used, the largest conformation, P_{II} is sampled less dominant than with the MD integrator (NVT: 36.18 % compared to 39.68 %, and NpT: 39.02 % compared to 44.93 %). Instead, the conformational weight is distributed more equally over the minor conformations α_R, α_P and α_L.

The results show an acceptable agreement with the reference weights that were extracted from the 200 ns MD trajectory for all runs using 500 ps or more minimum sampling time per node, at least for the runs conducted with the MD integrator (i.e. the same integrator that was used for the long-time trajectories used as reference). Long-time data from the SD integrator is not available, but it can be expected to
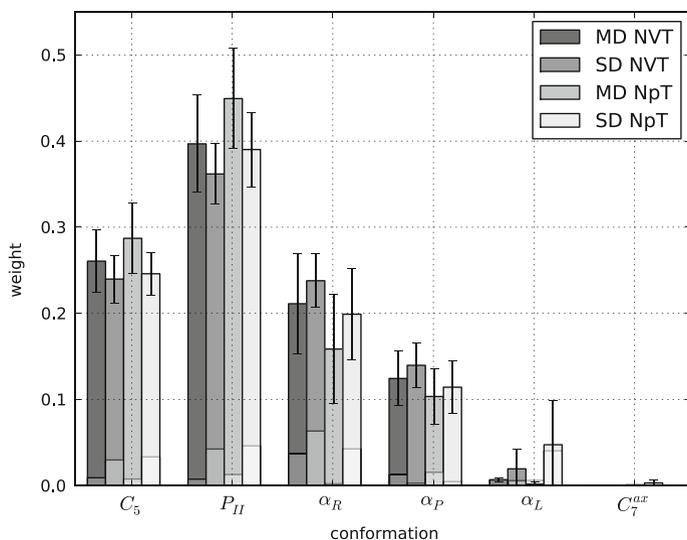
Fig. 10 Conformational weights of alanine dipeptide. *Error bars* indicate the standard deviation w.r.t. 10 runs. Deviation from the reference values is indicated as sub-bar plot. 15 nodes, 1,000 ps minimum sampling time per node, random node placement, with MD integrator (NVT), SD integrator (NVT), MD integrator (NpT), and SD integrator (NpT), *dark gray* to *light gray*

Table 4 Averaged conformational weights of alanine dipeptide at 300 K in the NVT and in the NpT ensemble, derived from ten runs of ZIBgridfree using the Amber-99SB force field (MD integrator, double precision GROMACS)

	c	C_5	P_{II}	α_R	α_P	α_L	C_7^α
NVT	$\tilde{w}_c$	0.2606	0.3968	0.2112	0.1244	0.0067	0.0003
NpT	$\tilde{w}_c$	0.2871	0.4493	0.1586	0.1035	0.0015	0.0001

deliver a slightly different distribution. In general, the largest deviation is found for the α_R conformation: ZIBgridfree tends to overweight α_R by about 4 %, a weight that is mostly drawn from the α_P, and partly from the α_L conformation. As the conformations of alanine dipeptide tend to have notable overlapping regions (as opposed to the well-separated conformations of n-pentane), the error might not only be due to insufficient sampling, but also to imperfect clustering of certain states in transient regions. For comparison, the conformational weights in the NVT and the NpT ensemble, obtained from ZIBgridfree with 15 nodes and 1000 ps minimum sampling time per node and averaged over ten runs, are given in Table 4.

4.3 Host-guest binding process in explicit solvent

In order to give a proof of concept for a different application of the algorithm, the analysis of a small crown ether-ammonium host-guest binding process is presented in the following. The system consists of an 18-crown-6 dimer host molecule (**C6**), and
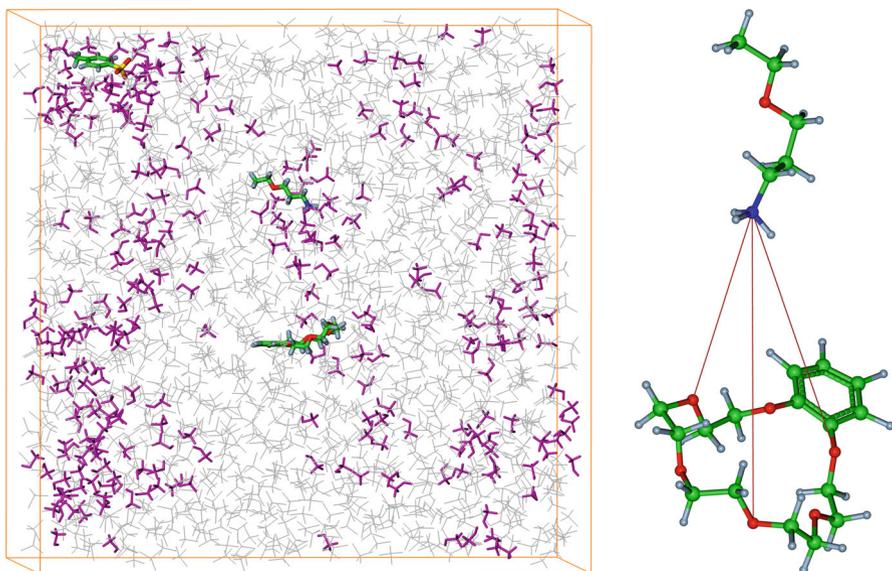
Fig. 11 *Left* Host guest system **C6**-(**MonoG1**+H)-OTs after 2 ns equilibration of the solvent mixture (10:1 chloroform-methanol) in the position-restrained unbound state at 298 K (chloroform = *gray*, methanol = *purple*). Polar clusters of methanol molecules are clearly visible. *Right* Three distances between ammonium moiety and binding site form the internal coordinates for the system (Color figure online)

an ammonium ion guest molecule incorporating a short flexible tail (**MonoG1**+H). The thermodynamics of the formation of complex (**MonoG1**+H)•**C6** in a mixture of chloroform and methanol and in the presence of tosylate counter ions (denoted as OTs) could be characterized recently, along with an analoguous bivalent system [42].

The presampling for this system was obtained by free diffusion MD simulations involving the complete explicit solvent and counter ion setup. One out of five 10 ns MD simulations starting from the unbound state (Fig. 11, left) with about 1.5 nm separation between host and guest molecule captured a binding event. The relatively low yield can be explained by the fact that (i) both host and guest molecule are rather small and mobile and therefore subject to rapid diffusion in the box and (ii) the complexation of host and guest is hindered by the counter ion associating with the ammonium moiety, obscuring the interaction site. Consequently, not every close contact between host and guest immediately induces complex formation. The trajectory which captured the binding event was prolonged to a total of 100 ns without showing indications for complex dissociation.

4.3.1 Discretization and metastability analysis

The conformational space discretization was based on a set of internal coordinates consisting of three strongly correlated distances between ammonium moiety and 18-crown-6 ring (Fig. 11, right). In order to remove the abundance of unbound states not related to the binding process from the presampling data, states with distances of more than 1.8 nm distance between the interaction sites were discarded. A total
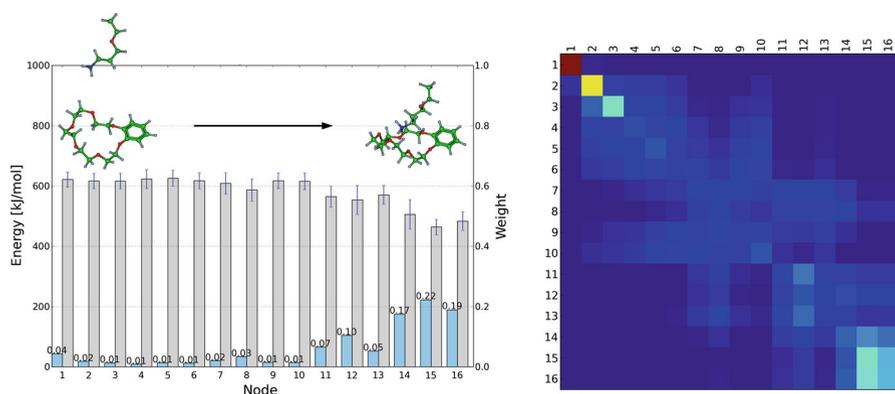
Fig. 12 *Left* Mean potential energy (*gray*) and corrected discretization node weights (*blue*) for the 16 discretization nodes of system **C6**-(**MonoG1**+H)-OTs. *Right* Overlap integral matrix *S* with 16 discretization nodes. Large matrix entries (*red*, *yellow*) indicate no or only minor overlap with neighboring discretization nodes and represent isolated and/or stable regions. Discretization nodes with many off-diagonal entries (*blueish*) exhibit a significant overlap with their neighborhood and thus mark transient regions (Color figure online)

number of 16 discretization nodes was placed equidistantly in the remaining part of conformational space. For each discretization node, 5×500 ps of MD in the NVT ensemble were simulated at a temperature of 298 K, with each 500 ps run starting at the initial position of the discretization node using a random starting impulse vector, leading to a joint sampling time of 40 ns for the complete discretization.

The thermodynamic reweighting of the partial distributions sampled for the 16 discretization nodes documents a decrease in potential energy that is directly related to the distance of the host to the guest molecule (Fig. 12, left). A notable improvement in the interaction energy sets in with node 11 at an approximate host-guest distance of 7.5 Å, and culminates in the bound state (nodes 14, 15 and 16). While nodes 1–10 have similar (and low) thermodynamic weights, nodes 11–16, covering host-guest distances of 7.5 Å and nearer, represent the largest share of the distribution.

The *S* matrix of the discretization (Fig. 12, right) exhibits an isolated unbound state represented by node 1, an articulate "block" for the bound state (nodes 14, 15 and 16 in the lower right corner) and a large transition region in between. Accordingly, the clustering with PCCA+ identifies three metastable states, namely the unbound state (UB) with a weight of 4.58 %, the almost bound state (AB) with a weight of 9.04 %, and the bound state (SB) with a weight of 86.38 % (Fig. 13). State UB is detached from the rest of the system except for a small degree of communication involving nodes 2 and 3 that leads into state AB. State AB, in turn, exhibits a fluent transition into state SB. Nodes 6, 9, and in particular 10 mark the transition region between the two clusters AB and SB. Nodes 14, 15 and 16 have the highest membership w.r.t. to state SB, and represent the proper bound state.

In order to look into the transition behavior on the level of the metastable states, additional unrestrained short-time MD simulations in the NVE ensemble were conducted. The unrestrained "transition nodes" (as opposed to the discretization nodes used for sampling the stationary distribution) were placed in regions of conformational space that mark interfaces between the different metastable states, and thus are
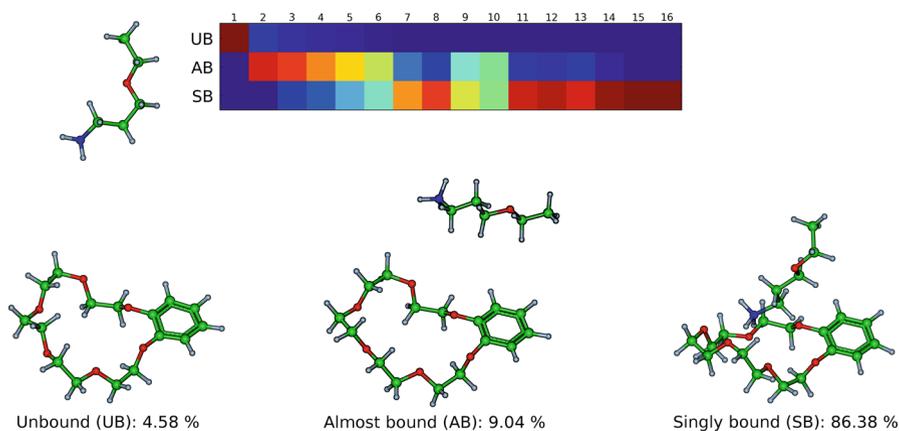
Fig. 13 The $\chi^\top$ matrix (*top*) of system **C6**-(**MonoG1**+H)-OTs groups the 16 discretization nodes into three metastable states: The unbound state (UB, left), the almost bound state (AB, center), and the singly bound state (SB, right). The *colors* in the $\chi^\top$ matrix indicate the degree of membership of a discretization node to a given metastable state: *dark red* = highest degree of membership, *dark blue* = no membership. Nodes 6, 9 and 10 represent transition regions that belong almost evenly to the two metastable states AB and SB (Color figure online)

prone to reveal the associated transition behavior more readily than simulations that are started exactly within the center of a metastable region.

$$P_c(\tau) = \begin{matrix} \\ UB \\ AB \\ SB \end{matrix} \begin{pmatrix} UB & AB & SB \\ 0.9868 & 0.0132 & 0. \\ 0.0489 & 0.7584 & 0.1928 \\ 0.0002 & 0.0029 & 0.9969 \end{pmatrix}, \text{ with } \tau = 100 \text{ ps} \qquad (16)$$

Using a total of 45 transition nodes started for ten runs of 100 ps each using a random starting impulse (45 ns additional sampling time), the transition probability matrix $P_c(\tau)$ is obtained (Matrix 16). Within the short time span of 100 ps, the system has a very high probability to remain in either state UB or state SB. Given the system is in state AB, it is more likely to make the transition into the bound state ($\approx 19\%$) than into the unbound state ($\approx 5\%$).

5 Conclusion

As far as the limited number of test cases allows, it was shown that algorithm and software perform reasonably well in determining the conformational weights and inter-conformational transition probabilities of small molecular systems in both vacuum and explicit solvent. The performance of the method in comparison to other approaches was not evaluated explicitly, but, given that a similar algorithmic framework is used, should be in the order of available umbrella sampling approaches. Due to the fact that ZIBgridfree is dependent on the availability of a presampling of conformational space from which discretization nodes can be selected, the cost of obtaining the presampling

would have to be added to the overall performance balance. The cost of generating an adequate presampling is dependent on the system in question. For instance, a series of docking poses of a small molecule in a protein binding pocket would also serve as a valid starting point for using ZIBgridfree.

Given the efficiency of current MD code in generating even very long trajectories, the need for a relatively complex algorithm like ZIBgridfree can be questioned. We see the advantage of using ZIBgridfree mainly in the more directed generation of sampling data in transient regions of conformational space (reducing the amount of redundant sampling data) and the possibility to add another level of parallelization to the sampling process, namely parallel sampling of the discretization nodes (i.e. conformational space regions), which can be used to complement the parallel force field evaluation in order to increase the overall sampling efficiency. Furthermore, the use of collective variables (i.e. internal coordinates) and the integrated clustering approach lead to a level of abstraction that significantly facilitates the analysis of the sampling data, the identification of relevant events and their biological or chemical interpretation.

In upcoming work, we would like to improve the usability of software and algorithm. In particular, we would like to eliminate certain discretization parameters that currently have to be set by the user. Ideally, for a given system, an optimal number of discretization nodes is proposed beforehand. The ZIBgridfree scheme is also a suitable discretization of the infinitesimal generator described in [43]. Further invesigation in this direction will also be done in future.

6 Supporting information available

The source code of ZIBgridfree is available at https://github.com/CMD-at-ZIB/ZIBMolPy.

References

1. G. Torrie, J. Valleau, Nonphysical sampling distributions in Monte Carlo free-energy estimation: umbrella sampling. J. Comput. Phys. **23**, 187–199 (1977)
2. A. Amadei, A. Linssen, H. Berendsen, Essential dynamics of proteins. Proteins: Struct. Funct. Genet. **17**, 412–425 (1993)
3. Y. Sugita, Y. Okamoto, Replica-exchange molecular dynamics method for protein folding. Chem. Phys. Lett. **314**, 141–151 (1999)
4. M. Bonomi, D. Branduardi, G. Bussi, C. Camilloni, D. Provasi, P. Raiteri, D. Donadio, F. Marinelli, F. Pietrucci, R. Broglia, M. Parrinello, PLUMED: a portable plugin for free energy calculations with molecular dynamics. Comput. Phys. Comm. **180**, 1961–1972 (2009)
5. M. Weber, H. Meyer, ZIBgridfree—Adaptive Conformation Analysis with Qualified Support of Transition States and Thermodynamic Weights (2005)
6. P. Deuflhard, in *Trends in Nonlinear Analysis*, eds. by M. Kirkilionis, S. Krömker, R. Rannacher, F. Tomi (Springer, Berlin, 2003), pp. 269–288

7. S. Kube, M. Weber, A coarse graining method for the identification of transition rates between molecular conformations. J. Chem. Phys. **126**, 024103 (2007)
8. M. Sarich, F. Noe, C. Schuette, On the approximation quality of Markov state model. Multiscale Model. Simul. **8**, 1154–1177 (2010)
9. K. Fackeldey, A. Bujotzek, M. Weber, A meshless discretization method for Markov state models applied to explicit water peptide folding simulations, in *Lecture Notes in Computational Science and Engineering*, vol. 89 (Springer, 2012), pp. 141–154
10. J. Prinz, J. Chodera, V. Pande, W. Swope, J.C. Smith, Noe F., Optimal use of data in parallel tempering simulations for the construction of discrete-state Markov models of biomolecular dynamicsReplica-exchange molecular dynamics method for protein folding. J Chem Phys. **134**, 244108 (2011)
11. M. Weber, Meshless Methods in Conformation Dynamics. Doctoral thesis, Freie Universität Berlin, Department of Mathematics and Computer Science (2006)
12. B. de Groot, D. van Aalten, R. Scheek, A. Amadei, G. Vriend, H. Berendsen, Prediction of protein conformational freedom from distance constraints. Proteins: Struct. Funct. Genet. **29**, 240–251 (1997)
13. D. Shepard, A two-dimensional interpolation function for irregularly-spaced data, in *Proceedings of 23rd ACM National Conference* (1968), pp. 517–524
14. B. Hess, C. Kutzner, D. van der Spoel, E. Lindahl, GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. J. Chem. Theory Comput. **4**, 435–447 (2008)
15. A. Gelman, D. Rubin, Inference from iterative simulation using multiple sequences. Stat. Sci. **7**, 457–472 (1992)
16. M. Klimm, A. Bujotzek, M. Weber, Direct reweighting strategies in conformation dynamics. MATCH Commun. Math. Comput. Chem. **65**, 333–346 (2011)
17. S. Kumar, J.M. Rosenberg, D. Bouzida, R.H. Swendsen, P.A. Kollman, Multidimensional free-energy calculations using the weighted histogram analysis method. J. Comput. Chem. **16**, 1339–1350 (1995)
18. B. Roux, The calculation of the potential of mean force using computer simulations. Comput. Phys. Commun. **91**, 275–282 (1995)
19. M. Weber, S. Kube, L. Walter, P. Deuflhard, Stable computation of probability densities for metastable dynamical systems. Multiscale Model. Simul. **6**, 396–416 (2007)
20. R. Sinkhorn, A relationship between arbitrary positive matrices and doubly stochastic matrices. Ann. Math. Stat. **35**, 876–879 (1964)
21. P. Deuflhard, M. Weber, Robust Perron cluster analysis in conformation dynamics. Linear Algebra Appl. **398**, 161–184 (2005)
22. M. Weber, S. Kube, Robust Perron cluster analysis for various applications in computational life science, in *Computational Life Sciences: First International Symposium, CompLife 2005*, ed. by M.R. Berthold et al. (Springer, Heidelberg, 2005), pp. 57–66
23. V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, C. Simmerling, Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins: Struct.: Funct. Bioinf. **65**, 712–725 (2006)
24. A. da Silva, W. Vranken, ACPYPE—AnteChamber PYthon Parser interfacE. BMC Res. Notes **5**, 367 (2012)
25. J. Wang, R. Wolf, J. Caldwell, P. Kollman, D. Case, Development and testing of a general Amber force field. J. Comput. Chem. **25**, 1157–1174 (2004)
26. J. Wang, W. Wang, P. Kollman, D. Case, Automatic atom type and bond type perception in molecular mechanical calculations. J. Mol. Graph. Model. **25**, 247–260 (2006)
27. D. Case, T. Cheatham III, T. Darden, H. Gohlke, R. Luo, K. Merz Jr, A. Onufriev, C. Simmerling, B. Wang, R. Woods, The Amber biomolecular simulation programs. J. Comput. Chem. **26**, 1668–1688 (2005)
28. A. Jakalian, B. Bush, D. Jack, C. Bayly, Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. J. Comput. Chem. **21**, 132–146 (2000)
29. A. Jakalian, D. Jack, C. Bayly, Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. J. Comput. Chem. **23**, 1623–1641 (2002)
30. H.W. Horn, W.C. Swope, J.W. Pitera, J.D. Madura, T.J. Dick, G.L. Hura, T. Head-Gordon, Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. J. Chem. Phys. **120**, 9665–9678 (2004)
31. H. Horn, W. Swope, J. Pitera, Characterization of the TIP4P-Ew water model: vapor pressure and boiling point. J. Chem. Phys. **123**, 194504 (2005)

32. C. Caleman, P. van Maaren, M. Hong, J. Hub, L. Costa, D. van der Spoel, Force field benchmark of organic liquids: density, enthalpy of vaporization, heat capacities, surface tension, isothermal compressibility, volumetric expansion coefficient, and dielectric constant. J. Chem. Theory Comput. **8**, 61–74 (2011)
33. D. van der Spoel, P. van Maaren, C. Caleman, GROMACS molecule & liquid database. Bioinformatics **28**, 752–753 (2012)
34. U. Essmann, L. Perera, M. Berkowitz, T. Darden, H. Lee, L. Pedersen, A smooth particle mesh Ewald method. J. Chem. Phys. **103**, 8577 (1995)
35. G. Bussi, D. Donadio, M. Parrinello, Canonical sampling through velocity rescaling. J. Chem. Phys. **126**, 014101 (2007)
36. W. Van Gunsteren, H. Berendsen, A leap-frog algorithm for stochastic dynamics. Mol. Simul. **1**, 173–185 (1988)
37. M. Parrinello, A. Rahman, Polymorphic transitions in single crystals: a new molecular dynamics method. J. Appl. Phys. **52**, 7182–7190 (1981)
38. S. Nose, M. Klein, Constant pressure molecular dynamics for molecular systems. Mol. Phys. **50**, 1055–1076 (1983)
39. C. Schütte, Conformational Dynamics: Modelling, Theory, Algorithm and Application to Biomolecules. Habilitation thesis, Freie Universität Berlin, Department of Mathematics and Computer Science (1999)
40. T. Halgren, Merck molecular force field: I–V. J. Comput. Chem. **17**, 490–641 (1996)
41. J. Chodera, W. Swope, J. Pitera, K. Dill, Long-time protein folding dynamics from short-time molecular dynamics simulations. Multiscale Model. Simul. **5**, 1214 (2006)
42. L. von Krbek, *Multivalente Krone–Ammonium–Komplexe* (Freie Universität Berlin, Fachbereich Chemie, Biologie, Pharmazie, Master's thesis, 2012)
43. M. Weber, A Subspace Approach to Molecular Markov State Models via a New Infinitesimal Generator. Habilitation thesis, Freie Universität Berlin, 2011; Department of Mathematics and Computer Science. This material is available free of charge via the Internet at http://pubs.acs.org