Traversing the free-energy pathways of intricate biomolecular processes

Enhanced simulation development and applications

Pérez de Alba Ortíz, A.

Publication date
2021

2

Methods

If in the first act you have hung a pistol on the wall,
then in the following one it should be fired.
Otherwise don't put it there.

Anton Chekhov

We present an overview of the key methods used in this thesis. First, fundamentals of statistical mechanics are introduced in order to connect molecular simulation measurements to experimental observables. Then, we establish the principles of molecular dynamics (MD) simulations, with particle interactions handled at force-field and density-functional-theory levels, as well as hybrid. Since transitions over energy barriers—in which one is often interested—are *rare events* on MD simulation timescales, we introduce enhanced sampling methods. In particular, we focus on well-established biasing, or free-energy, techniques: umbrella sampling, constrained MD, steered MD and metadynamics. Finally, we introduce the workhorse of this thesis, path-based free-energy methods, which boost the capabilities of standard schemes to efficiently tackle complex transitions in high-dimensional free-energy landscapes.

Section 2.3.2 based on: A. Pérez de Alba Ortíz, J. Vreede, and B. Ensing. *The Adaptive Path Collective Variable: A Versatile Biasing Approach to Compute the Average Transition Path and Free Energy of Molecular Transitions.* In: Bonomi M., Camilloni C. (eds) Biomolecular Simulations. Methods in Molecular Biology, **vol. 2022**, pp. 255-290. Humana, New York, NY, (2019).

2.1. Statistical mechanics

Much of the detailed atomic information that is routinely produced during a molecular simulation, such as an atomic positions, velocities or forces, are not directly measurable in an experiment. The mathematical theory that connects the observable macroscopic behavior of a molecular system to its underlying microscopic components is statistical mechanics. Here, we present the main fundamentals on how observable properties can be computed as averages of microscopic quantities over the sampled configurations. More involved derivations of statistical mechanics can be found in [1].

Consider a system of N particles, within a volume V, at an absolute temperature T, and in the canonical ensemble, i.e. at constant N, V and T. The probability of finding the system in a given microstate in phase space with positions $\mathbf{r}^N$ and momenta $\mathbf{p}^N$ is given by a Boltzmann distribution [2] depending on the Hamiltonian, or total energy, $\hat{H}(\mathbf{r}^N, \mathbf{p}^N)$:

$$P(\mathbf{r}^N, \mathbf{p}^N) = \frac{1}{h^{3N}N!}\frac{e^{-\beta \hat{H}(\mathbf{r}^N,\mathbf{p}^N)}}{Z(N,V,T)}, \tag{2.1}$$

with $\beta = \frac{1}{k_\mathrm{B}T}$, the Boltzmann constant k_B, and the canonical partition function:

$$Z(N,V,T) = \frac{1}{h^{3N}N!}\iint e^{-\beta \hat{H}(\mathbf{r}^N,\mathbf{p}^N)} d\mathbf{r}^N d\mathbf{p}^N, \tag{2.2}$$

where the factor $\frac{1}{h^{3N}}$, with the Planck constant h, is inserted to keep the partition function dimensionless, and the factor $\frac{1}{N!}$ accounts for the indistinguishability of identical particles.

One can also demonstrate that the Helmholtz free energy [3] is equal to:

$$F = -k_\mathrm{B}T \ln Z = U - TS, \tag{2.3}$$

with the potential energy U and entropy S. The relation of the free energy with the partition function, and with other key quantities, has posed it as a common "currency" in molecular simulations.

Considering Equation 2.1, the expected value of a macroscopic observable $\mathcal{O}$ can be obtained via the ensemble average:

$$\langle \mathcal{O} \rangle = \frac{\iint \mathcal{O}(\mathbf{r}^N, \mathbf{p}^N) e^{-\beta \hat{H}(\mathbf{r}^N,\mathbf{p}^N)} d\mathbf{r}^N d\mathbf{p}^N}{\iint e^{-\beta \hat{H}(\mathbf{r}^N,\mathbf{p}^N)} d\mathbf{r}^N d\mathbf{p}^N}, \tag{2.4}$$

which can be evaluated by stochastic sampling, via Monte Carlo methods [4].

Alternatively, the ensemble average can be replaced with a time average by considering the ergodic hypothesis [5]. That is, the assumption that after a sufficiently long time, a representative part of the entire phase space of equiprobable microstates will be sampled. Then, we calculate:

$$\langle \mathcal{O} \rangle = \lim_{t\to\infty} \frac{1}{t}\int_0^t \mathcal{O}(\mathbf{r}^N, \mathbf{p}^N) dt' \tag{2.5}$$

Such sampling in time can be performed via molecular dynamics (MD) simulations [6]. Apart from sampling states, MD also generates the atomic motions that drive molecular processes; thus providing unique insight about the inner workings of, for example, biomolecules. This key feature of MD is used in this thesis to observe molecular transition mechanisms. In the following section, we provide details of how to perform MD simulations, and how to ensure sufficient sampling.

2.2. Molecular dynamics simulations

The first general method for MD simulations was proposed in 1959 by the late B. Alder [6]. In the next 50 years, and counting, the technique has enabled innumerable studies of a huge variety of systems in chemistry, biophysics and material science. Here, we explain the fundamentals, which can be consulted further in [1].

A system of N particles with positions $\mathbf{r}^N$ can be propagated in time, t, according to Newton's equations of motion [7]. For particle i with mass m_i, the force is given by:

$$\mathbf{f}_i = m_i \ddot{\mathbf{r}}_i = -\nabla_{\mathbf{r}_i} U(\mathbf{r}^N) \tag{2.6}$$

where the double over-dot denotes the second derivative of the position with respect to time, i.e. the acceleration, in Newton's notation.

The equation above can be integrated to propagate the N particles in time. To do so, several numerical integrators have been developed. Simply put, most schemes are based on Taylor expansions of $\mathbf{r}(t+\Delta t)$ combined in such way that some higher-order terms are cancelled. The Verlet integration method [8] has good numerical stability, time-reversibility and symplecticity, i.e. conservation of the phase space volume, at a modest computational cost [1]. A popular version is the velocity Verlet algorithm [9]:

$$\mathbf{r}_i(t+\Delta t) = \mathbf{r}_i(t) + \dot{\ddot{\mathbf{r}}}_i(t)\Delta t + \frac{\mathbf{f}_i(t)}{2m_i}\Delta t^2 \tag{2.7}$$

$$\dot{\ddot{\mathbf{r}}}_i(t+\Delta t) = \dot{\ddot{\mathbf{r}}}_i(t) + \frac{\mathbf{f}_i(t) + \mathbf{f}_i(t+\Delta t)}{2m_i}\Delta t \tag{2.8}$$

In this thesis, we use the velocity Verlet algorithm and its close variant, the leapfrog algorithm [10], with staggered steps for the positions $\mathbf{r}_i(t+\Delta t)$ and velocities $\dot{\ddot{\mathbf{r}}}_i(t+\Delta t/2)$.

In MD simulations, the system usually requires to be maintained at certain conditions corresponding to the chosen ensemble. For example, to keep a constant average temperature T, we employ schemes called thermostats, which modulate the kinetic energy of the atoms. In this thesis, we mainly apply the canonical sampling through velocity rescaling (CSVR) thermostat [11]. Moreover, in an NPT ensemble, the pressure of the system can also be controlled by a barostat that scales the size of the simulation box. In this thesis, we use the Raman-Parrinello barostat [12].

The potential energy gradient $-\nabla_{\mathbf{r}_i} U(\mathbf{r}^N)$ in Equation 2.7 can be calculated based on the interactions between the particles of the system. These interactions can be modeled at different resolutions, spanning from electronic, to atomic, to coarse-grained. The first two scales are employed in this thesis, and explained in the next sections.

2.2.1. Force fields

A force field is a model, consisting of a set of equations representing interatomic interactions, which enables the calculation of the potential energy of a system and its derivative, i.e. forces. The interatomic interactions are typically categorized into bonded and nonbonded terms:

$$U = \underbrace{U_{\text{bonds}} + U_{\text{angles}} + U_{\text{torsions}}}_{U_{\text{bonded}}} + \underbrace{U_{\text{Coulomb}} + U_{\text{van der Waals}}}_{U_{\text{nonbonded}}} \tag{2.9}$$

The bonded terms include:

$$U_{\text{bonds}} = \sum_{\text{bonds}} k_b (r_b - r_{b,0})^2, \tag{2.10}$$

$$U_{\text{angles}} = \sum_{\text{angles}} k_a (\theta_a - \theta_{a,0})^2, \tag{2.11}$$

$$U_{\text{torsions}} = \sum_{\text{torsions}} k_{\tilde{t}} (1 + \cos(n_{\tilde{t}} \phi_{\tilde{t}}))^2 \tag{2.12}$$

where the b, a and $\tilde{t}$, indices loop over all bonds, angles and torsions, respectively; the k factors are the force constants for each interaction; and the zero subindex indicates the reference value of each bond distance r, angle value θ, or torsion value ϕ.

The nonbonded terms are typically:

$$U_{\text{Coulomb}} = \sum_{i<j} \frac{1}{4\pi\epsilon_0} \frac{q_i q_j}{r_{ij}}, \tag{2.13}$$

$$U_{\text{van der Waals}} = \sum_{i<j} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{2.14}$$

Coulomb's electrostatic interaction [13] is given by the charges q of atoms i and j, the distance r_{ij} between them, and the vacuum permittivity ϵ_0. Van der Waals interactions [14] are modeled by a Lennard-Jones potential [15] with a well depth ϵ_{ij} and width σ_{ij}. There are several strategies to handle long-range nonbonded interactions and to account for periodic simulation boxes. We have used one of the fast and well-parallelized versions of Ewald summation throughout our studies, i.e. particle mesh Ewald (PME) [16, 17], as implemented in the GROMACS software [18].

Nowadays, well-known and robust force fields have been parametrized, based on experimental or *ab initio* data. Two well-known force-field families, AMBER [19] and

CHARMM [20], are used in this thesis. We also employ other force fields designed for nucleic acids [21], and for polysaccharides [22]. These parameter sets are typically made available to the scientific community via files that are compatible with robust software packages to run MD, such as GROMACS [18], which is prominently used in this thesis, or LAMMPS [23]. Standard force fields provide cost-efficient potential energies for simulations of molecules motions. However, chemical changes with bonds forming and breaking, as well as information about the electronic structure are out of their scope.

2.2.2. Density functional theory

Density functional theory (DFT) is a quantum mechanical method based on electron density. Its cost-efficiency has posed it as a wide-spread choice for *ab initio* MD. Here, we give a brief overview of its foundations. More details can be found in [24, 25]

Fundamentally, the behavior of electrons and atomic nuclei can be described by the non-relativistic time-independent Schrödinger equation [26]:

$$\hat{H} \left|\Psi_i\right\rangle = E_i \left|\Psi_i\right\rangle , \tag{2.15}$$

where the wave function $\left|\Psi_i\right\rangle$ is a stationary state with energy E_i, and $\hat{H}$ is the Hamiltonian of the system, whose full form is:

$$\hat{H} = \hat{T}_\mathrm{e} + \hat{V}_\mathrm{ee} + \hat{T}_\mathrm{n} + \hat{V}_\mathrm{nn} + \hat{V}_\mathrm{ne} \tag{2.16}$$

The operators of the Hamiltonian correspond to: the kinetic energy of electrons $\hat{T}_\mathrm{e}$, the electrostatic potential between electrons $\hat{V}_\mathrm{ee}$, the kinetic energy of nuclei $\hat{T}_\mathrm{n}$, the electrostatic potential between nuclei $\hat{V}_\mathrm{nn}$, and the electrostatic potential between electrons and nuclei $\hat{V}_\mathrm{ne}$,

The Hamiltonian can be simplified when considering the Born-Oppenheimer approximation [27], i.e. the assumption that nuclei, which are much slower and more massive than the electrons, can be regarded as immobile from the point of view of the electrons. One can therefore separate the nuclei kinetic energy $\hat{T}_\mathrm{n}$ and the interactions between nuclei $\hat{V}_\mathrm{nn}$ from the quantum mechanical Hamiltonian. The electron-nuclei interactions are then considered as an external potential $\hat{V}_\mathrm{ext}$. For consistency, the electron-electron interactions are relabeled as an internal potential $\hat{V}_\mathrm{int}$. The Hamiltonian is then rewritten as:

$$\hat{H} = \hat{T}_\mathrm{e} + \hat{V}_\mathrm{int} + \hat{V}_\mathrm{ext} \tag{2.17}$$

Even after this simplification, the Schrödinger equation is still not easily tractable. Analytical solutions are only available for the simplest of systems, such as the hydrogen atom. Numerical approaches remain extremely computationally expensive, due to the high-dimensionality of the many-body wave-function, $\Psi_i(\mathbf{r}_1, ..., \mathbf{r}_N)$ with N electrons.

In DFT, the many-body wavefunctions $\Psi_i(\mathbf{r}_1, ..., \mathbf{r}_N)$ are circumvented, and the problem is handled in terms of the electronic density $n(\mathbf{r})$. The Hohenberg-Kohn

(HK) theorems [28] state that: (1) the external potential V_{ext} is a unique functional of the ground-state electronic density $n(\mathbf{r})$; and (2) the functional returns the minimum ground-state energy if and only if the input is the exact ground-state density. In turn, the Kohn-Sham (KS) *ansatz* [29] maps the HK problem onto an auxiliary non-interacting electron system. Similar to the Hartree-Fock (HF) approach, one considers the electrons as moving through an averaged effective potential, rather than interacting between themselves. The V_{int} term is then split into a Hartree term containing all electron-electron interactions—even spurious self-interactions—and an exchange-correlation (xc) term, accounting for self-interaction corrections and Pauli's exclusion principle. Then, we express the KS DFT total energy as:

$$E_{\text{KS}}[n] = T_{\text{non-interacting}}[n] + E_{\text{Hartree}}[n] + E_{\text{xc}}[n] + \underbrace{\int V_{\text{ext}}(\mathbf{r})n(\mathbf{r})d\mathbf{r}}_{E_{\text{ext}}}, \tag{2.18}$$

with the non-interacting electron density depending on the orbitals $\varphi_i(\mathbf{r})$:

$$n(\mathbf{r}) = \sum_i^N |\varphi_i(\mathbf{r})|^2, \tag{2.19}$$

the kinetic energy:

$$T_{\text{non-interacting}}[n] = -\frac{1}{2}\sum_i^N \langle\varphi_i|\,\nabla^2\,|\varphi_i\rangle \tag{2.20}$$

and the Hartree term:

$$E_{\text{Hartree}}[n] = \frac{1}{2}\iint \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}d\mathbf{r}' \tag{2.21}$$

The KS DFT total energy can be obtained by solving iteratively:

$$\Big(-\frac{1}{2}\nabla^2 + V_{\text{Hartree}}[n] + V_{\text{xc}}[n] + V_{\text{ext}}\Big)\varphi_i(\mathbf{r}) = \epsilon_i\varphi_i(\mathbf{r}), \tag{2.22}$$

with $V_{\text{Hartree}}(\mathbf{r}) = \int \frac{n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}'$ and $V_{\text{xc}}(\mathbf{r}) = \frac{\delta E_{\text{xc}}[n]}{\delta n(\mathbf{r})}$

There is no exact formulation for the exchange-correlation functional. Several approximations have been developed, and the generation of novel, more accurate and cost-effective functionals is still an active field. In this thesis we employ a generalized gradient approximation (GGA) functional that depends on the local density and its gradient.

Moreover, the electron orbitals $\varphi_i(\mathbf{r})$ are usually expressed in terms of basis functions, for which in principle complete basis sets are required. In this thesis, we employ the Gaussian and plane waves method as implemented in CP2K's Quickstep [30–32]. This well-known software package also includes pseudopotentials to account for the interaction between core and valence electrons, as well as dispersion corrections [33] to account for van der Waals interactions.

2.2.3. Hybrid quantum mechanics/molecular mechanics

Many biomolecular processes involve chemical reactions, which occur in complex environments containing protein residues, nucleic acids, water molecules, ions, etc. Including the entire system in a quantum mechanical calculation is extremely expensive, even with today's algorithms and computational resources. Multiscale schemes, such as hybrid quantum mechanics/molecular mechanics (QM/MM) [34, 35] have made studies of biochemical processes feasible.

In QM/MM, a molecular system is divided into a QM—e.g. DFT—calculation, containing the crucial atoms involved in the reaction, embedded inside a more affordable MM—i.e. force field—simulation that includes the surrounding environment. The Hamiltonian for a QM/MM system can be constructed by an additive scheme, yielding the total energy:

$$E_{\mathrm{total}} = E_{\mathrm{QM}} + E_{\mathrm{MM}} + E_{\mathrm{QM/MM}}, \tag{2.23}$$

where the QM and MM terms account for their respective regions, and the QM/MM term includes the interactions between the two, which are typically divided into bonded and nonbonded contributions.

$$E_{\mathrm{QM/MM}} = E_{\mathrm{QM/MM}}^{\mathrm{bonded}} + E_{\mathrm{QM/MM}}^{\mathrm{nonbonded}} \tag{2.24}$$

The bonded interactions are usually handled by *link*—also called *capping*—atoms, which saturate the covalent bonds that are “cut” by the boundary between the QM and the MM zones. The nonbonded interactions are handled by electrostatic embedding—i.e. including the MM nuclei into the V_{ext} term of the QM calculation—and by Lennard-Jones potentials.

In this thesis we use the additive QM/MM implementation of the CP2K software package [31, 32]. There are several further alternatives for QM/MM models [36], including—but not limited to—coupling schemes based on force rather than on the Hamiltonian, adaptive QM regions from which atoms can enter or leave, or polarizable MM regions.

2.3. Enhanced sampling

As explained in Section 2.1, relevant macroscopic properties of a system can be extracted from MD simulations by means of time-averaging. However, the sampling of a system can be hindered by free energy barriers. In fact, the probability of observing a transition over an energy barrier decreases exponentially with the barrier height (see Equation 2.1). Therefore, crosses over barriers higher that a couple of $k_{\mathrm{B}}T$—i.e. the energy of thermal fluctuations—are typically *rare events*; occurring extremely infrequently and quickly within timescales accessible *in silico*. However, these events are not rare *in vivo*, and many biologically relevant processes involve conformational or chemical changes that overcome energy barriers. Given that brute-force MD is not a sustainable solution, the scientific community has devised several so-called enhanced sampling methods over the last decades.

Enhanced sampling techniques can be roughly divided into three categories: (1) in biasing methods, the dynamics of the system are biased by an external potential to favor a desired transition [37–41]; (2) in transition path methods, the dynamics of the system are focused on reactive trajectories, or transition paths [42–49]; and (3) in tempering methods, the temperature of (part of) the system is increased, thus enhancing the thermal fluctuations and accelerating barrier-crossing [50–52]. In this thesis, we employ biasing, or free-energy, methods in combination with concepts from transition path methods.

2.3.1. Free-energy methods

In order to sample rare transitions, biasing methods introduce an external potential to influence a set of key descriptive degrees of freedom that drive the molecular transition. We refer to these set of n descriptors as the collective variables (CVs)[1], $\{z_i(\mathbf{r})\}$, with $i = 1...n$, which are functions of the atomic positions. For example, in a chemical reaction that involves one bond breaking and another one forming, the CVs could be the two corresponding bond distances. Similarly, in a conformational change involving three residues rotating, three dihedral angles could be the CVs.

The main idea of biasing methods is to use the external potential to force the occurrence of the rare transition, but do it in such a way that the effect of the bias can be quantified and corrected for afterward. By analyzing and post-processing the biased sampling, one can recover valuable free-energy profiles (for one CV), or surfaces (for several CVs), with interpretable (meta)stable state valleys and barriers separating them. The population ratios between reactants and products can be derived from their free energies, and rate constants can be calculated based on the barriers. Such quantities can then be compared with—or predict—experimental measurements.

In this thesis, we used the PLUMED package [53, 54]—which can be used in combination with several force-field and DFT MD codes—to calculate CVs, exert biasing potentials, and analyze our enhanced sampling simulations.

Umbrella sampling

Umbrella sampling (US) was first proposed in 1977 by Torrie and Valleau [37]. The method essentially works by adding external potentials, such that the probability of sampling high-energy configurations is increased. The transition progress is divided into several *windows*, to which MD simulations are harmonically restrained (see Fig. 2.1). For each window, the restraint is centered at a particular value of the CVs, $\{z_i(\mathbf{r})\}$, with $i = 1...n$, which is specified a priori. By tuning the stiffness of the restraints, the conformations near the free-energy barrier can be sampled. The restraints are usually harmonic, with the form:

$$V_{ij}(z_i) = \frac{k_{ij}}{2}(z_i - \tilde{z}_{ij})^2, \tag{2.25}$$

[1] At this point, it is pertinent to make a distinction between a CV and a reaction coordinate (RC). For purposes of this thesis, while the CVs compose a set of several transition descriptors—in which none alone can express the transition entirely—the RC is an abstract one-dimensional parameter that fully describes the progress of the transition. Another closely related term, an order parameter (OP), is a function of the atomic positions that can distinguish states, but not necessarily be used for biasing.

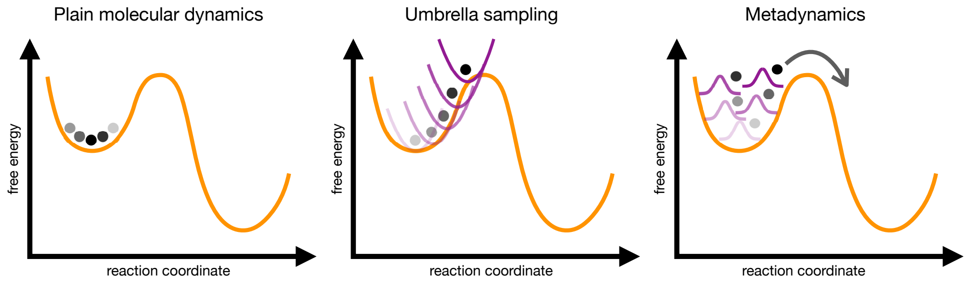

Figure 2.1: Schemes of the sampling near a free-energy barrier performed by: plain molecular dynamics, showing no barrier-crossing and remaining in the initial valley; umbrella sampling, showing windows with harmonic potentials that restrain the sampling near the barrier; and metadynamics, showing Gaussian potentials that repel the sampling from the initial valley, and drive it over the barrier.

where the index i indicates the CV, the index j indicates the window, $\tilde{z}_{ij}$ is the restraint center, and k_{ij} is the corresponding force constant.

For one CV, the unbiased free energy within each window, $F^{\mathrm{u}}_{ij}(z_i)$, can be calculated based on the biased sampling distribution, $P^{\mathrm{b}}_{ij}(z_i)$, and the corresponding biasing potential, $V_{ij}(z_i)$.

$$F^{\mathrm{u}}_{ij}(z_i) = -k_B T \ln P^{\mathrm{b}}_{ij}(z_i) - V_{ij}(z_i) + C_{ij} \tag{2.26}$$

The term C_{ij} is an unknown constant that differs for each window. In order to combine the several F^{u}_{ij} estimates from the different windows into a single free-energy profile, one can use schemes such as the weighted histogram analysis method [55, 56]. In general, to obtain a sensible free-energy profile, some sampling overlap is required across neighboring windows.

Constrained MD

In 1989, Carter and coworkers proposed constrained MD as a method to calculate free-energy profiles [38, 39]. Similar to US, the sampling of high-energy configurations is done by subdividing the transition progress, but instead of allowing the sampling of successive parts of the CV range, here the sampling is constrained at fixed values of the CV. While in the original method the constraints are holonomic—i.e. allowing no deviations—in this thesis we employ stiff restraints, by setting large values of k_{ij} in Equation 2.25. This is also done in similar methods [57, 58].

For a unimodal distribution of the biased sampling within window j, we can assume that the time-average of the force exerted by the stiff restraint, $\langle f_j(z_i)\rangle$, is an estimator of the underlying free-energy gradient:

$$\langle f_j(z_i)\rangle = -\frac{\partial F_j}{\partial z_i} \tag{2.27}$$

After collecting enough samples, a free-energy profile can be obtained by numerical integration. Moreover, error analysis can be performed based on the convergence of the average forces.

Note that neither constrained MD, nor US, produces continuous trajectories of the transition between two stable states. If the CVs do not adequately describe the transition mechanism, the free-energy profile will not agree with experimental rate constants or population ratios. A basic test of the free-energy profile can be made a posteriori based on the transition state, by performing a committor analysis. In principle, starting simulations at the top of the free-energy barrier, with random initial momenta, should yield an equal probability, i.e. 50 %, of descending—or *committing*—to either the reactant of the product state. The concept of the committor is further explained in following sections.

Steered MD

A harmonic restraint, such as the one described in Equation 2.25, can also be steered in time [40, 59]:

$$V(z_i,t) = \frac{k_i(t)}{2}(z_i - \tilde{z}_i(t))^2 \tag{2.28}$$

Performing such a steered MD simulation, enables us to drive a system over a free-energy barrier. Furthermore, an estimate of the free-energy difference between any two configurations, and therefore a free-energy profile, can be obtained via Jarzynski's equality [40]:

$$\langle e^{-\beta \mathcal{W}_i} \rangle = e^{-\beta \Delta F_i}, \tag{2.29}$$

where $\mathcal{W}_i$ is the work done by the moving restraint along the CV z_i.

However, this expression requires numerous trajectories to be evaluated, and typically only the trajectories with the smallest values of $\mathcal{W}_i$ contribute significantly to the average. Moreover, mechanistic details are difficult to extract from such trajectories, given that the speed of the steering is typically much faster than that of the unbiased process. In this thesis, we use steered MD mainly for testing the ability of a set of CVs to drive a transition, and to obtain intermediate configurations that can be used in a US or constrained MD run.

Metadynamics

In contrast to the previous techniques, metadynamics [41, 60] does not use a restraint to direct the sampling to high-energy regions. Instead, it adds a history-dependent bias that drives the system away from already visited, low-energy, configurations. This biasing potential is gradually built by depositing Gaussian kernels of height H and width W at previously visited configurations $\tilde{z}_i(t)$ (see Fig. 2.1):

$$V_{\text{bias}}(\mathbf{z},t) = \sum_t H(t)\exp\left(-\sum_{i=1}^{n}\frac{(z_i - \tilde{z}_i(t))^2}{2W_i^2}\right), \tag{2.30}$$

In the long-time limit, after the valleys have been covered by Gaussian kernels, the biasing potential converges to the negative free-energy surface:

$$F(\mathbf{z}) = -V_{\text{bias}}(\mathbf{z},t) \tag{2.31}$$

The exploratory quality of metadynamics, which does not require the definition of windows a priori, has posed it as the method of choice for numerous studies in biophysics and chemistry. Moreover, the method has been extended with several powerful algorithms, e.g. multiple walkers, or adaptive Gaussian potentials to aid convergence [60].

In metadynamics, an inadequate set of CVs is identifiable by hysteresis in the sampling, that is, by constant overfilling of the free-energy valleys with Gaussian potentials, which leads to a non-convergent estimation. This issue can sometimes be fixed by adding more CVs to complete the description of the transition. The handling of extra CV parameters is relatively simple, because the width of the Gaussians along each CV, W_i, is the only parameter to consider. However, for standard biasing methods, the convergence time of a free-energy estimation scales exponentially with the number of CVs; meaning that for practical purposes most calculations are limited to two or three CVs. This means that the study of highly complex transitions, which involve many descriptors, is often unaffordable. In the next section, we describe our approach to overcome that challenge.

2.3.2. Path-based free-energy methods

To circumvent the CV-dimensionality limitation of biasing methods, inspiration can be taken from techniques that focus on transition pathways, such as transition path sampling (TPS) [44, 45], transition interface sampling (TIS) [61] or milestoning [62]. The basic premise of path-based free-energy methods is to exert the external bias not on the high-dimensional CV-space, but on the one-dimensional progress parameter along an adaptive path connecting two known stable states, i.e. a path-CV. The path-CV can be iteratively optimized based on the free-energy gradient until a minimum free-energy path (MFEP), and its associated profile, are converged. Some well-known path-based biasing methods include the string method [43, 63–65], the nudged elastic band (NEB) [42], and the scheme proposed by Branduardi and coworkers [48]. Here, we describe our in-house developed methodology, originally introduced in [66], and extended in [67, 68].

Defining the path collective variable

Let us consider an N-particle system with positions $\mathbf{r}(t) \in \mathbb{R}^{3N}$ and velocities $\mathbf{v}(t) \in \mathbb{R}^{3N}$. The dynamics of the system are governed by a potential $U(\mathbf{r})$ and follow a canonical distribution at a temperature T. Let us also assume that the system has an underlying free-energy surface (FES) with two stable states A and B. The FES can be fully described by a set of n key descriptive degrees of freedom, the CVs $\mathbf{z} = \{z_i(\mathbf{r})\}$, with $i = 1...n$. We aim to find the average transition path between A and B in the space of the CVs, z_i, and define the progress along it as a reaction coordinate. Provided that the CVs are sufficiently good descriptors of the system, the reaction coordinate is well-defined in terms of transition path theory [69] and the committor distribution [70, 71]. That is, along the path, we can determine the committor probability $p_B(\mathbf{r})$ that a trajectory starting with random Maxwell-Boltzmann distributed velocities arrives in state B before going through state A. As the system moves near the A or B basins in CV-space, $p_B(\mathbf{r})$ approaches respectively

0 or 1. In this CV-space it is possible to define an isocommittor surface comprising all points where $p_B(\mathbf{r}) = 0.5$. Furthermore, the isocommittor surfaces spanning all committor values from 0 to 1 provide a continuous foliation of CV-space from A to B. In these hyperplanes, we can define the transition flux density ρ as the number of trajectories going through the surface per unit area. This flux density peaks at the transition channel in the FES between A and B, resulting in the average transition path that we wish to localize.

To locate the average transition path, we make the following assumptions [66]:

1. The average transition path can be represented in CV-space by a path-CV: a curve defined by the vector function $\mathbf{P}(s) : \mathbb{R} \to \mathbb{R}^n$, where the parameter $s(\mathbf{z})|_{\mathbf{P}} : \mathbb{R}^n \to [0, 1]$ yields the progress along the path from A to B, such that $\mathbf{P}(0) \in A$ and $\mathbf{P}(1) \in B$. This quantity can in principle be connected to the committor value.

2. In the vicinity of the path, the isocommitor planes P_s are perpendicular to $\mathbf{P}(s)$.

3. In the vicinity of the path, the normalized transition flux density ρ can be represented by configurational probability $p(\mathbf{z}) = \exp(-F(\mathbf{z})/k_\mathrm{B}T)$, where F is the free energy, and k_B is the Boltzmann constant.

Given the first and second assumptions, it is possible to project any point $\mathbf{z}$ in CV-space onto its closest point on the path $\mathbf{P}(s)$, and derive the path progress parameter $s(\mathbf{z})$. Moreover, since we wish the curve $\mathbf{P}(s)$ to follow the transition channel of maximum flux density, we can take the third assumption and approximate the average transition path as:

$$\begin{aligned} \mathbf{P}(s) &= \int_{P_s} \mathbf{z}' p_s(z_1', ..., z_n') dP_s, \text{ with} \\ p(z_1', ..., z_n') &= \frac{1}{Z} \int e^{-\beta U(\mathbf{r})} \delta(z_1 - z_1') ... \delta(z_n - z_n') d\mathbf{r}, \end{aligned} \tag{2.32}$$

where $p_s(z_1', ..., z_n')$ is the flux probability density at the isocommitor surface perpendicular to $\mathbf{P}(s)$, with the Dirac delta function δ and the partition function Z.

Finding the average transition path

In principle, Equation 2.32 enables the calculation of the average transition path by sampling and making a histogram of $\mathbf{z}$ during an MD run, or a Monte Carlo simulation. However, the transition is a rare event on the time-scales accessible to standard simulations. The flux density away from neighborhood of the A or the B basins will be poorly sampled. This problem is typically overcome with enhanced sampling methods—e.g., metadynamics—by biasing directly the dynamics of the CVs, $\mathbf{z}$. But in practice, such an approach is limited to low-dimensional CV-spaces. The path-CV aims to overcome this limitation.

In our method, path-metadynamics (PMD) [66, 67], we exert a time-dependent metadynamics bias,

$$V_{\mathrm{bias}}(s_{\mathrm{g}}, t) = \sum_t H(t) \exp\left(\frac{-(s_{\mathrm{g}} - \tilde{s}_{\mathrm{g}}(t))^2}{2W^2} \right), \tag{2.33}$$

onto the one-dimensional path progress parameter, $s(t)$, along a guess transition path $s_{\mathrm{g}} = s(\mathbf{z})|_{\mathbf{P}_{\mathrm{g}}}$, with Gaussian height H and width W. This growing repulsive potential drives the system back and forth along the path, away from already visited configurations. If the path remains fixed, the metadynamics bias potential converges, eventually, to (minus) the free energy along the guess path $F_{\mathrm{g}} = -V_{\mathrm{bias}}(s_{\mathrm{g}}, t)$ [41].

To improve our guess path and use it to locate the average transition path, we replace in Equation 2.32 the ensemble average of transition points through the hyperplanes by a time average:

$$\langle \mathbf{z} \rangle_{s_{\mathrm{g}}} = \lim_{t \to \infty} \int_0^t \int_{P_{s_{\mathrm{g}}}} \mathbf{z}(t') dP_{s_{\mathrm{g}}} \, dt'. \tag{2.34}$$

We can now optimize the path toward the average transition path by iteratively relocating the guess path to the cumulative average density $\mathbf{P}_{\mathrm{g}}(s_{\mathrm{g}}) = \langle \mathbf{z} \rangle_{s_{\mathrm{g}}}$ (Figure 2.2). Simultaneously, the metadynamics algorithm will adapt the bias-potential after each path update as it keeps adding layers of Gaussian potentials to the total bias, overwriting the free energy at previous trial paths [72], and continuously converging toward the free energy at the average transition path[2].

Most of the extensive machinery developed in the last years for metadynamics can be applied directly with the PMD method. For example, the *well-tempered* method [73, 74] can aid in converging a free-energy profile by gradually reducing the height of the Gaussians on the fly[3]. The same effect can be obtained by manually reducing the Gaussian height at every recrossing from A to B [66]. Another

[2] In a PMD simulation, the parameters for the Gaussian deposition pace and the path update pace should be somewhat balanced for optimal efficiency. Setting a high metadynamics deposit frequency with a low path update frequency leads to many recrossings before finding an optimum path, while at the same time the initial crossings may take place at high-energy states outside the intrinsic transition valley. On the other hand, if one sets a slow metadynamics deposit pace and a high path update frequency, it may take a long time for the barrier crossing to occur, but it will most likely take place over paths that are already close to optimal. Note that the effect of increasing the metadynamics deposit frequency is generally the same as increasing the size of the Gaussian potentials. A sensible initial set up is to have the same pace for the metadynamics deposit frequency and for the path updates; the Gaussian height can be set to around two orders of magnitude less that the expected barrier (generally smaller than $k_{\mathrm{B}}T$ to allow for self-healing) and the width to around 0.05 normalized path units. Then, the path flexibility can be controlled using the half-life parameter.

[3] It is common practice in metadynamics to gradually reduce the Gaussian height to converge a free-energy profile. This can be done manually after each recrossing on the path, or via the well-tempered method. When using the latter option, one should use a somewhat larger bias factor than advertised for "normal" (non-adaptive) CVs, because path optimization requires generally the sampling of several barrier crossings, before which the Gaussian height should not have already been reduced too much. In our experience, a well-tempered bias factor of around 10 to 15 times the height of the barrier is a working choice. Another very interesting option is to use transition-tempered metadynamics (TTMetaD), a method that first fills the valleys in a non-tempered phase, and then converges the free-energy profile in a well-tempered fashion [75].

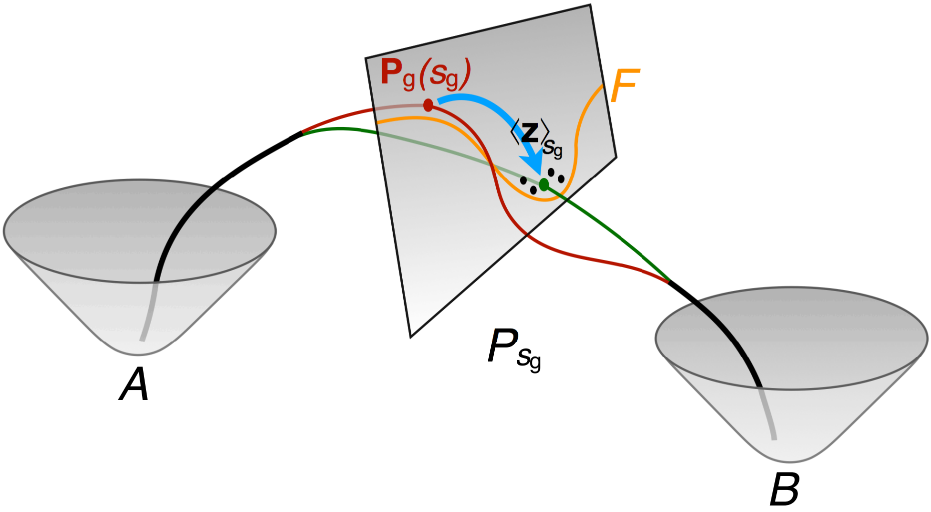

Figure 2.2: Graphical representation of an initial guess path-CV section (red) converging to the average transition path (green) between basins A and B. The curve points $\mathbf{P}_{\text{g}}(s_{\text{g}})$ are relocated to the cumulative average density $\langle \mathbf{z} \rangle_{s_{\text{g}}}$ in CV-space, which peaks at the valley in the free energy F (yellow) at each hyperplane $P_{s_{\text{g}}}$.

interesting feature is *multiple-walker* metadynamics [76], which can be applied with the path-CV to speed up the simulation. Here, several replicas of the system are simulated simultaneously in parallel for the exploration of different regions of CV-space, while each replica communicates its updates on both the path and on the bias potential to the other replicas. By initializing walkers both in the A and B basins, the shape of the path can be rendered significantly faster[4]. In Chapter 4, we apply multiple-walker PMD to conformational changes of oligopeptides.

The presented recipe to locate the average transition path with an adaptive path-CV is not exclusively coupled to metadynamics. Other enhanced sampling methods, such as steered MD, constrained MD, US, or even TPS, can be used with the path-CV, without changes to the path formalism. This flexibility of the path-CV is exemplified in Chapter 6.

Projection of CV-space onto a string of nodes

In order to implement the method numerically, we must provide a discrete definition of the path-CV as a parametrized curve. This is done by representing the curve as a

[4] When using multiple-walker metadynamics, it is always recommended to keep Gaussians narrow and small. Otherwise, we risk the walkers not actually exploring the underlying free-energy profile, but only feeling each other's repulsive potentials. This can be assessed by analyzing the diffusion of s over time for all walkers. There should indeed be some repulsion, but also crossings between walkers.

string of M ordered nodes[5], $\mathbf{P}_{\rm g}(s_{\rm g},t) \rightarrow \mathbf{P}_j^{t_i}$, with $j = 1, 2, ..., M$ labeling the nodes on the string and t_i representing the discrete time parameter at path update step i. Then, the projection of a point in CV-space, $\mathbf{z}$, onto the path—which yields the value of the path progress parameter s—is given by:

$$\begin{aligned}
s_{\rm g}(\mathbf{z}) &= \frac{m}{M} \pm \frac{\sqrt{(\mathbf{v}_1\cdot\mathbf{v}_3)^2 - |\mathbf{v}_3|^2(|\mathbf{v}_1|^2 - |\mathbf{v}_2|^2)} - (\mathbf{v}_1\cdot\mathbf{v}_3) + |\mathbf{v}_3|^2}{2M|\mathbf{v}_3|^2}, \text{ with}\\
\mathbf{v}_1 &= \mathbf{P}_m - \langle\mathbf{z}\rangle,\\
\mathbf{v}_2 &= \langle\mathbf{z}\rangle - \mathbf{P}_{m-1},\\
\mathbf{v}_3 &= \mathbf{P}_{m+1} - \mathbf{s}_m
\end{aligned} \tag{2.35}$$

where $\mathbf{P}_m$ is the closest path node to $\mathbf{z}$, and $\mathbf{P}_{m-1}$ and $\mathbf{P}_{m+1}$ are its neighboring nodes. This expression implies that points beyond the first or last nodes are mapped to values of $s_{\rm g} < 0$ and $s_{\rm g} > 1$ respectively. To have control over this mapping at and outside the stable states, extra trailing nodes can be added at both ends of the original path[6]. If necessary, wall potentials can be added to restrict the sampling on a particular $s_{\rm g}$-region. Note also that the projection in Equation 2.35 requires that the nodes are equidistant. This requirement is imposed by a reparametrization step [43] after each path update.

Evolution of the path-CV

The path update step, which sets $\mathbf{P}_{\rm g}(s_{\rm g}) = \langle\mathbf{z}\rangle_{s_{\rm g}}$, uses the time averaged distance between the sampled $\mathbf{z}$-points and their projected points on the path, $\mathbf{P}_{\rm g}(s_{\rm g}(\mathbf{z}))$. This distance is weighted by a weight, w, which is only non-zero for the two closest nodes, giving the following path node propagation equation:

[5] Determining a good number of nodes to capture a transition is a trial-and-error procedure. As a rule of thumb, one can start with a small number of transition nodes (20 to 30). If the resulting curve is able to capture all CV fluctuations of the transition, then one has succeeded. Otherwise, one can gradually add more nodes until all features of the transition are represented. In general, the path is resilient to changes in this parameter. We have increased the number of transition nodes up to 40% without affecting the final result. However, when too many nodes are added, the path tends to coil or loop around the stable states, and oscillate excessively around small fluctuations of the CVs [67].

[6] Typically, we set a small number of additional trailing nodes (10 to 20), as we require just enough of them to capture the valleys at both ends of the path. However, sometimes the trailing nodes can be exploited in other clever ways. For example, one can intentionally direct them to steep regions in the free-energy landscape and get a natural wall effect to restrain the sampling. This works as long as the trailing nodes are not relocated. Alternatively, one can have the trailing nodes probe a secondary relevant transition channel. We have performed simulations on other systems, in which the transition nodes capture the optimal path, while the trailing nodes fall into the second optimal path (although both ends do not touch and therefore that second path is not fully captured). In these cases, care must be taken that the trailing nodes do not approach the primary channel of the transition nodes. If this occurs, points in CV-space lying close to both sets of nodes can be suddenly mapped from one s value to the other, leading to ill-defined sampling. In some calculations where the sampling does not go beyond the stable basins (as we saw in the application of steered MD), trailing nodes are not needed.

$$\mathbf{P}_j^{t_{i+1}} = \mathbf{P}_j^{t_i} + \frac{\sum_k w_{j,k} \cdot (\mathbf{z}_k - \mathbf{P}^{t_i}(s(\mathbf{z}_k)))}{\sum_k w_{j,k}}, \text{ with}$$
$$w_{j,k} = \max\left[0, \left(1 - \frac{||\mathbf{P}_j^{t_i} - \mathbf{P}^{t_i}(s(\mathbf{z}_k))||}{||\mathbf{P}_j^{t_i} - \mathbf{P}_{j+1}^{t_i}||}\right)\right] \tag{2.36}$$

where k is the current MD step and $\Delta t = t_{i+1} - t_i$ is the time interval between two path updates. See Figure 2.3 for an illustration of the path update calculation. In order to slow down or accelerate the convergence of the path, an additional fade factor, $\xi = \exp(-\ln(2)/\tau)$, can be introduced, with a half-life parameter, τ, being the number of MD steps after which a distance measured from the path contributes only 50% of its original value to the average[7]. We reformulate:

$$\mathbf{P}_j^{t_{i+1}} = \mathbf{P}_j^{t_i} + \frac{\sum_k w_{j,k} \cdot (\mathbf{z}_k - \mathbf{P}^{t_i}(s(\mathbf{z}_k)))}{\sum_k \xi^{t_i - k} w_{j,k}} \tag{2.37}$$

Tuning the algorithm

When facing landscapes with multiple or ill-defined transition valleys, it can be beneficial to not only bias the sampling along the path, but also restrain the sampling to the path vicinity. A harmonic restraint potential on the perpendicular distance from the path, $z = ||\mathbf{z}_k - \mathbf{P}^{t_i}(s(\mathbf{z}_k))||$, set either at zero distance or allowing some freedom, can help in converging a transition. We refer to this restraint as a “tube” potential. In the limit of an infinitesimally narrow tube potential, the path is optimized by following the local free-energy gradient—in a similar fashion to the string method [43]—and PMD converges to the MFEP closest to the initial guess path instead of to the average transition path. Thus, the tube potential allows to control the behavior of PMD by switching between a path optimization based on CV density, quick for well-defined landscapes with a single channel, and a path optimization based on the free-energy gradient, suitable for more complex scenarios with multiple channels. This versatility is key to our adaptive path framework, especially when considering that the essential distinction between different path-CV implementations is the optimization rule [43, 48, 49, 66, 67, 77, 78]. Of course, when using a tube potential, care has to be taken regarding its effect on the entropic contribution to the free energy[8].

[7] There is no satisfactory default value for the half-life parameter. The general rule of thumb is to start with a relatively short half-life if the initial guess path is likely to be outside the intrinsic transition valley, and then switch to a large value, possibly even to infinite, once the (neighborhood of the) intrinsic transition valley is found. However, a too small half-life may yield a very flexible and dynamical evolution, which leads to curvy paths that do not guide the biased system over the transition barrier. With a too large half-life, the path evolves evermore slowly during the simulation, as each newly sampled transition density weights less due to the ever-growing history of previous samples. To check whether the path has stopped evolving because of the large half-life, or because it has actually found the transition valley, one should analyze the time evolution of the node weights from the path output files. If very large weight values appear early on in the simulation, a shorter half-life is probably needed.

[8] The tube potential is a convenient handle to temper a too flexible path evolution and to restrict the sampling to a specific transition pathway by blocking bifurcations. However, it should be noted

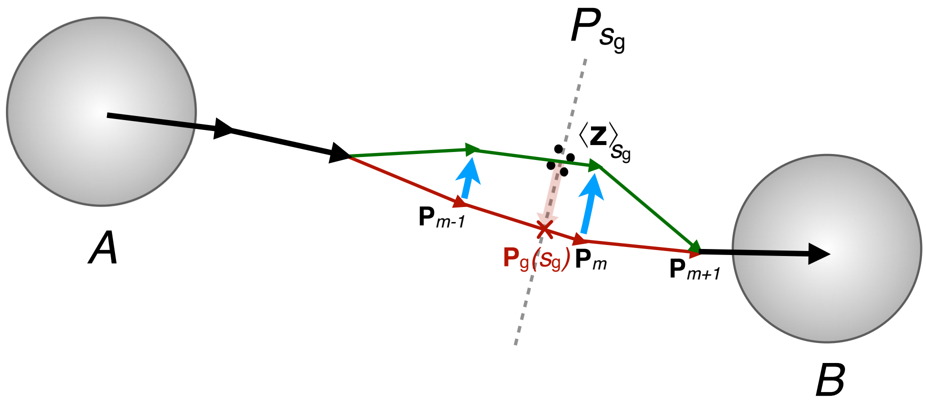
P_{s_g}
$\langle \mathbf{z} \rangle_{s_\mathrm{g}}$
A
$\mathbf{P}_{m\text{-}1}$
$\mathbf{P}_\mathrm{g}(s_\mathrm{g})$
$\mathbf{P}_m$
$\mathbf{P}_{m+1}$
B

Figure 2.3: Graphical explanation of a path node update. The sampled average density $\langle \mathbf{z} \rangle_{\sigma_g}$ is projected onto the path at $\mathbf{P}_\mathrm{g}(s_\mathrm{g})$. The two closest path nodes $\mathbf{s}_{m-1}$ and $\mathbf{s}_m$ are relocated according to weights that depend on their distances from $\mathbf{P}_\mathrm{g}(s_\mathrm{g})$. A subsequent reparametrization step redistributes the nodes along the path to make them equidistant again.

Another useful algorithmic extension is the scaling of CV-space. Imagine a set of CVs with numerical ranges differing several orders of magnitude. In order to keep an equidistant set of nodes under these conditions, the node distribution across dimensions would need to be severely unbalanced. As a consequence, the path progress parameter s would also be defined mostly by the most widely-ranging CV. To avoid this imbalance, one may rescale the CVs in a manner that the space to be sampled is in all dimensions normalized to one[9]. This is particularly helpful when dealing with CVs of different units (e.g., rad and deg) or dimensions (e.g., rad and nm). To rescale the CV-space, it is useful to have *a priori* knowledge of the minimum and maximum values that each CV can have.

A final remark on the algorithm concerns a side effect of the reparametrization step to ensure node equidistance. The implemented reparametrization algorithm [43] turns out to favor straight paths somewhat and displays a tendency to "cut corners" while redistributing the nodes. While this tendency is often beneficial, as it maintains a smooth, non-curling or self-intersecting curve, there are obvious drawbacks. In particular, when the metadynamics is temporarily sampling one end of the path, the repeating reparametrization after every path update redistributes

that, as long as the path is not yet optimal, the tube potential acts as an addition hurdle, forcing the system to cross outside the intrinsic transition valley. Secondly, after the path optimization is converged, the tube potential affects the sampling of the degrees of freedom orthogonal to the path, thus affecting the entropic contribution to the profile. This biasing by a tube potential can be relaxed somewhat by setting the harmonic wall at a non-zero value of z. In this spirit, it can be convenient to first measure the widths of the stable state basins, and then use this to set a tube potential that does not affect the stable state valleys.

[9] Scaling of CV-space can be done once the range of each CV during the transition is known. We simply calculate each scaling factor as $1/(z_{\mathrm{i,max}} - z_{\mathrm{i,min}})$ for each CV, z_i.

also the nodes at the other end of the path, moving them gradually back to a straight path and thus undoing previous path optimization. This side effect is much reduced in the multiple-walker PMD implementation, because in that case the sampling is more continuous along the entire path. Apart from preventing the information loss, of course the multiple-walker option also results in an almost trivial parallelization speedup for the sampling of the path and the free energy, thus providing a powerful extension to the original method [67][10].

In summary, the path-CV consists of a set of ordered nodes describing the transition from basin A to basin B in the high-dimensional CV-space. The system can be biased to move along the path, while the positions of the nodes in CV-space can be optimized by following the average density of the sampled points, which peaks at the free-energy valley. By means of this sampling along and around the path, we can converge the average transition path and the free energy along it. Additional actions can be taken to control the extent of the sampling and the flexibility of the path when facing challenging, forking free-energy landscapes. Namely, we can add a tube potential to restrain the sampling in the direction perpendicular to the path and switch from a density-based optimization toward the average transition path, to a gradient-based path optimization toward the closest MFEP to the initial guess path.

Path-CV implementation in PLUMED

The theoretical and numerical framework discussed above has been implemented into the PLUMED software [53] as a function of CVs. Invoking the action `PATHCV` in PLUMED requires that the following keywords are specified:

- `LABEL`: sets the identifier for this instance.
- `ARG`: sets the list of (priorly defined) CVs that span the space in which the path exists.
- `GENPATH`: generates a straight path between two points in CV-space; the two points typically marking the stable states. It takes 3 integers as arguments, corresponding to the number of anterior trailing nodes, actual transition nodes and posterior trailing nodes, followed by the CV-space coordinates of the initial and final transition nodes separated by commas.
- `INFILE`: points to a file containing an input path.
- `FIXED`: indicates the two fixed nodes corresponding to the initial and final states. The default values are the first and last nodes, thus assuming no trailing nodes.

[10] One can use multiple walkers to continuously explore all regions of the path and avoid the reparametrization step from undoing the node optimization in temporarily unsampled regions. When doing this, it is recommended to have at least as many walkers as expected stable and metastable states along the path. Another very interesting way to use multiple walkers is to include a special walker—which updates the path, but not the metadynamics bias—steered to a particular point in CV-space. Thus, we can find the optimal path which crosses that region (see Chapter 5).

- `OUTFILE (PATH)`: points to a file where the path updates are printed, concatenated one after the other.
- `SCALE`: lists the scaling factors to normalize the CV-space. The default value is one for each CV.
- `STRIDE (PACE)`: indicates the frequency for printing the path in MD steps.
- `PACE (0)`: indicates the frequency for optimizing the path nodes in MD steps.
- `HALFLIFE (-1)`: indicates the number of MD steps after which a previously measured path distance weighs only 50% in the average. A negative number sets it to infinity.

The parenthesized arguments indicate the default values for the keywords when relevant. The format of the `INFILE` and `OUTFILE` comprehends a first column with the numbering of the nodes, followed by N columns with the value of each CV at each node position, which in turn are followed by N columns showing the cumulative measured displacement of the system from the given node along each CV. The final column contains the cumulative weight $w_{j,k}$ for the corresponding node, such that the cumulative displacement divided by the cumulative weight gives the average distance between the path and the measured average transition density. After each path update, the cumulative displacement is reset to zero for the non-fixed nodes, but the cumulative weights remain.

To use the multiple-walker implementation, one should also provide:

- `WALKERS_ID`: indicates the ID of the current walker, starting from zero.
- `WALKERS_N`: indicates the total number of walkers.
- `WALKERS_DIR`: points to the directory where all walkers write and read each other's files.
- `WALKERS_RSTRIDE`: indicates the reading frequency for walkers in MD steps.

Two quantities can be extracted, and biased, from the path-CV: the components `s` and `z`, corresponding to the progress along the path, $s(\mathbf{z}_k)|_{\mathbf{P}^{t_i}}$, and the displacement from the path, $z = ||\mathbf{z}_k - \mathbf{P}^{t_i}(s(\mathbf{z}_k))||$. In the PLUMED syntax, the components are called by `LABEL.s` and `LABEL.z` respectively.

References

[1] D. Frenkel and B. Smit, *Understanding Molecular Simulations: From algorithms to applications* (Academic Press, 2002).

[2] L. Boltzmann, *Über die Beziehung zwischen dem zweiten Hauptsatze des mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung, respective den Sätzen über das Wärmegleichgewicht* (Sitzungberichte der Kaiserlichen Akademie der Wissenschaften. Mathematisch-Naturwissen Classe., 1877).

[3] H. Helmholtz, *Wissenschaftliche Abhandlungen* (J.A. Barth, 1882).

[4] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *Equation of state calculations by fast computing machines,* J. Chem. Phys. **21**, 1087 (1953).

[5] L. Boltzmann, *Vorlesungen über Gastheorie: 2. Teil* (J.B. Barth, Leipzig, Germany, 1898).

[6] B. J. Alder and T. E. Wainwright, *Studies in molecular dynamics. I. General method,* J. Chem. Phys. **31**, 459 (1959).

[7] I. Newton, *Philosophiae naturalis principia mathematica* (typis A. et J.M. Duncan, 1833).

[8] L. Verlet, *Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules,* Phys. Rev. **159**, 98 (1967).

[9] W. C. Swope, H. C. Andersen, P. H. Berens, and K. R. Wilson, *A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters,* J. Chem. Phys. **76**, 637 (1982).

[10] R. W. Hockney and J. W. Eastwood, *Computer simulation using particles* (crc Press, 1988).

[11] G. Bussi, D. Donadio, and M. Parrinello, *Canonical sampling through velocity rescaling,* J. Chem. Phys. **126**, 014101 (2007).

[12] M. Parrinello and A. Rahman, *Polymorphic transitions in single crystals: A new molecular dynamics method,* J. Appl. Phys. **52**, 7182 (1981).

[13] C. A. de Coulomb, *Premier-troisième mémoire sur l'electricité et le magnétisme* (Académie Royale des sciences, 1785).

[14] J. D. van der Waals, *Over de continuiteit van den gas- en vloeistoftoestand,* Ph.D. thesis, Universiteit Leiden (1873).

[15] J. E. Lennard-Jones, *On the determination of molecular fields. II. From the equation of state of gas,* Proc. R. Soc. Lond. A. **106**, 463 (1924).

[16] T. Darden, D. York, and L. Pedersen, *Particle mesh Ewald: An NLog(N) method for Ewald sums in large systems,* J. Chem. Phys. **98**, 10089 (1993).

[17] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, *A smooth particle mesh Ewald method,* J. Chem. Phys. **103**, 8577 (1995).

[18] H. J. C. Berendsen, D. van der Spoel, and R. van Drunen, *GROMACS: a message-passing parallel molecular dynamics implementation,* Comput. Phys. Commun. **91**, 43 (1995).

[19] Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang, and P. Kollman, *A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations,* J. Comput. Chem. **24**, 1999 (2003).

[20] A. D. MacKerell Jr, N. Banavali, and N. Foloppe, *Development and current status of the CHARMM force field for nucleic acids,* Biopolymers **56**, 257 (2000).

[21] I. Ivani, P. D. Dans, A. Noy, A. Pérez, I. Faustino, J. Walther, P. Andrio, R. Goñi, A. Balaceanu, G. Portella, *et al.*, *Parmbsc1: a refined force field for DNA simulations,* Nat. Methods **13**, 55 (2016).

[22] K. N. Kirschner, A. B. Yongye, S. M. Tschampel, J. González-Outeiriño, C. R. Daniels, B. L. Foley, and R. J. Woods, *GLYCAM06: a generalizable biomolecular force field. Carbohydrates,* J. Comput. Chem. **29**, 622 (2008).

[23] S. Plimpton, *Fast parallel algorithms for short-range molecular dynamics,* J. Comput. Phys. **117**, 1 (1995).

[24] F. Jensen, *Introduction to Computational Chemistry*, 2nd ed. (John Wiley & Sons, Ltd, 2007).

[25] C. J. Cramer, *Essentials of Computational Chemistry. Theories and Models*, 2nd ed. (John Wiley & Sons, Ltd, 2004).

[26] E. Schrödinger, *An undulatory theory of the mechanics of atoms and molecules,* Phys. Rev. **28**, 1049 (1926).

[27] M. Born and R. Oppenheimer, *Zur quantentheorie der molekeln,* Ann. Phys. (Berl.) **389**, 457 (1927).

[28] P. Hohenberg and W. Kohn, *Inhomogeneous electron gas,* Phys. Rev. **136**, B864 (1964).

[29] W. Kohn and L. J. Sham, *Self-consistent equations including exchange and correlation effects,* Phys. Rev. **140**, A1133 (1965).

[30] J. VandeVondele, M. Krack, F. Mohamed, M. Parrinello, T. Chassaing, and J. Hutter, *Quickstep: Fast and accurate density functional calculations using a mixed gaussian and plane waves approach,* Comput. Phys. Comm. **167**, 103 (2005).

[31] J. Hutter, M. Iannuzzi, F. Schiffmann, and J. VandeVondele, *cp2k: atomistic simulations of condensed matter systems,* Wiley Interdiscip. Rev. Comput. Mol. Sci. **4**, 15 (2014).

[32] T. D. Kühne, M. Iannuzzi, M. Del Ben, V. V. Rybkin, P. Seewald, F. Stein, T. Laino, R. Z. Khaliullin, O. Schütt, F. Schiffmann, *et al.*, *CP2K: An electronic structure and molecular dynamics software package-Quickstep: Efficient and accurate electronic structure calculations,* J. Chem. Phys. **152**, 194103 (2020).

[33] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, *A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu,* J. Chem. Phys. **132**, 154104 (2010).

[34] A. Warshel and M. Levitt, *Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme,* J. Mol. Biol. **103**, 227 (1976).

[35] M. J. Field, P. A. Bash, and M. Karplus, *A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations,* J. Comput. Chem **11**, 700 (1990).

[36] H. M. Senn and W. Thiel, *QM/MM methods for biomolecular systems,* Angew. Chem. Int. Edit. **48**, 1198 (2009).

[37] G. M. Torrie and J. P. Valleau, *Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling,* J. Comput. Phys. **23**, 187 (1977).

[38] E. A. Carter, G. Ciccotti, J. T. Hynes, and R. Kapral, *Constrained reaction coordinate dynamics for the simulation of rare events,* Chem. Phys. Lett **156**, 472 (1989).

[39] W. K. den Otter and W. J. Briels, *The calculation of free-energy differences by constrained molecular dynamics simulations,* J. Chem. Phys. **109**, 4139 (1998).

[40] C. Jarzynski, *Nonequilibrium equality for free energy differences,* Phys. Rev. Lett **78**, 2690 (1997).

[41] A. Laio and M. Parrinello, *Escaping free-energy minima,* Proc. Natl. Acad. Sci. U.S.A. **99**, 12562 (2002).

[42] H. Jónsson, G. Mills, and K. W. Jacobsen, *Nudged elastic band method for finding minimum energy paths of transitions,* in *Classical and Quantum Dynamics in Condensed Phase Simulations*, edited by B. Berne, G. Ciccotti, and D. F. Coker (World Scientific, 1998) pp. 385–404.

[43] L. Maragliano, A. Fischer, E. Vanden-Eijnden, and G. Ciccotti, *String method in collective variables: Minimum free energy paths and isocommittor surfaces,* J. Chem. Phys. **125**, 024106 (2006).

[44] C. Dellago, P. G. Bolhuis, F. S. Csajka, and D. Chandler, *Transition path sampling and the calculation of rate constants,* J. Chem. Phys. **108**, 1964 (1998).

[45] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler, *Transition path sampling: Throwing ropes over rough mountain passes, in the dark,* Annu. Rev. Phys. Chem. **53**, 291 (2002).

[46] S. Park, M. K. Sener, D. Lu, and K. Schulten, *Reaction paths based on mean first-passage times,* J. Chem. Phys. **119**, 1313 (2003).

[47] P. Y. Ayala and H. B. Schlegel, *A combined method for determining reaction paths, minima, and transition state geometries,* J. Chem. Phys. **107**, 375 (1997).

[48] D. Branduardi, F. L. Gervasio, and M. Parrinello, *From A to B in free energy space,* J. Chem. Phys. **126**, 054103 (2007).

[49] A. C. Pan, D. Sezer, and B. Roux, *Finding transition pathways using the string method with swarms of trajectories,* J. Phys. Chem. B **112**, 3432 (2008).

[50] B. Berg and T. Neuhaus, *Multicanonical ensemble: A new approach to simulate first-order phase transitions,* Phys. Rev. Lett. **68**, 9 (1992).

[51] Y. Sugita and Y. Okamoto, *Replica-exchange molecular dynamics method for protein folding,* Chem. Phys. Lett. **314**, 141–151 (1999).

[52] L. Maragliano and E. Vanden-Eijnden, *A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations,* Chem. Phys. Lett. **426**, 168 (2006).

[53] G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni, and G. Bussi, *PLUMED 2: New feathers for an old bird,* Comput. Phys. Commun. **185**, 604 (2014).

[54] M. Bonomi, G. Bussi, C. Camilloni, G. A. Tribello, P. Banáš, A. Barducci, M. Bernetti, P. G. Bolhuis, S. Bottaro, D. Branduardi, *et al.*, *Promoting transparency and reproducibility in enhanced molecular simulations,* Nat. Methods **16**, 670 (2019).

[55] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, *The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method,* J. Comput. Chem. **13**, 1011 (1992).

[56] A. Grossfield, *WHAM: the weighted histogram analysis method, version 2.0.9,* http://membrane.urmc.rochester.edu/content/wham (2013).

[57] J. G. Kirkwood, *Statistical mechanics of fluid mixtures,* J. Chem. Phys. **3**, 300 (1935).

[58] J. Kästner and W. Thiel, *Bridging the gap between thermodynamic integration and umbrella sampling provides a novel analysis method: "Umbrella integration",* J. Chem. Phys. **123**, 144104 (2005).

[59] H. Grubmüller, B. Heymann, and P. Tavan, *Ligand binding: molecular mechanics calculation of the streptavidin-biotin rupture force,* Science **271**, 997 (1996).

[60] G. Bussi and A. Laio, *Using metadynamics to explore complex free-energy landscapes,* Nat. Rev. Phys. **2**, 1 (2020).

[61] T. S. Van Erp, D. Moroni, and P. G. Bolhuis, *A novel path sampling method for the calculation of rate constants,* J. Chem. Phys. **118**, 7762 (2003).

[62] A. K. Faradjian and R. Elber, *Computing time scales from reaction coordinates by milestoning,* J. Chem. Phys. **120**, 10880 (2004).

[63] E. Weinan, W. Ren, and E. Vanden-Eijnden, *String method for the study of rare events,* Phys. Rev. B **66**, 052301 (2002).

[64] E. Weinan, W. Ren, and E. Vanden-Eijnden, *Finite temperature string method for the study of rare events,* J. Phys. Chem. B **109**, 6688 (2005).

[65] E. Vanden-Eijnden and M. Venturoli, *Revisiting the finite temperature string method for the calculation of reaction tubes and free energies,* J. Chem. Phys. **130**, 194103 (2009).

[66] G. Díaz Leines and B. Ensing, *Path finding on high-dimensional free energy landscapes,* Phys. Rev. Lett **109**, 020601 (2012).

[67] A. Pérez de Alba Ortíz, A. Tiwari, R. Puthenkalathil, and B. Ensing, *Advances in enhanced sampling along adaptive paths of collective variables,* J. Chem. Phys. **149**, 072320 (2018).

[68] A. Pérez de Alba Ortíz, J. Vreede, and B. Ensing, *The adaptive path collective variable: a versatile biasing approach to compute the average transition path and free energy of molecular transitions,* in *Biomolecular Simulation*, edited by M. Bonomi and C. Camilloni (Springer, 2019) pp. 255–290.

[69] M. Ferrario, G. Ciccotti, and K. Binder, *Computer Simulations in Condensed Matter: from Materials to Chemical Biology* (Springer, 2007).

[70] L. Onsager, *Initial recombination of ions,* Phys. Rev. **54**, 554 (1938).

[71] P. G. Bolhuis, C. Dellago, and D. Chandler, *Reaction coordinates of biomolecular isomerization,* Proc. Natl. Acad. Sci. U.S.A. **97**, 5877 (2000).

[72] B. Ensing, A. Laio, M. Parrinello, and M. L. Klein, *A recipe for the computation of the free energy barrier and the lowest free energy path of concerted reactions,* J. Phys. Chem. B **109**, 6676 (2005).

[73] A. Barducci, G. Bussi, and M. Parrinello, *Well-tempered metadynamics: a smoothly converging and tunable free-energy method,* Phys. Rev. Lett **100**, 020603 (2008).

[74] M. Bonomi, A. Barducci, and M. Parrinello, *Reconstructing the equilibrium Boltzmann distribution from well-tempered metadynamics,* J. Comput. Chem. **30**, 1615 (2009).

[75] J. F. Dama, G. Rotskoff, M. Parrinello, and G. A. Voth, *Transition-tempered metadynamics: robust, convergent metadynamics via on-the-fly transition barrier estimation,* J. Chem. Theory Comput. **10**, 3626 (2014).

[76] P. Raiteri, A. Laio, F. L. Gervasio, C. Micheletti, and M. Parrinello, *Efficient reconstruction of complex free energy landscapes by multiple walkers metadynamics,* J. Phys. Chem. B **110**, 3533 (2006).

[77] G. A. Gallet, F. Pietrucci, and W. Andreoni, *Bridging static and dynamical descriptions of chemical reactions: An ab initio study of CO2 interacting with water molecules,* J. Chem. Theory Comput. **8**, 4029 (2012).

[78] F. Pietrucci and A. M. Saitta, *Formamide reaction network in gas phase and solution via a unified theoretical approach: Toward a reconciliation of different prebiotic scenarios,* Proc. Natl. Acad. Sci. U.S.A. **112**, 15030 (2015).