Improved Sampling Techniques for the Simulation of Biocatalytic Reaction Mechanisms

Andreas Valentin Hulm

München 2025

Dissertation zur Erlangung des Doktorgrades
der Fakultät für Chemie und Pharmazie
der Ludwig-Maximilians-Universität München

Improved Sampling Techniques for the Simulation of Biocatalytic Reaction Mechanisms

Andreas Valentin Hulm

aus

Freising

2025

Erklärung

Diese Dissertation wurde im Sinne von §7 der Promotionsordnung vom 28. November 2011 von Herrn Prof. Dr. Christian Ochsenfeld betreut.

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, 25.08.2025 _____

(Andreas Hulm)

Dissertation eingereicht am:	26.08.2025
1. Gutachter:	Prof. Dr. Christian Ochsenfeld
2. Gutachter:	Prof. Dr. Benjamin P. Fingerhut
Mündliche Prüfung am:	15.10.2025

Danksagung

An dieser Stelle möchte ich all jenen danken, durch die diese Arbeit möglich wurde. Ihr habt mich im Laufe der Jahre unterstützt, inspiriert und motiviert.

Zunächst möchte ich mich bei **Prof. Dr. Christian Ochsenfeld** dafür bedanken, dass ich meine Doktorarbeit in seinem Arbeitskreis anfertigen konnte. Ich danke ihm für die spannenden Projekte und die Unterstützung im Laufe der Jahre.

Des Weiteren bedanke ich mich bei **Prof. Dr. Benjamin Fingerhut** für die Anfertigung des Zweitgutachtens.

Ich danke dem gesamten **Arbeitskreis Ochsenfeld** für die herzliche Arbeitsatmosphäre, sowie viele inspirierende und lehrreiche Diskussionen.

Ein besonderer Dank gilt **Dr. Johannes Dietschreit**, der mir als Mentor den Einstieg in die Promotion leicht gemacht hat, **Dr. Beatriz von der Esch**, die mich mit ihrer Energie angesteckt und motiviert hat, **Yannick Lemke** für spannende Schachpartien, die den Arbeitsalltag bereichert haben, **Judit Katalin Szántó** und **Robert Schiller** für die großartige Zusammenarbeit, sowie **Alexandra Stan-Bernhardt** für die Freundschaft, die in vielen Gesprächen in Kaffeepausen entstanden ist.

Zuletzt das Wichtigste: die **Familie**, ohne die das Anfertigen einer Doktorarbeit zwar möglich, aber sinnlos wäre. Der größte Dank gilt **Katja**, die mich auf meinem Weg begleitet und immer unterstützt.

Abstract

In recent years, the dramatic improvement of computer hardware, as well as remarkable algorithmic advances, enabled accurate computer simulations of ever-growing molecular systems. Nowadays, with the help of composite quantum-mechanical/molecular-mechanical (QM/MM) methods, it is possible to model 'electronic' processes, like chemical reaction mechanisms that involve covalent bond rearrangements, in extended biological macromolecules with thousands of atoms. Such systems are frequently characterized by remarkable structural flexibility and potential energy surfaces full of local minima. Therefore, to enable the calculation of accurate ensemble properties, representative sampling of their configurational landscape is required, usually employing molecular dynamics (MD) simulations. However, it is crucial that the trajectories obtained are sufficiently long to resolve rare events, *i.e.*, phenomena that occur on macromolecular timescales and are beyond the reach of conventional all-atom molecular dynamics (MD) simulations. Hence, importance sampling techniques are applied to accelerate transitions between metastable states. In addition to accelerating sampling, such methods must provide accurate estimates of reaction free energy differences and kinetic rates of chemical transitions, such that the most likely reaction mechanisms can be found.

In this dissertation, several improvements of importance sampling techniques are presented. The focus lies on highly efficient algorithms, which are suitable for use together with an accurate QM/MM treatment of the electronic structure based on density functional theory (DFT), which still limits MD simulations to timescales of only a few hundred picoseconds due to their computational demand. The result is a versatile toolbox of sampling algorithms, which is made publicly available in the open-source `adaptive-sampling` Python package, that is highly useful for modeling intricate biocatalytic reaction mechanisms in explicit protein environments. This is demonstrated by its application to the simulation of challenging enzymatic systems that catalyze intricate biochemical transitions, such as ribonucleic acid (RNA) modification, adenosine triphosphate (ATP) hydrolysis, or long-range biological protonation dynamics.

List of Publications

This is a cumulative dissertation, comprising five articles in peer-reviewed journals. Their complete contents and corresponding supporting information can be found in Chapter 3. In the following, the articles are listed together with the author's contribution to each of them, followed by a list of further publications.

I **A. Hulm**; J. C. B. Dietschreit; C. Ochsenfeld. "Statistically Optimal Analysis of the Extended-System Adaptive Biasing Force (eABF) Method."
J. Chem. Phys. **2022**, 157, 024110.

Contribution by the Author: *Conception, implementation of the algorithms, performing and analyzing the simulations, and writing the manuscript.*

II **A. Hulm**; C. Ochsenfeld. "Improved Sampling of Adaptive Path Collective Variables by Stabilized Extended-System Dynamics."
J. Chem. Theory Comput. **2023**, 19, 9202-9210.

Contribution by the Author: *Conception, implementation of the algorithms, performing and analyzing the simulations, and writing the manuscript.*

III J. K. Szántó; A. Hulm; C. Ochsenfeld. "Molecular Mechanism of ATP Hydrolysis Catalyzed by p97: a QM/MM Study"
J. Chem. Theory Comput. **2025**, 21, 19, 9459–9469.

Contribution by the Author: *Participation in the conception, implementation of the algorithms, scientific consulting in performing and analyzing the simulations, and participation in writing the manuscript.*

IV M. C. Pöverlein; **A. Hulm**; J. C. B. Dietschreit; J. Kussmann; C. Ochsenfeld; V. R. I. Kaila. "QM/MM Free Energy Calculations of Long-Range Biological Protonation Dynamics by Adaptive and Focused Sampling."
J. Chem. Theory Comput. **2024**, 20, 13, 5751–5762.

Contribution by the Author: *Participation in the conception, implementation of the algorithms, performing and analyzing simulations jointly with M. C. Pöverlein, and participation in writing the manuscript.*

V **A. Hulm**; R. P. Schiller; C. Ochsenfeld. "Combining Fast Exploration with Accurate Reweighting in the OPES-eABF Hybrid Sampling Method."
J. Chem. Theory Comput. **2025**, 21, 6434–6445.

Contribution by the Author: *Conception, implementation of the algorithms jointly with Robert P. Schiller, performing and analyzing the simulations, and writing the manuscript.*

Further publications:

VI J. C. B. Dietschreit; D. J. Diestler; **A. Hulm**; C. Ochsenfeld; R. Gómez-Bombarelli. "From Free-Energy Profiles to Activation Free Energies."
J. Chem. Phys. **2022**, 157, 084113.

VII A. Stan-Bernhardt; L. Glinkina; **A. Hulm**; C. Ochsenfeld. "Exploring Chemical Space Using Ab Initio Hyperreactor Dynamics."
ACS Cent. Sci. **2024**, 10, 2, 302–314.

Table of Contents

Chapter 1

Introduction

In recent years, biomolecular simulations have become an invaluable tool for the understanding of biological systems. Their worth is demonstrated by significant contributions to diverse fields such as drug discovery and development [1–4], biocatalysis [5–7], biotechnology [8, 9], nanotechnology [10], and chemical biology [11]. In all these research domains, physics-based simulations serve to complement experimental methods, thereby facilitating a robust understanding of biological processes at the molecular level.

Due to the sheer size and structural flexibility of macromolecules in both the protein and nucleic acid worlds, for reliable in-silico predictions it is essential to shift the paradigm from studying single structures to analyzing conformational ensembles [12]. For this purpose, over the years, molecular dynamics (MD) [13–15] simulations have emerged as a mature tool [16, 17], which generate time-dependent trajectories of atomic configurations. To make such simulations practical, several severe approximations are commonly employed. For example, for many biochemical problems, classical molecular mechanics (MM) is a reasonable choice, assuming that molecules always remain in the instantaneous electronic ground state [18]. Hence, molecular motion is viewed as dynamics on a static potential energy surface (PES). It is this most powerful idea that makes the simulation of extended biological macromolecules feasible by empirically approximating the PES with simple mathematical functions (*i.e.*, force fields) that model atomic interactions and fully replace electronic degrees of freedom, like for example demonstrated in the pioneering AMBER protein force fields [19, 20].

However, the accurate description of processes that involve significant electronic rearrangements still requires quantum mechanics (QM). Examples are chemical reactions that involve covalent bond breaking and/or formation. Fortunately, such processes are usually rather local, such that it is often sufficient to use QM/MM hybrid schemes [21], for the development of which Karplus, Levitt, and Warshel share a Nobel prize [22]. Therefore, only the reaction center of moderate size must be investigated by using expensive QM methods, while the much larger rest of the system can be modeled using less expensive empirical force fields. In recent years, it has become possible to perform density functional theory (DFT) based QM/MM MD trajectories spanning hundreds of picoseconds thanks

to algorithmic advances in the Ochsenfeld [23–28] and other groups [29–33], *e.g.*, employing density fitting approximations and seminumerical integration techniques together with massively parallel implementations to harness the power of modern computer hardware optimally.

Despite these successes, the brute force MD simulation of chemical transitions that cross high barriers and, hence, occur on macroscopic time scales remains elusive. Therefore, it is key to accelerate target processes in MD simulations using importance sampling techniques, which are the main focus of this work. The most popular class of methods relies on the definition of collective variables (CVs) (*i.e.*, reaction coordinates), which provide a low-dimensional description of the given process, focusing the sampling on the relevant parts of configuration space. Popular examples are Umbrella Sampling (US) [34, 35], (Well-tempered) metadynamics (WTM/MtD) [36, 37] or the adaptive-biasing force (ABF) method [38, 39]. In this thesis, extended-Lagrangian techniques [40, 41] like the extended-system ABF (eABF [42, 43]) are applied, which, instead of applying a bias directly to CVs, couple the CV to a fictitious particle that is biased vicariously. The induced algorithmic flexibility is harnessed by highly effective hybrid methods, like the WTM-eABF proposed by Fu *et al.* [44, 45], which can further be combined with Gaussian accelerated MD (GaMD) [46] to reduce the CV dependence [47]. Generally, all these methods aim for the characterization of target processes in terms of their reaction and activation free energy by calculating the associated potential of mean force (PMF) (*i.e.*, free energy surface).

This cumulative dissertation is based on five publications, which are summarized below and provided in full in Chapter 3. Furthermore, a thorough theoretical discussion on the sampling problem is given in Chapter 2, putting the presented works into perspective, while Chapter 4 concludes this work and gives a brief outlook..

Publication **I** builds the foundation for the later works. Its goal is to show that extended-system-based importance sampling methods (*e.g.*, eABF, WTM-eABF, or GaWTM-eABF) can be combined with the multistate Bennett acceptance ratio (MBAR) [48] estimator, which enables the recovery of unbiased statistical weights of individual simulation frames. For the example of nuclear magnetic resonance (NMR) shieldings, it is demonstrated how this extends the application of these methods from the computation of the PMF alone to the calculation of ensemble averages of arbitrary properties. Additionally, it is shown how PMF reweighting is enabled, *i.e.*, mapping the probability density on different reaction coordinates, a concept that the further publications will heavily rely on. Lastly, as part of this project, the open-source `adaptive-sampling` program package was initiated, which grew to a versatile toolbox for importance sampling simulations over the following years [49].

Publication **II** addresses one of the biggest practical challenges in the application of most importance sampling algorithms, which is their inherent dependency on the *a priori* choice of a suitable set of CVs that can capture all slow degrees of freedom of the target process. To mitigate this problem, we apply path CVs (PCVs) [50–52], which replace the choice of the CVs with the choice of a path of discrete nodes that connect the reactant and product states. Suitable guess paths can systematically be obtained from interpolation between metastable states or from minimum energy path optimization approaches like the

Nudged Elastic Band (NEB) method [53]. Additionally, the guess path can iteratively be converged to the minimum free energy path (MFEP) on-the-fly in an adaptive PCV formalism. With a new stabilization algorithm, we can combine extended-system-based sampling algorithms with adaptive PCVs. Further, we show that the WTM-eABF method, which is able to rapidly adapt to changes in the PCV, together with the MBAR estimator, which allows for reweighting of the whole simulation data to any path, is highly beneficial in combination with adaptive PCVs. The motivation for this project is the investigation of the reaction mechanism of pseudouridine synthases (PUS), where in the process of transforming the uridine C1'-N1 into a C1'-C5 bond, the unbound uridilate ion undergoes a non-linear rotating motion, which is hard to capture using conventional linear CVs. Using the proposed path WTM-eABF method, the first step of the PUS reaction mechanism was successfully simulated in line with the experimental evidence for a reaction that runs over a glycal intermediate.

In Publication **III**, a workflow including path WTM-eABF simulations is applied to study the reaction mechanism of adenosine triphosphate (ATP) hydrolysis by p97, a member of the AAA+ protein family. In these simulations, the catalytic role of highly conserved protein residues in the active site is deduced, providing a complete understanding of the catalytic process. Experimental results like kinetic turnover rates, cryo-EM structures of the final state, and phosphate NMR chemical shieldings are compared to calculated properties and show good agreement. Overall, this project highlights how the developed methods enable the in-depth understanding of catalysis in extended biological systems.

In Publication **IV**, the presented sampling approaches are taken to their limit on the highly challenging problem of biological long-range proton transfer (pT). These water-mediated processes are catalyzed by titrable amino acids and can cover distances of tens of Angstrom. Hence, the excess proton might undergo a high number of different protonation pathways involving all participating water molecules and amino acids. With the application of the modified center-of-excess charge (mCEC) [54] as a CV, which provides a global description of pT pathways, together with fast sampling of transitions using the WTM-eABF method, we aim for the on-the-fly exploration of new protonation pathways. At the same time, the efficient exploration of parallel pathways naturally prohibits the convergence of PMF estimates, as whenever new regions of configurational space are visited, the PMF changes. Therefore, it is important to be able to obtain sufficient sampling along protonation pathways that were only sparsely visited during the WTM-eABF simulation. Hence, we show that US windows can be added in such regions to obtain focused sampling. In contrast to performing the exploration using conventional MtD/WTM, both the WTM-eABF and US data are processed all in one using the MBAR estimator, such that no sampling data is lost. Under the line, a workflow is presented that, by combining global and local sampling approaches, provides a promising basis for the reliable simulation of highly challenging long-range pT processes.

Finally, Publication **V** suggests further improvements to the WTM-eABF method by replacing WTM with the more recent on-the-fly probability enhanced sampling (OPES) [55], which converges more rapidly. Additionally, an algorithm is devised to obtain suitable coupling widths for the extended system from short unbiased MDs, which is the most critical

parameter for all extended-system-based methods. We show that the resulting OPES-eABF converges faster than its predecessors WTM-eABF and OPES, requires minimal user input, and is highly robust. Three test cases are discussed, starting with the case of high reaction free energy barriers, for which we identify a weakness in OPES that is resolved by OPES-eABF. Further, results for sampling poor CVs and parallel reaction pathways are shown. Altogether, we show that OPES-eABF unites the benefits of its building blocks, providing a unified tool to mitigate diverse sampling problems.

Chapter 2

Theoretical Background

We begin this chapter with a brief introduction to the basics of statistical mechanics as they can be found in many standard textbooks [40, 56–58]. Particular emphasis is given to the connection of statistical mechanics to molecular dynamics (MD), which builds the theoretical foundation of this work. Later, an overview of the ideas behind importance sampling is given, followed by a thorough discussion of the practical problem of dimensionality reduction and a deep dive into adaptive importance sampling techniques. Throughout these sections, the new developments of this thesis are illuminated with a focus on the application to (bio)catalysis.

2.1 Molecular Dynamics and Statistical Mechanics

In this work, the classical Hamiltonian description of the nuclear motion will be applied. Assuming the adiabatic Born-Oppenheimer approximation holds, the lighter electrons adjust adiabatically to the nuclear motion, always remaining in their instantaneous ground state. Compact vectorial notations will be frequently employed for an N particle system in $\mathbb{R}^3$ with $N_c = 3N$ cartesian coordinates $\mathbf{x}$ and corresponding velocities $\dot{\mathbf{x}} = \mathbf{v}$.

With the diagonal mass matrix $\mathbf{M}$ and some potential energy function U the Hamiltonian is defined as the sum of kinetic and potential energies

$$\mathcal{H}(\mathbf{x}, \mathbf{p}) \coloneqq \frac{\mathbf{p}^T \mathbf{M}^{-1} \mathbf{p}}{2} + U(\mathbf{x}), \tag{2.1}$$

where $\mathbf{p} = \mathbf{M}\mathbf{v}$ is the conjugate momentum vector. Thus, the complete microscopic state, *i.e.*, microstate, of a system can be described by a point in *phase space* $\{\mathbf{x}, \mathbf{p}\}$, which is the set of all coordinates and momenta for which $\mathcal{H}(\mathbf{x}, \mathbf{p})$ is finite. We refer to the *configuration space* when only looking at the configurations of the system, ignoring the momenta. From the Hamiltonian, the dynamic equations

$$\begin{cases} \delta\mathbf{x} = \mathbf{M}^{-1}\mathbf{p}\,\delta t \\ \delta\mathbf{p} = -\frac{\partial U}{\partial \mathbf{x}}\,\delta t = -\nabla_{\mathbf{x}} U(\mathbf{x})\,\delta t \end{cases} \tag{2.2}$$

are obtained, which are numerically integrated in time with a discrete time step δt. Note that more sophisticated integration schemes are used in practice, which increase the accuracy and stability [56]. This is the foundation for molecular dynamics (MD) calculations, which is a means of investigating the dynamic motion of molecular systems.

From the above, given a potential energy function $U(\mathbf{x})$ and initial conditions, time trajectories of microstates can be obtained. In practice, however, one is instead often interested in macroscopic observables that can be regarded as the statistical average over a subset of microstates $\mathcal{S}$, *i.e.*, an *ensemble*, that is constrained by macroscopic variables such as volume V, pressure P, absolute temperature T, total energy E, or number of particles N. The statistical distribution of microstates associated with a system with constant N, V, and E is called *microcanonical ensemble*. Here, as the energy is conserved, all possible states have the same probability.

However, the most important statistical ensemble is the *canonical ensemble*, as it is relevant for most experimental setups. The canonical ensemble represents all possible microstates of a system that are in thermal equilibrium with a heat bath at a fixed temperature. Therefore, instead of the total energy, the temperature is conserved. Analogous, Eq. 2.2, which on its own would sample from the microcanonical ensemble, has to be modified to include a thermostat that models an external bath with target temperature T. One example of such isothermal dynamics is Langevin dynamics

$$\begin{cases} \delta\mathbf{x} = \mathbf{M}^{-1}\mathbf{p}\,\delta t \\ \delta\mathbf{p} = \left(-\nabla_{\mathbf{x}} U(\mathbf{x}) - \gamma\mathbf{p} + \sqrt{\frac{2\gamma\mathbf{M}}{\beta\,\delta t}}\mathbf{G} \right) \delta t\ , \end{cases} \tag{2.3}$$

where $\beta = 1/(k_{\mathrm{B}}T)$ is the inverse temperature and k_{B} the Boltzmann constant. Here, compared to Hamiltonian dynamics, in the second line forces are modified with a damping constant γ, also known as collision frequency, and a Gaussian-distributed stochastic vector $\mathbf{G}$ [59]. Note that Hamiltonian dynamics is exactly recovered for $\gamma = 0$, while fully stochastic Brownian dynamics is retrieved for $\gamma \to \infty$. In this work, underdamped Langevin dynamics is always employed, using small values of γ. Hence, by sampling from the canonical ensemble, the probability $\rho(\mathbf{x}_i, \mathbf{p}_i)$ of individual states i is Boltzmann distributed, and depends on the energy via

$$\rho(\mathbf{x}_i, \mathbf{p}_i) = \frac{e^{-\beta\mathcal{H}(\mathbf{x}_i,\mathbf{p}_i)}}{\sum_{j\in\mathcal{S}} e^{-\beta\mathcal{H}(\mathbf{x}_j,\mathbf{p}_j)}} = Q^{-1} e^{-\beta\mathcal{H}(\mathbf{x}_i,\mathbf{p}_i)} \tag{2.4}$$

where Q is called *canonical partition function*. It acts as a normalization constant such that

$$\sum_{i\in\mathcal{S}} \rho(\mathbf{x}_i, \mathbf{p}_i) = 1 \tag{2.5}$$

and is the central quantity of statistical thermodynamics. As an alternative to the sum over all accessible microstates, one can write Q in terms of an integral over phase space.

$$Q = \frac{1}{N!}\frac{1}{h^{3N}} \iint \mathrm{d}\mathbf{p}\,\mathrm{d}\mathbf{x}\, e^{-\beta\mathcal{H}(\mathbf{x},\mathbf{p})} \tag{2.6}$$

Here, the factor h^{-3N}, where h is the Planck constant, is necessary to obtain a dimensionless quantity, while division by $N!$ accounts for individual particles being indistinguishable, which would otherwise lead to over-counting of microstates. Inserting Eq. 2.1, where the potential energy is assumed to be independent of the momenta, the integral over phase space can be split into separate integrals over the kinetic and potential energy.

$$Q = \frac{1}{N!}\frac{1}{h^{3N}} \int \mathrm{d}\mathbf{p}\, e^{-\frac{\beta}{2}\left(\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p}\right)} \int \mathrm{d}\mathbf{x}\, e^{-\beta U(\mathbf{x})} \tag{2.7}$$

Since in the canonical ensemble the temperature is conserved, one can take advantage of the fact that the momentum distribution obeys a Maxwell-Boltzmann distribution at thermal equilibrium to compute the momentum integral analytically, which is only an additive constant as long as no atoms are destroyed, created, or change their mass. Hence, we will often focus on the configurational integral

$$Z \coloneqq \int \mathrm{d}\mathbf{x}\, e^{-\beta U(\mathbf{x})}. \tag{2.8}$$

From the above, ensemble averages, which will be denoted by $\langle ... \rangle$, of any observable O can be calculated as

$$\langle O \rangle = \sum_{i \in \mathcal{S}} \rho(\mathbf{x}_i, \mathbf{p}_i) O_i \,, \tag{2.9}$$

and the (Helmholtz) free energy, one of the most important properties connected to the partition function, is given by

$$A = \langle \mathcal{H} \rangle - TS = -\beta^{-1} \ln Q, \tag{2.10}$$

where $\langle \mathcal{H} \rangle$ denotes the inner energy and S the entropy. The free energy is one of the central quantities of statistical thermodynamics, as according to the minimum free energy principle, systems will always tend to the state that minimizes A. Therefore, in the context of catalysis, one is often interested in calculating the relative free energy difference of two metastable states, *e.g.*, reactant R and product P of a chemical reaction,

$$\Delta A_{R \to P} = A_P - A_R = -\beta^{-1} \ln \frac{Z_P}{Z_R} \,, \tag{2.11}$$

where Z_R and Z_P are obtained by only integrating over the configuration space associated with R and P, respectively. Note that the momentum integral cancels out such that $Q_P/Q_R = Z_P/Z_R$. Exothermic reactions are characterized by a negative free energy difference, while positive values are obtained for endothermic reactions.

Unfortunately, looking at Eq. 2.6, which involves an integral over the enormous phase space, one can grasp the daunting challenge of computing the partition function and, hence, all the above properties. Indeed, it is virtually impossible to fully explore the configuration space for most chemical systems. Instead, in practice, one relies on the concept of *ergodicity*,

which states that the integral over configuration space can be replaced by a simple average over time trajectories, as obtained through MD simulations.

$$Z = \int \mathrm{d}\mathbf{x}\, e^{-\beta U(\mathbf{x})} = \lim_{t\to\infty} \frac{1}{t} \int_0^t \mathrm{d}t'\, e^{-\beta U(\mathbf{x}(t'))} \tag{2.12}$$

Ergodicity is often assumed to be valid for experiments as well as simulations, but can only rarely be proven. For MD trajectories, ergodicity is generally not fulfilled, as according to Eq. 2.4, visiting higher energy regions becomes exponentially more unlikely. Thus, simulations are trapped in metastable states from which they cannot escape on computationally feasible time scales. This problem motivates the field of importance sampling, which tries to restore ergodicity in MD simulations at least for selected target regions of configuration space.

2.2 Importance Sampling and Catalysis

In the following sections, the field of enhanced sampling is reviewed in order to put the presented work into perspective. Generally, one has to distinguish purely exploratory methods, discovering uncharted regions in configuration space, from those that allow quantitative estimation of probability distributions and free energies from the sampled space. The hyperreactor dynamics, which we introduce in Publication **VII**, falls into the former category as it allows for efficient exploration of reaction networks from a given starting point, but does not directly provide the free energy of the investigated processes. However, the main focus of this thesis lies on the latter. Specifically, methods are discussed that can characterize (bio)catalytic reaction mechanisms in terms of their reaction and activation free energy.

For this purpose, one relies on the subclass of *importance sampling* techniques, which can accurately recover the original probability distribution, while sampling from another *biased* distribution. This is achieved by modifying the potential energy to behave a certain way by adding some biasing potential U^{bias} to the physical potential energy

$$\widetilde{U}(\mathbf{x}) = U(\mathbf{x}) + U^{\mathrm{bias}}(\mathbf{x})\,, \tag{2.13}$$

which gives rise to the new bias force $-\nabla_{\mathbf{x}} U^{\mathrm{bias}}(\mathbf{x})$. Exceptions are expanded ensemble approaches [60] like replica exchange MD [61], which will not be covered in this thesis. These methods follow an orthogonal strategy by preserving the original probability distribution while exploiting transitions to other ensembles, *e.g.*, at higher temperature.

From the modified potential $\widetilde{U}(\mathbf{x})$ one obtains the biased configurational probability distribution

$$\begin{aligned} \widetilde{\rho}(\mathbf{x}) &= \frac{e^{-\beta(U(\mathbf{x})+U^{\mathrm{bias}}(\mathbf{x}))}}{\int e^{-\beta(U(\mathbf{x})+U^{\mathrm{bias}}(\mathbf{x}))}\mathrm{d}\mathbf{x}} = \frac{1}{\widetilde{Z}}\, e^{-\beta U(\mathbf{x})}\, e^{-\beta U^{\mathrm{bias}}(\mathbf{x})} = \\ &= \frac{1}{Z}\, e^{-\beta U(\mathbf{x})}\, \frac{Z}{\widetilde{Z}}\, e^{-\beta U^{\mathrm{bias}}(\mathbf{x})} = \rho(\mathbf{x})\, \frac{Z}{\widetilde{Z}}\, e^{-\beta U^{\mathrm{bias}}(\mathbf{x})}\,, \end{aligned} \tag{2.14}$$

with biased configurational integral $\widetilde{Z}$. Therefore, from Eq. 2.14 the unbiased probability distribution can be recovered from the biased distribution as

$$\rho(\mathbf{x}) = \widetilde{\rho}(\mathbf{x})\,\frac{\widetilde{Z}}{Z}\,e^{+\beta U^{\mathrm{bias}}(\mathbf{x})}. \tag{2.15}$$

The main difficulty of computing Eq. 2.15 is the evaluation of the ratio of original and biased configurational integrals, for which, after a few transformations, one arrives at

$$\frac{Z}{\widetilde{Z}} = \frac{\int e^{-\beta U(\mathbf{x})}\,\mathrm{d}\mathbf{x}}{\int e^{-\beta \widetilde{U}(\mathbf{x})}\,\mathrm{d}\mathbf{x}} = \frac{\int e^{-\beta U(\mathbf{x})}\,e^{-\beta U^{\mathrm{bias}}(\mathbf{x})}\,\mathrm{d}\mathbf{x}}{\int e^{-\beta \widetilde{U}(\mathbf{x})}\,e^{-\beta U^{\mathrm{bias}}(\mathbf{x})}\,\mathrm{d}\mathbf{x}} = \frac{\int e^{-\beta \widetilde{U}(\mathbf{x})}\,e^{+\beta U^{\mathrm{bias}}(\mathbf{x})}\,\mathrm{d}\mathbf{x}}{\int e^{-\beta \widetilde{U}(\mathbf{x})}\,\mathrm{d}\mathbf{x}} = \left\langle e^{+\beta U^{\mathrm{bias}}(\mathbf{x})} \right\rangle_{\widetilde{U}}, \tag{2.16}$$

with the ensemble average under the biased distribution denoted by $\langle ... \rangle_{\widetilde{U}}$. Hence, the true ensemble average can be recovered from biased simulations via

$$\begin{aligned} \langle O(\mathbf{x}) \rangle &= \int O(\mathbf{x})\,\rho(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int O(\mathbf{x})\,\widetilde{\rho}(\mathbf{x})\,\frac{\widetilde{Z}}{Z}\,e^{+\beta U^{\mathrm{bias}}(\mathbf{x})}\,\mathrm{d}\mathbf{x} = \\ &= \left\langle O(\mathbf{x})\,\frac{\widetilde{Z}}{Z}\,e^{+\beta U^{\mathrm{bias}}(\mathbf{x})} \right\rangle_{\widetilde{U}} = \frac{\left\langle O(\mathbf{x})\,e^{+\beta U^{\mathrm{bias}}(\mathbf{x})} \right\rangle_{\widetilde{U}}}{\left\langle e^{+\beta U^{\mathrm{bias}}(\mathbf{x})} \right\rangle_{\widetilde{U}}}. \end{aligned} \tag{2.17}$$

One of the most important concepts of importance sampling methods are collective variables (CVs), which are functions $\xi : \mathbb{R}^{N_c} \rightarrow \mathbb{R}^d$ with $d << N_c$, that map high dimensional configurations onto a low-dimensional representation

$$\mathbf{z} = (z_1, z_2, ..., z_d) = \xi(\mathbf{x})\,. \tag{2.18}$$

The marginal probability density $\rho(\mathbf{z})$ is defined by the integral over all variables except $\mathbf{z}$

$$\rho(\mathbf{z}) = \int \delta[\mathbf{z} - \xi(\mathbf{x})]\rho(\mathbf{x})\,d\mathbf{x} = \langle \delta[\mathbf{z} - \xi(\mathbf{x})] \rangle\,, \tag{2.19}$$

where $\delta[...]$ denotes the multivariate Dirac delta distribution. The potential of mean force (PMF), which is also often referred to as the free energy surface (FES), is defined via

$$A(\mathbf{z}) = -\beta^{-1} \ln \rho(\mathbf{z})\,. \tag{2.20}$$

Although by itself not gauge invariant against the choice of CV, $\rho(\mathbf{z})$ and $A(\mathbf{z})$ are of central importance for the simulation of (bio)catalytic reaction mechanisms, as they are related to the reaction and activation free energy, as will be shown further below. The unbiased PMF can be recovered from biased simulations by again inserting Eq. 2.15

$$\begin{aligned} A(\mathbf{z}) &= -\beta^{-1}\,\ln\left[\widetilde{\rho}(\mathbf{z})\,\frac{\widetilde{Z}}{Z}\,e^{+\beta U^{\mathrm{bias}}(\mathbf{z})}\right] \\ &= \widetilde{A}(\mathbf{z}) - U^{\mathrm{bias}}(\mathbf{z}) - \beta^{-1}\,\ln\frac{\widetilde{Z}}{Z} \\ &= \widetilde{A}(\mathbf{z}) - U^{\mathrm{bias}}(\mathbf{z}) + \Delta A_{U\rightarrow\widetilde{U}}\,, \end{aligned} \tag{2.21}$$

with biased PMF $\widetilde{A}(\mathbf{z})$. Here, apart from $U^{\text{bias}}(\mathbf{z})$, the free energy difference between the biased and unbiased systems appears as a second correction term. If only one biasing potential is considered, this term is only an additive constant that can be neglected. However, for example, in Umbrella Sampling (US) [35] (see section 2.4.1), one runs multiple simulations with different biasing potentials, each contributing only a fragment to the PMF. Therefore, statistical estimators are required that can recover a continuous PMF, ideally globally with similar precision for all contributing simulations.

For this purpose, many popular methods were developed, which can roughly be divided into two main categories. Methods of the first category yield fast and accurate estimates of the PMF, but are not able to recover other ensemble properties. Examples are techniques based on thermodynamic integration like Umbrella Integration (UI) [62], which instead of the PMF computes its derivative to avoid the unknown third term of Eq. 2.21.

$$\frac{\partial A(\mathbf{z})}{\partial \mathbf{z}} = \frac{\partial \widetilde{A}(\mathbf{z})}{\partial \mathbf{z}} - \frac{\partial U^{\text{bias}}(\mathbf{z})}{\partial \mathbf{z}} = \frac{\beta^{-1}\,\partial \ln \widetilde{\rho}(\mathbf{z})}{\partial \mathbf{z}} - \frac{\partial U^{\text{bias}}(\mathbf{z})}{\partial \mathbf{z}} \tag{2.22}$$

Subsequently, Eq. 2.22 is integrated to obtain the PMF up to an additive constant, which remains elusive.

In contrast, methods of the second category explicitly calculate the critical free energy differences of biased ensembles and are hence able to recover the full statistical information from biased simulations. Popular examples are the weighted histogram analysis method (WHAM) [63] or the multistate Bennett acceptance ratio (MBAR) [48], a bin-less version of WHAM. The MBAR is also the multi-state generalization of BAR [64], which aims at a minimization of the variance in the free energy estimate from only two states. As BAR for two states, MBAR is the lowest variance unbiased estimator for both free energies and ensemble averages from multiple thermodynamic states. Since it is of central importance for this work, the MBAR equations are derived below based on the new derivation published by Michael Shirts [65], which is more intuitive than the one presented in the original publication (Ref. [48]).

Assume M samples have been collected from biased potential U_i proportional to their configurational probability distribution

$$\rho_i(\mathbf{x}) = \frac{e^{-\beta U_i(\mathbf{x})}}{Z_i}\,. \tag{2.23}$$

Then ensemble averages of observables can be obtained from ρ_i via Eq. 2.9. Dividing and multiplying by a second normalized and well-behaved distribution $\rho_j(\mathbf{x})$ leads to

$$\langle O \rangle_i = \sum_{m=1}^{M} O(\mathbf{x}_m) \left(\frac{\rho_i(\mathbf{x}_m)}{\rho_j(\mathbf{x}_m)} \right) \rho_j(\mathbf{x}_m)\,. \tag{2.24}$$

If samples are instead picked proportional to $\rho_j(\mathbf{x})$, the same average can be calculated from the Monte Carlo integration

$$\langle O \rangle_i = \frac{1}{M} \sum_{m=1}^{M} O(\mathbf{x}_m) \left(\frac{\rho_i(\mathbf{x}_m)}{\rho_j(\mathbf{x}_m)} \right)\,. \tag{2.25}$$

At this point, a similar problem as in Eq. 2.21 arises, namely that $\rho_i(\mathbf{x})$ and $\rho_j(\mathbf{x})$ are only known up to the unknown constants Z_i and Z_j. A solution can be found by simply choosing the artificial observable $O(\mathbf{x}) = 1$, which yields, after a few transformations

$$\begin{aligned} \langle 1 \rangle_i &= \frac{1}{M} \sum_{m=1}^{M} (1) \left(\frac{\rho_i(\mathbf{x}_m)}{\rho_j(\mathbf{x}_m)} \right) \\ 1 &= \frac{1}{M} \sum_{m=1}^{M} \left(\frac{e^{-\beta U_i(\mathbf{x}_m)} Z_i^{-1}}{e^{-\beta U_j(\mathbf{x}_m)} Z_j^{-1}} \right) \\ \frac{Z_i}{Z_j} &= \frac{1}{M} \sum_{m=1}^{M} \left(\frac{e^{-\beta U_i(\mathbf{x}_m)}}{e^{-\beta U_j(\mathbf{x}_m)}} \right) . \end{aligned} \tag{2.26}$$

From the definition of the free energy, one can insert $Z_i = e^{-\beta A_i}$. Taking the logarithm, one arrives at the standard reweighting method by Zwanzig from 1954 [66], which today is well known as the free energy perturbation (FEP) method.

$$\begin{aligned} \ln \left[\frac{e^{-\beta A_i}}{e^{-\beta A_j}} \right] &= \ln \left[\frac{1}{M} \sum_{m=1}^{M} \left(\frac{e^{-\beta U_i(\mathbf{x}_m)}}{e^{-\beta U_j(\mathbf{x}_m)}} \right) \right] \\ A_j - A_i &= -\beta^{-1} \ln \left[\frac{1}{M} \sum_{m=1}^{M} e^{-\beta (U_i(\mathbf{x}_m) - U_j(\mathbf{x}_m))} \right] \\ \Delta A_{ij} &= -\beta^{-1} \ln \left\langle e^{-\beta \Delta U_{ij}(\mathbf{x}_m)} \right\rangle_j \end{aligned} \tag{2.27}$$

Now the key step is to realize that the MBAR can be seen as reweighting from a *mixture distribution*. Assume M_k samples have been collected from K distributions, *e.g.*, by means of US, the mixture distribution is obtained by simply combining all $M = \sum_{k=1}^{K} M_k$ samples

$$\rho^{\text{mix}}(\mathbf{x}) = \frac{1}{M} \sum_{k=1}^{K} M_k \rho_k(\mathbf{x}) = \frac{1}{M} \sum_{k=1}^{K} M_k \frac{e^{-\beta U_k(\mathbf{x})}}{Z_k} . \tag{2.28}$$

Note, that if all individual distributions $\rho_k(\mathbf{x})$ are normalized, $\rho^{\text{mix}}(\mathbf{x})$ must also be normalized and that there is an M_k/M chance to draw a sample from any distribution k. Using the same trick as above, but for reweighting from $\rho^{\text{mix}}(\mathbf{x})$ to $\rho_i(\mathbf{x})$, one obtains

$$\begin{aligned} 1 &= \frac{1}{M} \sum_{m=1}^{M} \left(\frac{\rho_i(\mathbf{x}_m)}{\rho^{\text{mix}}(\mathbf{x}_m)} \right) \\ 1 &= \frac{1}{M} \sum_{m=1}^{M} \left(\frac{e^{-\beta U_i(\mathbf{x}_m)} \, Z_i^{-1}}{M^{-1} \sum_{k=1}^{K} M_k e^{-\beta U_k(\mathbf{x}_m)} Z_k^{-1}} \right) \\ Z_i &= \sum_{m=1}^{M} \frac{e^{-\beta U_i(\mathbf{x}_m)}}{\sum_{k=1}^{K} M_k e^{-\beta U_k(\mathbf{x}_m)} Z_k^{-1}} , \end{aligned} \tag{2.29}$$

where before solving for Z_i in the second line Eq. 2.23 and Eq. 2.28 are inserted. Finally, after again inserting $Z_i = e^{-\beta A_i}$, one arrives at the MBAR

$$e^{-\beta A_i} = \sum_{m=1}^{M} \frac{e^{-\beta U_i(\mathbf{x}_m)}}{\sum_{k=1}^{K} M_k e^{\beta(A_k - U_k(\mathbf{x}_m))}}\,, \tag{2.30}$$

which is a system of K equations that can be solved for $\mathbf{A} = (A_1, A_2, ..., A_K)$. Typically, as one is only interested in relative free energies, one chooses $A_1 = 0$ and solves the remaining $K-1$ equations. Note, that as $\mathbf{A}$ enters both sides Eq. 2.30 has to be solved self-consistently, or can alternatively be recast into a minimization problem [48]. Reweighting (*i.e.*, importance sampling) of observables at any state from the mixture distribution can be written

$$\begin{aligned} \langle O \rangle_i &= \frac{1}{M} \sum_{m=1}^{M} O(\mathbf{x}_m) \frac{\rho_i(\mathbf{x}_m)}{\rho^{mix}(\mathbf{x}_m)} \\ &= \sum_{m=1}^{M} O(\mathbf{x}_m) \frac{e^{\beta(A_i - U_i(\mathbf{x}_m))}}{\sum_{k=1}^{K} M_k e^{\beta(A_k - U_k(\mathbf{x}_m))}} \\ &= \sum_{m=1}^{M} O(\mathbf{x}_m) W_{im} \end{aligned} \tag{2.31}$$

Therefore, the full statistical information is recovered as equilibrium expectations of any observable can be obtained.

Independent of the statistical estimator of choice, if the CV space is capable of distinguishing two metastable states like the reactant R and product P of a chemical transition, the reaction free energy $\Delta A_{R\to P}$ can be obtained by integrating the marginal probability density over the corresponding domains of CV space [67].

$$\Delta A_{R\to P} = -\beta^{-1} \ln \frac{\int_P \rho(\mathbf{z})\, d\mathbf{z}}{\int_R \rho(\mathbf{z})\, d\mathbf{z}} = -\beta^{-1} \ln \frac{Z_P}{Z_R} \tag{2.32}$$

Additionally, for a one dimensional CV space $\xi(\mathbf{x}) = z$, an expression that relates the activation free energy $\Delta A^{\ddagger}_{R\to P}$ to $\rho(z)$ is provided in Publication **VI**:

$$\Delta A^{\ddagger}_{R\to P} = -\beta^{-1} \ln \frac{\rho(z^{\ddagger}) \left\langle \lambda_\xi \right\rangle_{z^{\ddagger}}}{Z_R}\,, \tag{2.33}$$

where $z^{\ddagger}$ denotes the transition state (TS) in CV space and $\langle ... \rangle_{z^{\ddagger}}$ denotes averages over the transition state ensemble (TSE). $\lambda_\xi = \sqrt{\beta h^2 / 2\pi m_\xi}$ is the thermal de-Broglie wave length of the pseudo-particle associated with the CV with mass

$$m_\xi^{-1} = (\nabla_{\mathbf{x}} \xi)^{\mathrm{T}}\, \mathbf{M}^{-1}\, (\nabla_{\mathbf{x}} \xi) \tag{2.34}$$

Note that m_ξ^{-1} depends on the gradient of the CV and is only constant if the CV is linear in Cartesian coordinates. Therefore, besides accounting for the mass of atoms involved

in the transition, capturing, for example, kinetic isotope effects, it removes distortions of the Cartesian space by non-linear CVs. The frequently employed difference of maxima and minima of the PMF $\Delta A^{\ddagger}_{R\to P} \approx A(z^{\ddagger}) - A(z^{\min})$ can be seen as an approximation of Eq. 2.32, ignoring distortions of the coordinate system and the influence of mass while assuming the probability density to be sharply peaked in the reactant minimum. Reaction rate constants $k_{R\to P}$, which are often experimentally accessible as well, can be obtained from the $\Delta A^{\ddagger}_{R\to P}$ via Eyring's equation

$$k_{R\to P} = \frac{1}{\tau_{R\to P}} = \frac{\kappa}{\beta h} e^{-\beta \Delta A^{\ddagger}_{R\to P}} , \tag{2.35}$$

where κ is the transition coefficient, which is often ignored under the assumption that all trajectories that reach the transition state also cross it ($\kappa = 1$). This corresponds to computing an upper bound to the true reaction rate constant.

A different strategy for estimating kinetic rates can be found by directly observing the first passage time $\tau_{R\to P}$ from multiple biased simulations that cross the barrier. Inserting Eq. 2.33 into Eq. 2.35 one obtains

$$\frac{1}{\tau_{R\to P}} \propto \frac{Z_{\mathrm{TS}}}{Z_R} . \tag{2.36}$$

Now, think of adding a biasing potential that only affects the reactant state, reducing the effective barrier, but is zero at the TSE. Looking at Eq. 2.36, such a bias only affects Z_R. Therefore, as everything else cancels, one obtains the ratio of unbiased to accelerated escape times

$$\alpha = \frac{\tau_{R\to P}}{\widetilde{\tau}_{R\to P}} = \frac{Z_R}{\widetilde{Z}_R} = \frac{\int_R e^{-\beta U(\mathbf{x})} \mathrm{d}\mathbf{x}}{\int_R e^{-\beta \widetilde{U}(\mathbf{x})} \mathrm{d}\mathbf{x}} = \left\langle e^{+\beta U^{\mathrm{bias}}(\mathbf{x})} \right\rangle_{\widetilde{U},R} , \tag{2.37}$$

which is also known as *acceleration factor*. Note that Eq. 2.37 is identical to Eq. 2.16 except for the configurational integral that now only runs over the reactant state. Hence, once transitions are observed, biased transition times can be rescaled via

$$\tau_{R\to P} = \sum_{i=1}^{M_{\mathrm{tot}}} \Delta t \, e^{\beta U^{\mathrm{bias}}(\mathbf{x}_i)} , \tag{2.38}$$

with discrete time step Δt and total number of steps until a transition is observed M_{tot}. While such an approach was first introduced in 1997 by Voter in the context of hyperdynamics [68], today there are many methods for constructing adequate *flooding bias* potentials that leave the TSE bias free [69–73]. The advantage is that, in contrast to Eq. 2.33, no extensive sampling at the TS is required as TSEs enter in no form [74]. On the other hand, compared to 2.33, significant computational overhead might arise as the estimation $\Delta A_{R\to P}$ requires a separate calculation. Additionally, before constructing the flooding bias, one often still has to approximately compute $\Delta A^{\ddagger}_{R\to P}$ anyway to obtain an initial estimate of barrier heights one wishes to overcome and ensure that the TSE remains bias-free. Note that because of the exponential in Eq. 2.35, the underlying potential energy has to be

obtained very precise to get quantitative results of $k_{R\to P}$. Therefore, independent of the chosen approach, in practice, a qualitative agreement of experimental and predicted rate constants is often the only achievable goal.

Hence, the general strategy for investigating reaction mechanisms in terms of importance sampling is summarized in Fig. 2.1. While the third step, reweighting, was thoroughly discussed in this section, a comprehensive discussion of the first two steps is still missing. Therefore, in the following section, we will take a closer look at step 1, dimensionality reduction, giving some insight into the practical choice of CVs and their implications for the results of simulations. Thereafter, step 2 will be reviewed, which involves the actual sampling. Here, we dive into the large and sometimes confusing zoo of different sampling algorithms to illuminate their similarities and differences.

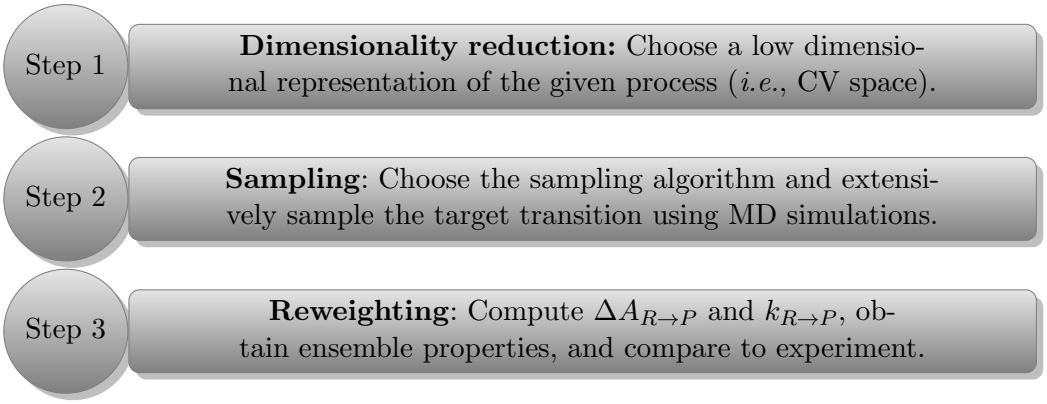


Fig. 2.1. General workflow for the characterization of (bio)chemical reaction mechanisms through importance sampling.

2.3 Dimensionality Reduction

The definition of a suitable CV space that includes all slow degrees of freedom is often crucial for successful sampling. Many methods apply only low-dimensional bias potentials because of the steep rise in computational cost associated with sampling higher-dimensional spaces. In such cases, bias forces are given by the chain rule

$$-\nabla_{\mathbf{x}} U^{\text{bias}}(\xi(\mathbf{x})) = -\frac{\partial U^{\text{bias}}}{\partial \xi}\frac{\partial \xi}{\partial \mathbf{x}}\,, \tag{2.39}$$

which illuminates that metastability will only be removed along ξ. Important characteristics that have to be fulfilled for optimal CVs include that they

(a) distinguish between long-lived metastable states,

(b) are orthogonal to the dividing surface between both metastable states to correctly capture the TSE,

(c) provide a separation of timescales, such that the dynamics of remaining fast variables is unmixed and negligible,

(d) are not degenerate, such that each CV describes a different slow mode,

(e) are Markovian, such that the slow dynamics are described as an evolution in a free-energy landscape with a configuration-dependent diffusion coefficient,

(f) are smooth and differentiable, such that they can be applied together with importance sampling algorithms according to Eq. 2.39.

One always has to consider that results obtained with poor CVs might be wrong and misleading. For example, in Publication **VI** we show, that while the ΔA is relatively robust to suboptimal CVs, $\Delta A^{\ddagger}$ is highly sensitive. This is because the latter explicitly depends on the probability density at the TSE (Eq. 2.33), which is only estimated correctly if point (b) is fulfilled. The robustness of the former can be understood by considering that the probability density is commonly sharply peaked in metastable states, such that Eq. 2.32 is accurate as long as the CV successfully discriminates metastable states.

The quality of CVs can be tested based on computing the *committor function* $q_{\mathrm{P}}(\mathbf{x})$, as first introduced by Kolmogorov [75], which is a statistical measure of how committed a trajectory is to state P on an energy surface with metastable states R and P. Specifically, it gives the probability of a trajectory starting in $\mathbf{x}$, ending in state P without having first passed through R. Hence, for the TSE it must be fulfilled that

$$q_{\mathrm{P}}(\mathbf{x}) \simeq q_{\mathrm{R}}(\mathbf{x}) \simeq \frac{1}{2} \tag{2.40}$$

However, the computation of $q_{\mathrm{P}}(\mathbf{x})$ is challenging, as it has to be numerically estimated from selected snapshots of the sampled TSE via *committor analysis* [76]. For this purpose, many short trajectories are started from configurations at $\mathbf{z}^{\ddagger}$ with Maxwell-Boltzmann velocities, and the number that reaches P before R is counted. This approach is expensive and depends on the criteria that are applied to choose which basin a trajectory is committed to.

Therefore, in Publication **VI**, a different criterion is proposed to check if it is safe to calculate activation free energies from the PMFs along the chosen CV

$$D_s(\mathbf{z}) = \left\langle \left| \frac{\nabla_{\mathbf{x}}\xi(\mathbf{x})}{|\,\nabla_{\mathbf{x}}\xi(\mathbf{x})\,|} \cdot \frac{\nabla_{\mathbf{x}}U(\mathbf{x})}{|\,\nabla_{\mathbf{x}}U(\mathbf{x})\,|} \right| \right\rangle_{\mathbf{z}} , \tag{2.41}$$

which is related to an earlier measure introduced by Dietschreit *et al.* [67]. Later, in Publication **II**, $D_s(\mathbf{z})$ is successfully applied to monitor the quality of an adaptive CV on-the-fly. For an ideal CV $D_s(\mathbf{z}^{\ddagger}) = 0$, which indicates that the CV is orthogonal to the dividing surface (*i.e.*, gradients of U and ξ are perpendicular). In practical cases, $D_s(\mathbf{z})$ should at least show a clear minimum at $\mathbf{z} = \mathbf{z}^{\ddagger}$. An illustration of both the committor and $D_s(\mathbf{z})$ is given in Fig. 2.2.

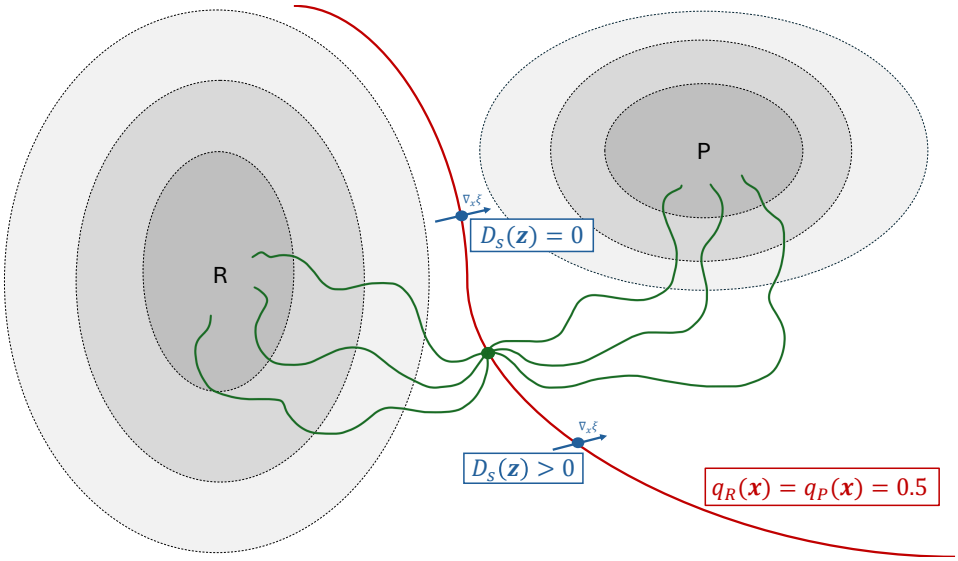


Fig. 2.2. Illustration of analyzing the transition state ensemble (TSE). Probability densities of two metastable states, R and P, are indicated in gray. Their dividing surface is given in red, for which Eq. 2.40 holds. Green lines denote trajectories from commitor analysis that start from a point on the dividing surface and half commit to each state. In contrast, the CV measure $D_s(\mathbf{z})$ is only zero where the gradient of the CV $\nabla_{\mathbf{x}}\xi$, indicated by a blue arrow, is orthogonal to the dividing surface.

CVs are commonly chosen *a priori* based on the chemical intuition of the practitioner. For example, a linear combination of the breaking and forming bond lengths might be considered for a simple chemical reaction where one covalent bond is broken and another is formed. Unfortunately, for complicated biochemical transitions, the process of manually identifying suitable CVs can be highly cumbersome or even impractical. For example, in Publication **II** we show that applied to an inherently non-linear transition, such a simple linear combination of breaking and forming bond distances leads to significant sampling artifacts.
Alternatively, for some processes, physics-inspired CVs might be considered. One example is our study of biological long-range proton transfer (pT) reactions in cooperation with the group of Ville R. I. Kaila (Publication **IV**). These highly intricate processes are catalyzed by titratable amino acids that form proton wires together with water molecules, as indicated in Fig. 2.3. Water-mediated proton transfer occurs via bond rearrangement in a Grotthuss-type mechanism. Therefore, the CV can intuitively be chosen as a linear combination of all breaking and forming covalent bonds of hydrogen atoms. However, such a CV explicitly includes only one hydrogen atom per water molecule, while the other cannot participate in the pT. Additionally, pT pathways must be selected manually before the simulation, and each pathway must be probed in a unique simulation. As an alternative, we decided to apply the modified center of excess charge coordinate (mCEC), as introduced by König *et*

al. [54]:

$$\zeta_{\mathrm{CEC}}(\mathbf{x}) = \sum_{i}^{N_H} \mathbf{x}_i - \sum_{j}^{N_X} w_j \mathbf{x}_j - \sum_{i}^{N_H} \sum_{j}^{N_X} f_{\mathrm{SW}}(\mid \mathbf{x}_i - \mathbf{x}_j \mid)(\mathbf{x}_i - \mathbf{x}_j), \tag{2.42}$$

where N_H, N_X are the number of involved protons and proton acceptors, respectively, and $\mathbf{x}_i$ denotes the Cartesian coordinates of atom i. The weights w_j are given by the minimal number of protons associated with atom j during pT, and f_{SW} is the switching function

$$f_{\mathrm{SW}}(r) = \left(1 + e^{(r - r_{\mathrm{SW}})/d_{\mathrm{SW}}}\right) \tag{2.43}$$

where r_{SW} controls the switching distance and d_{SW} how fast the function switches from one to zero. Note that ζ_{CEC} is a three-dimensional quantity. Therefore, we choose to project ζ_{CEC} onto the direction of pT via

$$\xi_{\mathrm{CEC}}(\mathbf{x}) = (\zeta_{\mathrm{CEC}} - \mathbf{x}_d) \cdot \frac{\mathbf{x}_a - \mathbf{x}_d}{\mid \mathbf{x}_a - \mathbf{x}_d \mid}, \tag{2.44}$$

with coordinates of first donor and last acceptor atom $\mathbf{x}_d$ and $\mathbf{x}_a$, respectively. The first two terms of Eq. 2.42 reflect the position of the excess proton in a water wire, while the third term is a correction that removes all contributions from the CEC that are irrelevant to pT. Thus, in contrast to the unmodified CEC (first two terms), the mCEC coordinate is close to zero if there is no excess proton. The advantage of ξ_{CEC} is that it provides a global description of all possible pT pathways. Thus, one can explore new pT pathways on-the-fly during the simulation without manually adapting the CV.

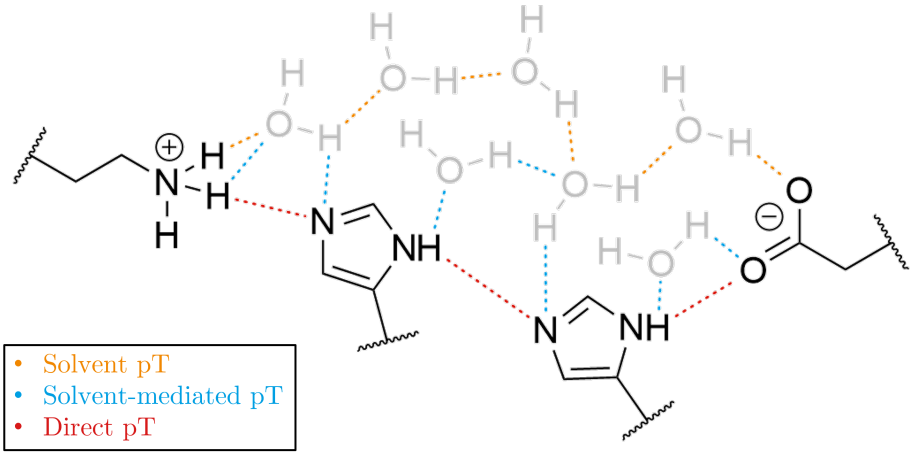


Fig. 2.3. Chemical drawing of long-range pT in the mammalian respiratory Complex I, as investigated in Publication **IV**. Three possible reaction pathways are marked, denoting direct pT between titrable amino acids (red), water-mediated pT between amino acids (blue), and pure solvent pT (orange). Unlike traditional CVs, the mCEC provides a global description of all these processes and their permutations, enabling the on-the-fly exploration of low-energy pT pathways through importance sampling.

As a third strategy towards dimensionality reduction, one can use adaptive CVs that learn to describe the given process more and more accurately before or during the simulation. Here, two main strategies were developed by the community [77]: adaptive path CVs (PCVs) and machine-learning-derived CVs. In publications **II** and **III**, the former are applied in a highly efficient manner to resolve the intricate reaction mechanisms of RNA modification by PUS and ATP hydrolysis by p97. As first demonstrated by Weinan *et. al.* in the string method [50], the idea is to connect two metastable states with a path, which is represented by a discrete string of nodes. The CVs are given by a progress parameter, often denoted $s(\mathbf{x})$, and a distance metric, often denoted $z(\mathbf{x})$, for which arithmetic [51] or geometric [52, 78] formulations exist. Additionally, there are combinations of PCVs with machine learning [79, 80]. The biggest advantage is that even non-linear transitions are trivial to describe by non-linear paths. For this thesis, mainly the geometric PCV was applied, which is a function of a pre-defined CV space $\mathbf{z}$ that has to include all important degrees of freedom. The geometric PCV is given by the progress parameter

$$s(\mathbf{z}) = \frac{m}{M} \pm \frac{1}{2M}\left(\frac{\sqrt{\mathbf{v}_1 \cdot \mathbf{v}_3)^2 - |\mathbf{v}_3|^2(|\mathbf{v}_1|^2 - |\mathbf{v}_2|^2)} - (\mathbf{v}_1 \cdot \mathbf{v}_3)}{|\mathbf{v}_3|^2} - 1\right), \tag{2.45}$$

and the distance to the path

$$z(\mathbf{z}) = \left|\mathbf{v}_1 + \frac{1}{2}\left(\frac{\sqrt{\mathbf{v}_1 \cdot \mathbf{v}_3)^2 - |\mathbf{v}_3|^2(|\mathbf{v}_1|^2 - |\mathbf{v}_2|^2)} - (\mathbf{v}_1 \cdot \mathbf{v}_3)}{|\mathbf{v}_3|^2} - 1\right)\mathbf{v}_4\right|, \tag{2.46}$$

with index of the first, second and third-closest nodes m, n, k, respectively, and vectors $\mathbf{v}_1 = \mathbf{z}_m - \mathbf{z}$, $\mathbf{v}_2 = \mathbf{z} - \mathbf{z}_{m-1}$, $\mathbf{v}_3 = \mathbf{z}_{m+1} - \mathbf{z}_m$ and $\mathbf{v}_4 = \mathbf{z} - \mathbf{z}_{m-1}$. The $\pm$ in Eq. 2.45 is negative if $\mathbf{z}$ is left of the closest path node, and positive otherwise. The path can be refined during the simulation by moving nodes closer to the average transition path in an iterative manner, which yields convergence to the minimum free energy path (MFEP) [52]. This classifies adaptive PCVs as a sort of 'self-learning' algorithm [77] and, thus, as an alternative to machine-learning CVs. However, in contrast to many machine learning approaches, no initial data is required, such that the simulation can be started directly using any guess path, which can easily be obtained, *e.g.*, from linear interpolation or minimum energy path optimization methods [81]. The former strategy was used in Publication **II**, where linear guess paths were adaptively converged towards the MFEPs. The latter strategy was employed in Publication **III**, where PMFs were computed along a static path that was optimized beforehand using the improved nudged elastic band (NEB) method [82]. An illustration of one of the obtained reaction pathways is given in Fig. 2.4.

2.4 Importance Sampling Algorithms

In the last section, the problem of dimensionality reduction was addressed. Now, assuming a good CV that describes the full slow dynamics of the given process has been identified, methods are discussed that can remove metastability along this CV. Note that all the

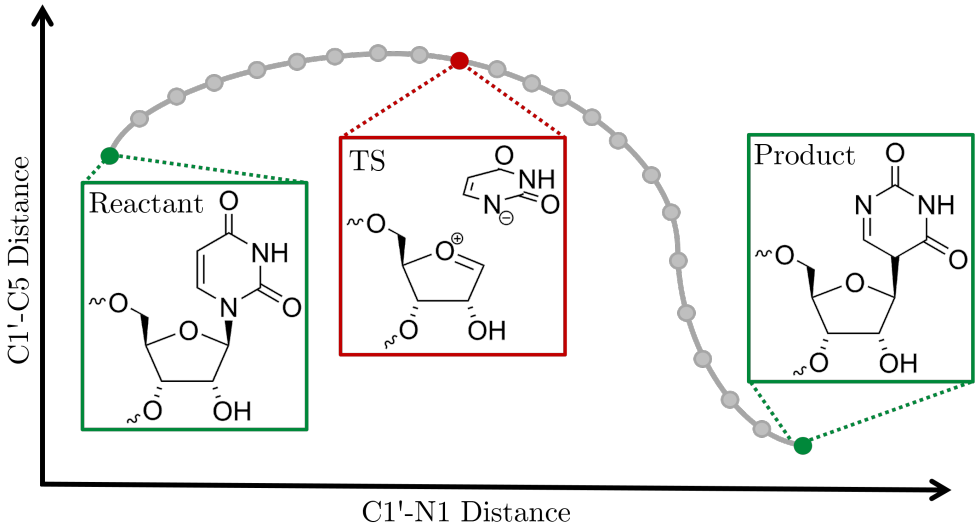


Fig. 2.4. Illustration of the rebound mechanism of PUS as investigated in Publication **II**. In gray, a path connecting uridin with the rebound intermediate is shown, in the space of the breaking C1'-N1 bond on the x-axis and the forming C1'-C5 bond on the y-axis. The reactant and product nodes are marked in green, while the node located at the transition state (TS) is marked in red. The inserts show chemical drawings of configurations that correspond to the respective path nodes.

discussed approaches were publicly implemented in the open-source `adaptive-sampling` package as part of this thesis.

2.4.1 Static Biasing Methods

The simplest sampling strategy is the application of static biasing potentials, which increase the probability of visiting certain regions of configuration space. Such an approach was first introduced by Torrie and Valleau in 1976 under the term Umbrella Sampling (US) [34]. Today, typically harmonic potentials are applied

$$U_i^{\mathrm{US}}(\mathbf{z}) = \frac{k_i}{2}(\mathbf{z} - \mathbf{z}_i^0)^2\,, \tag{2.47}$$

with force constant k_i and equilibrium value of the CV $\mathbf{z}_i^0$ [35]. Hence, for sufficiently strong coupling (large k_i), each simulation, which in US is also often referred to as *window*, will be in a local equilibrium close to $\mathbf{z}_i^0$, and one can simply place multiple overlapping windows i to sample selected regions of CV space. In Publication **IV**, US is applied to obtain local sampling in regions of configuration space that were only sparsely visited by a second sampling approach, which, on the contrary, is designed to explore new transition pathways. As described in section 2.2, the PMF can be recovered by several different methods, including UI [62], WHAM [63], or MBAR [48].

A significant drawback of this approach is that local constraints naturally lead to quasi-nonergodic effects, particularly when multiple competing reaction pathways are present. In other words, by design, U_i^{US} limits the search space, which is generally detrimental to efficient sampling. As a result, free energies might seriously depend on initial conditions, as demonstrated in a recent study by Aho *et al.* [83] that found differences of 2–20 kcal/mol in computed binding free energies for repeating a workflow based on US with changing starting configurations. Additionally, already in their original paper Torrie and Valleau [34] note, that there is no straightforward way to determine good biasing (or in their language weighting) functions and conclude, that „A possible embellishment of the technique would be to program the computer to carry out itself the trial-and-error development of a good weighting function". Addressing both problems is the aim of adaptive sampling techniques, which are the main focus of this work and will be discussed in the following sections.

But first, a special case of static biasing potentials will be addressed. These sampling approaches, instead of depending on any CV, only depend on the potential energy itself. This idea originates from the hyperdynamics method by Voter [68], which on its own is unfeasible for large systems because of its explicit dependence on the Hessian matrix. Today, several more efficient methods have been developed, including accelerated MD (aMD) [84], Gaussian accelerated MD (GaMD) [46], and Sigmoid accelerated MD (SaMD) [85]. All of them modify the potential energy below some threshold E with an energy-dependent boost potential ΔU:

$$U^{\mathrm{aMD}}(\mathbf{x}, U) = \begin{cases} U(\mathbf{x}) & \text{if } U(\mathbf{x}) \geq E, \\ U(\mathbf{x}) + \Delta U(U(\mathbf{x})) & \text{if } U(\mathbf{x}) < E\,. \end{cases} \tag{2.48}$$

Hence, the underlying topology of the energy surface is roughly preserved, but all energy barriers are lowered, and MD simulations can escape from kinetic traps. Again, reweighting of such simulations to obtain unbiased probability distributions and reaction rate constants is possible using standard importance sampling (see section 2.2). Often, to remove energetic noise in PMF estimates, instead of directly reweighting from Eq. 2.17, one approximates $\mathbf{z}$-conditioned ensemble averages of the boost potential with a cumulant expansion

$$\left\langle e^{\beta \Delta U(U(\mathbf{x}))} \right\rangle_{\Delta U, \mathbf{z}} = \exp\left[\sum_{k=1}^{\infty} \frac{\beta^k}{k!} C_k \right] , \tag{2.49}$$

with expansion coefficients C_k, that for the first two orders are given by the mean and standard deviation of the sampled boost potential [46]. Usually, the cumulant expansion is truncated after the second order, essentially assuming a Gaussian distribution of the boost potential at $\mathbf{z}$. In Publication **I**, we show that if GaMD (and in principle also the other forms of the boost potential) is combined with CV-dependent sampling approaches that allow for the application of the MBAR, it is more accurate to also directly incorporate the boost potential into the MBAR equations. Additionally, in Publication **VII**, we combine aMD, GaMD, and SaMD with a pressure-inducing potential as inspired by the ab initio nanoroeactor [86, 87] to devise a much more efficient and robust approach for reaction network exploration, which we term hyperreactor dynamics (HRD) [88].

2.4.2 Adaptive Biasing Potential Methods

In this section, methods are discussed that learn the biasing potential during the simulation based on information from the trajectory. Therefore, the biasing potential depends on the statistical properties of the trajectory and is progressively updated as more information becomes available. A large zoo of different methods is available to build such time-dependent biasing potentials. To name only a few of the more exotic ones, there is local elevation [89], self-healing umbrella sampling [90], basis function sampling [91], artificial neural network sampling [92]. The interested reader is referred to the LiveCoMS perpetual review on enhanced sampling [93], which is a continuously maintained and updated review on sampling methods. All of these methods share common ideas, but differ in the concrete way the biasing potential is built. Hence, for the sake of this thesis, we will focus on metadynamics (MtD), which was first introduced by Laio and Parrinello [36], and has since become an invaluable tool for molecular simulation [37, 94, 95].

The core idea of MtD is to introduce a repulsive biasing potential, which is written as a superposition of Gaussian hills

$$U^{\mathrm{MtD}}(\mathbf{z},t) = \sum_{t=0,\tau_{\mathrm{G}},2\tau_{\mathrm{G}},\ldots} h_{\mathrm{G}} e^{-(\mathbf{z}-\mathbf{z}_i)^2/2\sigma_{\mathrm{G}}^2}\,, \tag{2.50}$$

where h_{G} is the hill height and $\sigma_{\mathrm{G}} = (\sigma_1, ..., \sigma_d)$ are standard deviations corresponding to the CVs. Note that one could also apply Gaussians with a non-diagonal variance matrix. Hills are deposited during the simulation in fixed time intervals τ_{G}. Hence, the system is pushed away from regions that were already visited, and minima on the PMF are gradually filled. After enough time has passed, the biasing potential becomes quasi-stationary, fluctuating around the negative PMF, and the dynamics will effectively experience a random walk on the PMF. To ensure smooth convergence, Well-Tempered MtD (WTM) [37] was introduced, which scales down the height of the new Gaussian hills based on the previously deposited potential

$$h_{\mathrm{G}}^{\mathrm{WTM}}(\mathbf{z},t) = h_{\mathrm{G}}\, e^{-\frac{\beta}{\gamma-1}U^{\mathrm{MtD}}(\mathbf{z},t-1)}\,, \tag{2.51}$$

where γ is a parameter called the bias factor. Normal MtD is exactly retrieved for the limit $\gamma \to \infty$. As the scaling factor decreases over time, $h_{\mathrm{G}}^{\mathrm{WTM}}(\mathbf{z},t)$ goes to zero. For MtD and WTM, it has been mathematically proven [96, 97], that the biasing potential eventually converges to the asymptotic solution

$$\begin{aligned} U^{\mathrm{MtD}}(\mathbf{z},t) &= -\left(1-\frac{1}{\gamma}\right) A(\mathbf{z}) + \beta^{-1} \ln \frac{\int \mathrm{d}\mathbf{z}\, e^{-\beta A(\mathbf{z})}}{\int \mathrm{d}\mathbf{z}\, e^{-\beta(A(\mathbf{z})+U^{\mathrm{MtD}}(\mathbf{z},t))}} \\ &= -\left(1-\frac{1}{\gamma}\right) A(\mathbf{z}) + \beta^{-1} \ln \frac{Z}{\widetilde{Z}(t)} \\ &= -\left(1-\frac{1}{\gamma}\right) A(\mathbf{z}) + C(t)\,, \end{aligned} \tag{2.52}$$

with time-dependent constant $C(t)$. Solving Eq. 2.52 for the PMF leads to the unbiased

estimate

$$\begin{aligned} A(\mathbf{z}) &= -\left(\frac{\gamma}{\gamma-1}\right) U^{\mathrm{MtD}}(\mathbf{z},t) + C(t) \\ &= -\left(\frac{\gamma}{\gamma-1}\right) U^{\mathrm{MtD}}(\mathbf{z},t) + \beta^{-1} \ln \frac{Z}{\widetilde{Z}(t)} \,. \end{aligned} \tag{2.53}$$

At this point, it becomes obvious that $C(t)$ again corresponds to the problematic third term of Eq. 2.21, with the only difference that $\widetilde{Z}$ now is time-dependent because of the time-dependence of the biasing potential. Note that the overall estimate of $A(\mathbf{z})$ is time-independent, as the time-dependencies of both terms cancel [98]. While $C(t)$ is typically ignored for calculating the PMF, where one is only interested in relative free energy differences, it can be used to monitor the local quality of convergence of the PMF, as was shown by Tiwary and Parrinello [98]. Additionally, from Section 2.2 we know that it has to be considered for proper reweighting of other observables from the biased to the original probability distribution [94], for which, analogous to Eq. 2.17, one obtains

$$\langle O(\mathbf{x}) \rangle = \left\langle O(\mathbf{x}) e^{\beta(U^{\mathrm{MtD}}(\mathbf{z},t) - C(t))} \right\rangle_{\widetilde{U}} \,. \tag{2.54}$$

In recent years, many methods have been devised for calculating $C(t)$ (*e.g.*, [98–101]), which will not be further discussed here.

Note that in WTM, the PMF is only partially canceled by U^{MtD}, such that instead of a uniform distribution, in the long time limit, one samples from the so-called well-tempered distribution

$$\widetilde{\rho}^{\mathrm{WTM}}(\mathbf{z}) = \frac{e^{-\beta(A(\mathbf{z})-(1-\frac{1}{\gamma})A(\mathbf{z}))}}{\int \mathrm{d}\mathbf{z}\, e^{-\beta(A(\mathbf{z})-(1-\frac{1}{\gamma})A(\mathbf{z}))}} = \frac{e^{-\beta\frac{1}{\gamma}A(\mathbf{z})}}{\int \mathrm{d}\mathbf{z}\, e^{-\beta\frac{1}{\gamma}A(\mathbf{z})}} = \frac{[\rho(\mathbf{z})]^{\frac{1}{\gamma}}}{\int \mathrm{d}\mathbf{z}\, [\rho(\mathbf{z})]^{\frac{1}{\gamma}}} \,. \tag{2.55}$$

Only for the limit $\gamma \to \infty$, one again samples from a uniform distribution, as in normal MtD. Historically, instead of γ often the bias temperature ΔT was used as parameter, which is related to γ via $\gamma = (T + \Delta T)/T$. Therefore, one might view the well-tempered distribution as sampling the CVs at higher temperatures while keeping the PMF fixed.

One conceptual drawback of the metadynamics potential is that it has to be carefully parametrized. If the bias potential adapts too slowly, sampling will be inefficient. On the other hand, one also has to make sure that its growth rate is not too high. In this case, the system is pushed out of equilibrium by the harsh and discontinuous repulsion from its current state, which will introduce artifacts [102]. This is especially critical for reweighting via Eq. 2.54. Hence, for such a non-static bias, the initial non-adiabatic part of the trajectory where the biasing potential is still changing drastically often has to be discarded, and reweighting is only performed for the remaining part of the trajectory. In this regard, a significant improvement over MtD/WTM was recently introduced by Invernizzi and Parrinello with the on-the-fly probability enhanced sampling (OPES) method [55], where instead of the PMF the probability density itself is estimated by the kernel density

approximation

$$\widetilde{\rho}(\mathbf{z},t) = \frac{\sum_{t=\tau_\mathrm{G},2\tau_\mathrm{G},\ldots} w_t\, G(\mathbf{z},\mathbf{z}_t)}{\sum_{t=\tau_\mathrm{G},2\tau_\mathrm{G},\ldots} w_t} \tag{2.56}$$

with Gaussian kernels G and weights $w_t = e^{\beta U^{\mathrm{OPES}}(\mathbf{z}_t,t-1)}$. The biasing potential U^{OPES} is obtained from $\widetilde{\rho}(\mathbf{z},t)$ as

$$U^{\mathrm{OPES}}(\mathbf{z},t) = \left(1-\frac{1}{\gamma}\right)\beta^{-1}\log\left(\frac{\widetilde{\rho}(\mathbf{z},t)}{Z_t}+\epsilon\right) \tag{2.57}$$

where the normalization factor Z_t is given by the integration over the currently explored configuration space

$$Z_t = \frac{1}{|\Omega_t|}\int_{\Omega_t}\widetilde{\rho}(\mathbf{z},t)\mathrm{d}\mathbf{z}\,. \tag{2.58}$$

The term ϵ of Eq. 2.57 ensures that the logarithm is always defined.

By choosing $\epsilon = e^{-\beta\Delta E/(1-1/\gamma)}$, the parameter ΔE defines an upper bound to the bias potential, such that only transitions crossing barriers that are smaller or equal to ΔE are efficiently induced. Note that already the first kernel gives rise to a bias potential that is in the order of ΔE, such that OPES provides a very fast initial build-up of the biasing potential, which in later stages is gradually refined. For this reason, the OPES potential is quasi-stationary, and reweighting is possible via Eq. 2.17 without explicitly taking into account any time-dependence as in Eq. 2.54. Additionally, the chosen height of the Gaussian hills only changes the normalization, and the variance of the Gaussian hills can be obtained adaptively from the trajectory [103], such that ΔE and τ_G are the only remaining free parameters, making OPES much easier to parameterize than its predecessors.

2.4.3 Adaptive Biasing Force Methods

An alternative to adaptive potential methods are adaptive biasing force (ABF) methods, first introduced by Darve and Pohorille [38], where instead of adding an artificial biasing potential, the system forces are directly modified to ensure ergodic sampling. Note that although it might seem arbitrary at first glance, there is a fundamental difference between biasing the potential or force. Namely, while potentials and probability distributions are global properties, gradients are defined locally. Therefore, to build an efficient biasing potential, knowledge of a wide range of the configuration space is required, while building the biasing force only requires local knowledge, which in principle enables faster adaptation [42]. Additionally, from a conceptual standpoint, ABF methods have a high appeal due to their practical simplicity, requiring minimal user intervention.

The idea of ABF is that applying the negative gradient of the PMF as a biasing force exactly cancels associated free energy barriers, such that the average force acting on the CV is zero. Thus, the system only experiences random fluctuations of the environment, which leads to a random walk on the flattened PMF and, in the long term, uniform sampling.

To estimate the gradient of the PMF, one relies on calculating the ensemble average of instantaneous force samples

$$\nabla_{\mathbf{z}} A(\mathbf{z}) = -\left\langle \mathbf{F}_{\xi}(\mathbf{x}) \right\rangle_{\xi(\mathbf{x})=\mathbf{z}} . \tag{2.59}$$

Hence, one just needs to find an algorithm for estimating Eq. 2.59 on-the-fly from the trajectory. Assuming ergodicity orthogonal to the CV, this is typically done on a grid with bin width $\Delta\mathbf{z}$ from M_k currently available samples in bin k by the simple unweighted average

$$\overline{\mathbf{F}}_z(M_k, k) = \frac{1}{M_k} \sum_{i=1}^{M_k} \mathbf{F}_i^k , \tag{2.60}$$

with instantaneous force samples $\mathbf{F}_i^k$. Care has to be taken in the initial phase of the simulation, where Eq. 2.60 is a poor estimate of the true gradient of the PMF, and large fluctuations in the running average might drive the system away from equilibrium. Therefore, the biasing force initially has to be slowly scaled up, *e.g.*, by a linear ramp where the force is proportional to M_k/M_{full} if $M_k < M_{\text{full}}$ [39]. Hence, apart from the grid resolution, the ABF method only depends on M_{full} as a single empirical parameter, which controls how fast the bias force is scaled up.

The remaining open question is how to compute force samples $\mathbf{F}_i^k$. To derive the original expression for the instantaneous force, one needs to explicitly transform Cartesian coordinates with a set of $N-d$ functions $\mathbf{q} = (q_1, q_2, ..., q_{N-d})$, such that $(\xi_1, ..., \xi_d, q_1, ..., q_{N-d})$ is a complete set of generalized coordinates together with CVs. The Jacobian J of the transformation from Cartesian to generalized coordinates is

$$J := \begin{pmatrix} \frac{\partial \xi_1}{\partial x_1} & \cdots & \frac{\partial \xi_1}{\partial x_N} \\ \cdots & \cdots & \cdots \\ \frac{\partial \xi_d}{\partial x_1} & \cdots & \frac{\partial \xi_d}{\partial x_N} \\ \frac{\partial q_1}{\partial x_1} & \cdots & \frac{\partial q_1}{\partial x_N} \\ \cdots & \cdots & \cdots \\ \frac{\partial q_{N-d}}{\partial x_1} & \cdots & \frac{\partial q_{N-d}}{\partial x_N} \end{pmatrix} = \begin{pmatrix} J_{\xi} \\ J_q \end{pmatrix} . \tag{2.61}$$

From this, Carter *et al.* [104] obtain for the instantaneous force the expression

$$F_{\xi_i} = -\nabla_{\mathbf{x}} U(\mathbf{x}) \cdot \frac{\partial \mathbf{x}}{\partial \xi_i} + \beta^{-1} \frac{\partial \ln |J|}{\partial \xi_i} \tag{2.62}$$

where the force is projected into the direction of the *inverse gradient* $\frac{\partial \mathbf{x}}{\partial \xi_i}$ of CVs. In plain words, the inverse gradient corresponds to the vector along which a change in ξ is propagated in Cartesian coordinates, keeping other generalized coordinates $\mathbf{q}$ constant. This is the reason why the explicit coordinate transformation is needed. Note that the concrete choice of generalized coordinates is arbitrary, but yields a uniquely defined mean force [42]. Unfortunately, the requirement of a complete coordinate transformation and the computation

of second derivatives in the form of the Jacobian derivative render the implementation of Eq. 2.62 highly cumbersome.

To address these issues, the freedom of choice of generalized coordinates was harnessed by the idea of Den Otter [105] to replace the inverse gradient with an arbitrary vector field. Generalized to a multidimensional case and in the presence of a set of holonomic constraints of the form $\sigma_k(\mathbf{x}) = 0$, the vector field $\mathbf{v}_i$ just has to satisfy

$$\mathbf{v}_i \cdot \nabla_{\mathbf{x}} \xi_j = \delta_{ij} \,, \tag{2.63}$$

$$\mathbf{v}_i \cdot \nabla_{\mathbf{x}} \sigma_k = 0 \,, \tag{2.64}$$

for all j and k, meaning that it has to be orthogonal to the gradients of all other CVs and constraints, respectively [106]. From this, a much easier-to-compute expression of the instantaneous force can be obtained.

$$F_{\xi_i} = -\nabla_{\mathbf{x}} U(\mathbf{x}) \cdot \mathbf{v}_i + \beta^{-1} \mathrm{div}\,(\mathbf{v}_i) \tag{2.65}$$

Still, second derivatives are required in the form of the divergence of the vector field, but one can take advantage of the relative freedom of choice of $\mathbf{v}_i$ to make it practical. Usually one chooses

$$\mathbf{v}_i = \frac{\nabla_x \xi_i}{|\nabla_x \xi_i|^2} \,, \tag{2.66}$$

which is always valid, but certainly not an optimal choice for every CV [42].

For a one-dimensional CV, a time-dependent estimate of the PMF in bin k can be obtained from numerical integration, *e.g.*, using the rectangular rule

$$A_t(k) = -\sum_{i=1}^{k} \overline{F}_z(M_k, k) \Delta z + C \,. \tag{2.67}$$

Note that more sophisticated multidimensional integration schemes are available [39, 107]. Hence, other than in static or adaptive biasing potential methods, the constant C originates from the integration and is generally unknown. Therefore, only relative free energies can be recovered from ABF simulations, while general reweighting of the biased probability distribution remains elusive.

Other formulations of ABF methods exist, *e.g.*, using time derivatives [39] or projected forces [108], which are not further discussed here. Instead, in the following section, the extended-system ABF is introduced, which can lift all technical requirements on CVs and, as will be shown, combines the advantages of all the previously discussed sampling approaches.

2.4.4 Extended-System Dynamics

Above, static and adaptive importance sampling strategies based on biasing potentials and forces were discussed, all of them with their unique strengths and weaknesses. Now, the

extended-system dynamics will be presented, which combines ideas from all of them and is applied throughout this thesis. In the literature, this method was first introduced in the context of extended-system ABF (eABF) in a book by Leliévre *et al.* [40], but a similar idea was also reported under the name *dynamic reference restraining* by Yang *et al.* in the orthogonal space tempering (OST) method [41].

Generally, the term extended-system dynamics refers to methods where Langevin dynamics simulations are performed for the extended system $(\mathbf{x}, \lambda)$. The additional, nonphysical degrees of freedom $\lambda \in \mathbb{R}^d$ are harmonically coupled to the CVs such that the extended potential energy function reads

$$U^{\text{ext}}(\mathbf{x}, \lambda) = U(\mathbf{x}) + \sum_{i=1}^{d} \frac{1}{2\beta\sigma_i^2} \left(\xi_i(\mathbf{x}) - \lambda_i\right)^2 , \tag{2.68}$$

where $\sigma_i = \sqrt{1/\beta k_i}$ is the thermal coupling width, which is related to the coupling force constant k_i. Thus, for sufficiently strong coupling (small σ_i), one can achieve improved sampling of CV i by only biasing λ_i. As each λ_i is, by construction, orthogonal to all other degrees of freedom, Eq. 2.63 and Eq. 2.64 are always fulfilled. Therefore, the ABF can trivially be applied to the fictitious particles, with instantaneous force samples given by

$$F_{\lambda_i} = \frac{\partial}{\partial \lambda_i} U^{\text{ext}}(\mathbf{x}, \lambda) = \frac{1}{\beta\sigma_i^2} \left(\xi_i(\mathbf{x}) - \lambda_i\right) . \tag{2.69}$$

However, the true strength of extended-system dynamics lies in the decoupling of the problem of sampling acceleration from the problem of reweighting and free energy computation. Recalling the previously discussed adaptive sampling approaches, which may be based on biasing potentials of forces, it is striking that the free energy estimates always explicitly depend on the time-dependent bias. Hence, the convergence of free energy estimates is directly linked to the convergence of the adaptive bias. This is no longer the case in extended-system dynamics, where the coupling term can act as a buffer such that, independent of the biasing algorithm, the physical system only experiences the time-independent coupling potential. This key insight is harnessed in many of the developments presented in this thesis.

In terms of sampling acceleration, one gains high flexibility in the choice of algorithm. As shown by Fu *et al.* in the MtD-eABF [44] and WTM-eABF [45] methods, particularly efficient are combinations of complementary sampling strategies. The MtD/WTM pushes away from already visited regions, while the ABF flattens free energy barriers along the way, together allowing for highly efficient adaptation to the underlying PMF.

In terms of reweighting, in Publication **I** we show how importance sampling techniques like the MBAR can be applied to extended-system dynamics to recover the full statistical information. In the literature, in the spirit of conventional ABF only thermodynamic integration-based estimators were suggested, namely UI [41, 109] and the corrected z-averaged restraint (CZAR) [43]. Both of them yield fast and accurate on-the-fly estimates of the unbiased PMF, but, like previously discussed, are unable to recover other ensemble properties. Comparing Eq. 2.47 and Eq. 2.68, one can observe the similarity of the

extended system potential to the US potentials, both using a harmonic coupling term. Hence, one might alternatively view the extended-system dynamics as a special case of US with moving windows. Therefore, splitting the continuous trajectory of the extended system into biased states with constant $\lambda_i = \lambda_i^j$, one can obtain a similar set of K biased windows as would be obtained by performing K US simulations for each λ_i^j. In practice, states are defined by collecting frames with $\lambda_i^j - \Delta\lambda_i/2 \leq \lambda_i < \lambda_i^j + \Delta\lambda_i/2$ on a grid with bin width $\Delta\lambda_i$. Hence, in one window λ_i is approximated as always remaining fixed at bin centers λ_i^j, which, as we show numerically in Publication **I**, does not introduce any error if $\Delta\lambda_i \leq \sigma_i$. Hence, the MBAR estimator applies to extended-system dynamics in the same manner as to US, and the problem of accurately reweighting extended-system dynamics is solved. Note that, unlike in US, biasing functions do not have to be placed manually, and sufficient overlap of windows is always ensured as they are built from one or multiple continuous trajectories. Additionally, unlike for the direct reweighting from MtD or WTM, as the adaptive biasing potential does not enter the post-processing in any form, it is not necessary to cut the initial phase of the trajectory where the biasing potential still changes drastically.

In Publication **II**, we show that combining extended-system dynamics with adaptive PCVs is particularly useful. In the literature, adaptive PCVs are often combined with MtD/WTM, *e.g.*, by Branduardi *et al.* [51] or Leines *et al.* [52]. For these approaches, MtD/WTM is run along a continuously changing PCV, such that the PCV converges to the minimum free energy path (MFEP), before the biasing potential converges to an unbiased estimate of the PMF along the quasi-static final MFEP. We show that applying WTM-eABF instead has two big advantages. Firstly, due to the application of two complementary biasing strategies, adaptation of the bias to changes in the PCV is much faster, and secondly, one can use the MBAR to easily project the sampled data onto any reaction coordinate (or path) of choice. Hence, instead of waiting for convergence of the biasing potential, one can map the full trajectory to the final PCV to obtain an accurate PMF. The only problem with this approach is that extended-system dynamics are not stable in combination with discontinuous CVs. This is prohibitive to its application in combination with PCVs, as they can show discontinuities due to path short-cutting or path updates, which cause the sudden appearance of large coupling forces. Therefore, we propose a stabilization algorithm, which at time step $t-1$ corrects positions of λ_i for the discontinuities before integrating to t. The position of λ_i is updated according to

$$\lambda_{\mathrm{stable},i}^{t-1} = \begin{cases} \xi_i(\mathbf{x}^t) + (\lambda_i^{t-1} - \xi_i(\mathbf{x}^{t-1})) & \text{if } |\xi_i(\mathbf{x}^t) - \xi_i(\mathbf{x}^{\mathrm{t}-1})| > \Delta\xi_i^{\max} \\ \lambda_i^{t-1} & \text{otherwise}\,, \end{cases} \tag{2.70}$$

where $\Delta\xi_i^{\max}$ is the maximum allowed step size of ξ_i before considered as discontinuous.

Finally, in Publication **V**, we propose to replace the WTM potential in WTM-eABF with the more recent OPES, which yields the new OPES-eABF sampling method. While OPES-eABF inherits all the above-discussed benefits of WTM-eABF, the quasi-static nature of the OPES potential introduces additional advantages. Most importantly, if the WTM potential is not parametrized carefully, WTM-eABF can induce artifacts by too

harsh repulsion from the current state, or be inefficient by too slow buildup of the bias potential, similar to stand-alone WTM. This is not the case for OPES, where only a single intuitive parameter has to be manually set, while the rest can be obtained from the trajectory itself, *e.g.*, using an adaptive bandwidth algorithm [103]. Additionally, we show that OPES-eABF is much more robust against user parameters, as inefficiency by too slow buildup of the bias potential or artifacts by too harsh repulsion from the current state cannot occur. Hence, as the coupling widths σ_i of the extended system are the remaining critical parameters of OPES-eABF simulations, we derive an automatic algorithm for obtaining suitable coupling widths based on a short initial MD. The performance of OPES-eABF is benchmarked against WTM, WTM-eABF, and OPES on three test systems, where OPES-eABF is always the most efficient if appropriate CVs are chosen. Additionally, challenging cases of systems with very high ΔA (where a weakness of OPES is identified that is removed by OPES-eABF), poor CVs, and the occurrence of parallel reaction pathways are analyzed.

Chapter 3

Publications

3.1 Publication I: Statistically optimal analysis of the extended-system adaptive biasing force (eABF) method

Abstract: The extended-system adaptive biasing force (eABF) method and its newer variants offer rapid exploration of the configuration space of chemical systems. Instead of directly applying the ABF bias to collective variables, they are harmonically coupled to fictitious particles, which separates the problem of enhanced sampling from that of free energy estimation. The prevalent analysis method to obtain the potential of mean force (PMF) from eABF is thermodynamic integration. However, besides the PMF, most information is lost as the unbiased probability of visited configurations is never recovered. In this contribution, we show how statistical weights of individual frames can be computed using the Multistate Bennett's Acceptance Ratio (MBAR), putting the post-processing of eABF on one level with other frequently used sampling methods. In addition, we apply this formalism to the prediction of nuclear magnetic resonance shieldings, which are very sensitive to molecular geometries and often require extensive sampling. The results show that the combination of enhanced sampling by means of extended-system dynamics with the MBAR estimator is a highly useful tool for the calculation of ensemble properties. Furthermore, the extension of the presented scheme to the recently published Gaussian-accelerated molecular dynamics eABF hybrid is straightforward and approximation free.

Reproduced from

A. Hulm; J. C. B. Dietschreit; C. Ochsenfeld. "Statistically optimal analysis of the extended-system adaptive biasing force (eABF) method." *J. Chem. Phys.* **2022**, 157, 024110.
URL: `https://pubs.aip.org/aip/jcp/article/157/2/024110/2841454`,

with the permission of AIP Publishing.

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

Statistically optimal analysis of the extended-system adaptive biasing force (eABF) method

Cite as: J. Chem. Phys. 157, 024110 (2022); doi: 10.1063/5.0095554
Submitted: 11 April 2022 • Accepted: 14 June 2022 •
Published Online: 13 July 2022

Andreas Hulm,[1] Johannes C. B. Dietschreit,[1,2] and Christian Ochsenfeld[1,3,a)]

AFFILIATIONS

[1] Chair of Theoretical Chemistry, Department of Chemistry, University of Munich (LMU), Butenandtstr. 7, D-81377 München, Germany
[2] Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
[3] Max Planck Institute for Solid State Research, Heisenbergstr. 1, D-70569 Stuttgart, Germany

[a)] Author to whom correspondence should be addressed: christian.ochsenfeld@uni-muenchen.de

ABSTRACT

The extended-system adaptive biasing force (eABF) method and its newer variants offer rapid exploration of the configuration space of chemical systems. Instead of directly applying the ABF bias to collective variables, they are harmonically coupled to fictitious particles, which separates the problem of enhanced sampling from that of free energy estimation. The prevalent analysis method to obtain the potential of mean force (PMF) from eABF is thermodynamic integration. However, besides the PMF, most information is lost as the unbiased probability of visited configurations is never recovered. In this contribution, we show how statistical weights of individual frames can be computed using the Multistate Bennett's Acceptance Ratio (MBAR), putting the post-processing of eABF on one level with other frequently used sampling methods. In addition, we apply this formalism to the prediction of nuclear magnetic resonance shieldings, which are very sensitive to molecular geometries and often require extensive sampling. The results show that the combination of enhanced sampling by means of extended-system dynamics with the MBAR estimator is a highly useful tool for the calculation of ensemble properties. Furthermore, the extension of the presented scheme to the recently published Gaussian-accelerated molecular dynamics eABF hybrid is straightforward and approximation free.

I. INTRODUCTION

Exploring high-dimensional potential energy surfaces of complex molecular systems poses a great challenge, as it often involves extensive sampling of regions separated by free energy barriers.[1,2] Molecular dynamics (MD) or Monte Carlo (MC) simulations usually remain trapped in free energy wells, leading to non-ergodic sampling. Therefore, it is paramount to use importance-sampling algorithms to improve sampling efficiency.[3–8] Many of these methods rely on the definition of a low-dimensional set of collective variables (CVs) that are sufficient to discriminate between the states of interest. Typically, one to three-dimensional variables are chosen, because of the massive growth of sampling needed in higher-dimensional space, known as curse of dimensionality,[9] which, in turn, leads to immense computational cost.

The adaptive biasing force (ABF)[4,10,11] method is one such algorithm that achieves uniform sampling along the CV. Here, a running estimate of the mean force is applied as bias. Subsequently, the potential of mean force (PMF) can directly be obtained from thermodynamic integration (TI)[12] of the biasing force. Despite its outstanding stability and beneficial formal convergence properties, the application of ABF has been limited by stringent technical requirements on CVs to be usable for ABF.[13,14] These could be lifted by Lesage *et al.*[15] in the extended-system ABF (eABF) method. Instead of directly applying a biasing force to the physical system, each CV is coupled to a fictitious particle, extending the system.

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

In addition to removing all technical constraints on the CV, this framework separates the problem of sampling from free energy estimation. The resulting flexibility in the choice of bias force applied to the fictitious particle has been exploited by combining ABF and metadynamics,[16–18] termed well-tempered metadynamics extended-system ABF (WTM-eABF),[19,20] a highly potent enhanced sampling scheme, which stands out due to its efficiency and robustness.[21] In line with the original ABF method, the PMF is typically obtained via thermodynamic integration of the estimated free energy gradient.[15,22] Very recently, Gaussian-accelerated WTM-eABF (GaWTM-eABF) was introduced by Chen *et al.*,[23] where WTM-eABF is coupled with Gaussian-accelerated MD (GaMD),[24] which helps to overcome barriers orthogonal to the CV.

Many CV-based advanced sampling techniques not only offer access to the PMF but also provide access to the unbiased probability of a configuration. The original weight of a trajectory frame can be obtained, e.g., for umbrella sampling[3] by means of Multistate Bennett's Acceptance Ratio (MBAR).[25–27] While formalisms have been derived for metadynamics by which one regains the unbiased probability of a visited configuration,[28–31] for the classes of ABF methods, such an algorithm is missing, barring users from computing unbiased ensemble averages.

In this contribution, we close this gap for the aforementioned family of eABF-based algorithms. We show that MBAR can be applied to extend the scope of eABF-based enhanced sampling from pure PMF calculation to the calculation of ensemble averages in general. By solving the MBAR equations, Boltzmann weights of individual configurations are obtained, which enables the direct reweighting of any coordinate-dependent observable of interest. Because no histograms or other form of finite numeric grids are introduced, MBAR also provably yields the lowest variance free energy estimate.[26,32]

In the following, we will first revise the theory behind enhanced sampling using the ABF and eABF algorithms. We then show how MBAR can be applied to the statistical analysis of eABF simulations and demonstrate through several numerical examples how eABF-trajectories are unbiased by means of MBAR. Finally, we showcase for a real chemical system how advanced sampling and unbiased averages can be useful for calculating equilibrium nuclear magnetic resonance (NMR) shieldings. Due to their sensitivity to the molecular geometry, the accurate prediction of NMR shieldings is particularly challenging and often requires extensive sampling of molecular configurations.[33,34] Further, we demonstrate how by reweighting, eABF-biased simulations can even yield CV-resolved NMR shieldings.

II. THEORY

We consider a system consisting of N_a atoms with Cartesian coordinates $\mathbf{x}^T = (x_1, \ldots, x_{3N_a})$, whose physical distribution for the potential energy surface (PES) $U(\mathbf{x})$ is the Boltzmann-distribution $\rho(\mathbf{x})$ at absolute temperature T.

We assume that a reaction coordinate can be represented by a collective variable (CV) $\xi(\mathbf{x}) : \mathbb{R}^{3N_a} \to \mathbb{R}$, which has the marginal probability distribution,

$$\rho(z) = \int \delta[\xi(\mathbf{x}) - z]\rho(\mathbf{x})\ \mathrm{d}\mathbf{x} = \langle \delta[\xi(\mathbf{x}) - z] \rangle, \tag{1}$$

where $\langle\rangle$ denotes the ensemble average and $\delta[x]$ the Dirac delta distribution.[1] The PMF $A(z)$, i.e., free energy profile, along a CV is usually defined as

$$A(z) = -\beta^{-1} \ln \rho(z), \tag{2}$$

where $\beta = (k_B T)^{-1}$ and k_B is the Boltzmann constant. Note that the shape of the PMF is not gauge invariant against the choice of CV function ξ. Therefore, features of the PMF like minima or transition barriers have no CV-independent meaning.[35] This is, however, not problematic for the calculation of free energy differences as long as the CV can discriminate between configurations of states R and P, because the CV-dependent local features are integrated out to obtain the probabilities of the system occupying either state R or P,[36]

$$\Delta A_{R\to P} = -\beta^{-1} \ln \frac{\int_P \rho(z)\ \mathrm{d}z}{\int_R \rho(z)\ \mathrm{d}z}. \tag{3}$$

In the case of chemical reactions, there often exist several metastable states that are separated by free energy barriers much larger than the thermal energy $k_B T$,[1] which are not surmounted within the typical time scale of molecular dynamics simulations. For the exploration of chemical space, it is, therefore, crucial to accelerate sampling, e.g., by means of ABF, which we will revise in the following.

A. Adaptive biasing force

The ABF algorithm enhances sampling of low probability states by introducing a force $\frac{\partial A(z)}{\partial z} \nabla \xi(\mathbf{x})$ that cancels the unbiased mean force and leads to uniform sampling along the CV.[11]

Often, one has no prior knowledge of the free energy derivative, and, thus, ABF uses an on-the-fly estimate of the mean force. For this purpose, the CV is divided in L equally spaced bins of size dz. The approximation of the bias force $\overline{F}(z_i)$ in bin i is the average of collected force samples F_i^μ,[11]

$$\overline{F}(z_i) = \frac{1}{N_i} \sum_{\mu=1}^{N_i} F_i^\mu. \tag{4}$$

Assuming ergodicity, $\overline{F}(z_i)$ will converge to the z-conditioned ensemble average $\langle F \rangle_{\xi(\mathbf{x})=z_i}$ for a sufficiently large sample size N_i.

Taking a more general perspective, this corresponds to biasing the system dynamics with a time-dependent approximation of the free energy surface projected onto a low-dimensional collective variable, $A_t(\mathbf{x}) \circ \xi(\mathbf{x})$.[11] Although this approximation is very poor at the beginning of a simulation and the bias based on the running average has to be downscaled in the low-sampling regime, it systematically improves over the course of the simulation. This biasing strategy connects ABF to metadynamics and its well-tempered variant, which follow a similar idea, but push the system away from visited regions with a time-dependent repulsive potential. However, both algorithms change the bias potential and, thereby, also the weights of each frame over time. Recovering the statistical information of the underlying unbiased system is, therefore, difficult. While reweigthing schemes have been proposed for metadynamics,[28–31] to our knowledge none has been reported for ABF simulations.

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

Free energy surfaces can directly be recovered from $\overline{F}(z_i)$ by TI.[10] Despite its appealing theoretical simplicity and efficiency, the application of ABF is hindered by some practical limitations: When using multiple CVs, they are required to be mutually orthogonal as well as orthogonal to any additional constraints on the system. Additionally, a Jacobian term depending on the second derivative of the coordinates has to be calculated.[11]

B. Extended-system ABF

To circumvent the technical requirements of ABF, Lesage *et al.*[15] proposed the extended-system ABF (eABF) method, where a fictitious particle λ is coupled to the CV by a harmonic spring. The extended system $(\mathbf{x}, \lambda)$ evolves according to the potential

$$U_{\text{ext}}(\mathbf{x}, \lambda, t) = U(\mathbf{x}) + \frac{1}{2\beta\sigma^2}(\xi(\mathbf{x}) - \lambda)^2 + U_b(\lambda, t), \tag{5}$$

with thermal coupling width to the extended system σ. The key intuition behind this choice of potential is that in the tight coupling (low σ) limit, efficient sampling of λ will result in efficient sampling of ξ. Therefore, it is sufficient to bias the dynamics of λ to accelerate sampling along ξ with any choice of bias potential $U_b(\lambda, t)$. This opens up great flexibility for the choice of the importance-sampling algorithm. As each CV is coupled to its own fictitious particle, orthogonality conditions are fulfilled by construction, and the application of ABF is trivial. An especially useful variant is the combination of eABF with (Well-Tempered) Metadynamics (*meta*-eABF, WTM-eABF), which offers rapid exploration of free energy surfaces.[19,20] An asymptotically unbiased estimator of $A(z)$, which is independent of $U_b(\lambda, t)$, can be derived by correcting the free energy gradient, obtained from the biased distribution $\rho^B(z)$ of the extended system, with the average harmonic force on z,

$$\frac{\partial A(z)}{\partial z} = -\beta^{-1}\frac{\partial \ln \rho^B(z)}{\partial z} + \frac{1}{\beta\sigma^2}(\langle\lambda\rangle_z - z). \tag{6}$$

This estimator is called *corrected z-averaged restraint* (CZAR) and can be calculated from the trajectories of ξ and λ alone.[15] In the spirit of the original ABF method, the CZAR estimator yields the free energy gradient, which is integrated using TI.

However, none of these algorithms are able to accelerate sampling of degrees of freedom orthogonal to the CV. In recent years, great effort went into the development of CVs that contain all slow degrees of freedom to circumvent this inherent problem, for example, by means of machine learning.[37–41] Chen *et al.*[23] recently reported an alternative approach by combining WTM-eABF with Gaussian-accelerated MD (GaWTM-eABF). The GaMD potential, which is only a function of the potential energy itself, reduces free energy barriers along all degrees of freedom. Then WTM-eABF gives sampling a specific direction on this modified potential. The full GaWTM-eABF potential reads

$$U_{\text{ext}}(\mathbf{x}, \lambda, t) = U(\mathbf{x}) + \Delta U(\mathbf{x}) + \frac{1}{2\beta\sigma^2}(\xi(\mathbf{x}) - \lambda)^2 + U_b(\lambda, t), \tag{7}$$

where the GaMD potential $\Delta U(\mathbf{x})$ is given by

$$\Delta U(\mathbf{x}) = \begin{cases} \frac{1}{2}k_0(E - U(\mathbf{x}))^2, & U(\mathbf{x}) < E, \\ 0, & U(\mathbf{x}) \geq E. \end{cases} \tag{8}$$

Parameters $E = U_{\max}$ and $k_0 = \min\left(1, \frac{\sigma_0}{\sigma_U}\frac{U_{\max} - U_{\min}}{U_{\max} - U_{\text{avg}}}\right)$ are obtained from an equilibration calculation, the only free parameter being σ_0, which defines the upper bound for the standard deviation of $\Delta U(\mathbf{x})$. $U_{\max}$, $U_{\min}$, U_{avg}, and σ_U denote the maximum, minimum, average, and standard deviation of $U(\mathbf{x})$ in the equilibration stage, respectively. The PMF is obtained by summing the contributions of both biases,

$$A(z) \approx -\beta^{-1}\ln \rho^B(z) + \frac{1}{\beta\sigma^2}\int(\langle\lambda\rangle_z - z)dz - C_1(z) - \frac{1}{2}\beta C_2(z), \tag{9}$$

where the CZAR estimate is used for WTM-eABF and the GaMD bias is approximated by a second order cumulant expansion with average and variance of ΔU, C1 and C2, respectively. This approximation requires the distribution of ΔU to be Gaussian-shaped, which can be ensured by an appropriate choice of σ_0.

For all methods discussed thus far, the unbiased Boltzmann weight of individual frames is lost. In Sec. II C, we will show how the MBAR estimator can be applied to recover statistical weights and calculate any ensemble average of interest from all eABF-based methods.

C. Statistically optimal analysis of eABF-biased trajectories

To gain a statistically optimal analysis of eABF-biased trajectories, we will make use of MBAR.[26] In eABF simulations, the physical system is effectively sampled under harmonic restraint for each value of the fictitious particle λ. We take advantage of this to split the extended-system $U_{\text{ext}}(\mathbf{x}, \lambda, t)$, where λ is a dynamic variable, into M biased systems $U_i(\mathbf{x})$ with constant $\lambda = \lambda_i$. Configurations of the continuously sampled trajectory are separated *post hoc* into M biased states. The modified potential in window i reads

$$U_i(\mathbf{x}) = U(\mathbf{x}) + \Delta U(\mathbf{x}) + \frac{1}{2\beta\sigma^2}(\xi(\mathbf{x}) - \lambda_i)^2, \tag{10}$$

where the GaMD potential $\Delta U(\mathbf{x})$ is only added for GaWTM-eABF. The biased probability distribution $\rho^B(z)$ of the eABF trajectory is converted to a mixture distribution of M overlapping, λ-conditioned, biased distributions $\rho^B(z|\lambda_i)$ (see also Sec. S1 of the supplementary material).

Note that one could obtain a similar set of biased windows by means of umbrella sampling[3] with M harmonic biasing potentials. However, this has several disadvantages: Overlap of biased probability distributions $\rho(z|\lambda_i)$ has to be ensured by hand prior to the simulation with an appropriate choice of λ windows. This can be far from trivial and often requires significant experience from the user, whereas eABF ensures a uniform distribution of $\rho(z)$ automatically. Moreover, when using eABF, the physical system diffuses between states, which does not happen in standard umbrella simulations that have a static λ_i value. For systems with parallel valleys, this freedom of motion helps overcoming orthogonal barriers,[11] especially when paired with multiple-walker techniques[42,43] or GaMD.[23]

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

To obtain an estimate of the reduced free energy differences f_i between λ-states, one can solve the MBAR equations,[26]

$$e^{-\beta\hat{f}_i} = \sum_{n=1}^{N} \frac{e^{-\beta(U_i-U)(\mathbf{x}_n)}}{\sum_{m=1}^{M} N_m e^{\beta\hat{f}_m-\beta(U_m-U)(\mathbf{x}_n)}}, \tag{11}$$

with N_m samples in the mth λ-state such that $N = \sum_{m=1}^{M} N_m$. Because we are only interested in relative free energies, $\hat{f}_1$ is typically fixed to zero. Note that unlike before [compare Eq. (9)], no Gaussian distribution of ΔU is required.

By solving the MBAR equations iteratively, we obtain the weights $W(\mathbf{x}_n)$ of individual samples as

$$W(\mathbf{x}_n) = \frac{\mathcal{N}}{\sum_{m=1}^{M} N_m e^{\beta\hat{f}_m-\beta(U_m-U)(\mathbf{x}_n)}}, \tag{12}$$

where $\mathcal{N}$ is a normalization constant to ensure that $\sum_n W(\mathbf{x}_n) = 1$. As noted by Shirts and Ferguson,[27] it is no longer relevant which sample belongs to which state as the normalized probability of each sample is known. An unbiased equilibrium ensemble average of a position-dependent operator $O(\mathbf{x}_n)$ from all collected samples is obtained via reweighting,

$$\langle O\rangle = \sum_{n=1}^{N} W(\mathbf{x}_n) O(\mathbf{x}_n). \tag{13}$$

The unbiased probability density $\rho(z)$ (and therefore the PMF) along some CV can be obtained by calculating the expectation of the indicator function[27]

$$\rho(z) = \sum_{n=1}^{N} W(\mathbf{x}_n) I(z, \Delta z, \mathbf{x}_n) \Delta z^{-1}, \tag{14}$$

where

$$I(z, \Delta z, \mathbf{x}_n) = \begin{cases} 1, & \xi(\mathbf{x}_n) \in \left[z - \frac{\Delta z}{2}, z + \frac{\Delta z}{2}\right], \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

with bin centers z_i and bin width Δz. Note that the indicator function can potentially map to any K-dimensional CV (which might be different from the original bias CVs). Similarly, free energy differences are directly computed from weights $W(\mathbf{x}_n)$, instead of numeric integration of $\rho(z)$ [compare Eq. (3)],

$$\Delta A_{R\to P} = -\beta^{-1} \ln \frac{\sum_{x_n\in P} W(\mathbf{x}_n)}{\sum_{x_n\in R} W(\mathbf{x}_n)}, \tag{16}$$

where R and P denote two metastable states divided by a free energy barrier.

III. COMPUTATIONAL DETAILS

All sampling algorithms and numerical simulations are implemented, performed, and analyzed using NumPy. PMFs are calculated by post-processing the MD trajectories of ξ and λ. The TI of the CZAR estimate is computed with the trapezoid rule. The MBAR equations are solved self-consistently and the PMF is computed as a combination of Eqs. (2) and (14). The starting guess for all $\beta\hat{f}_i$'s is zero and $\hat{f}_1$ is set to zero after every cycle. Convergence is reached when the largest change of $\beta\hat{f}_i$ compared to the last cycle drops under 10^{-6}. Subsamples in λ-windows with window size $\Delta\lambda$ are obtained by selecting all frames where $\lambda \in [\lambda - \Delta\lambda/2, \lambda + \Delta\lambda/2]$ from the full biased trajectory. Error bars for both estimators are calculated via bootstrapping.[32,44] For this purpose, the data are resampled 100 times with replacement and equilibrium ensembles are calculated for each bootstrap sample. A generalized Python class for various adaptive sampling methods and their analysis methods is available at https://github.com/ochsenfeld-lab/adaptive_sampling; the numerical examples from this study can be found there as well.

A. Numerical tests in a 2D potential

For numerical tests, we consider a single particle of mass 10 a.u. moving on two different two-dimensional double-well potentials given by

$$U_1(x, y) = a(x - c)^2(x + d)^2 + by^2, \tag{17}$$

$$U_2(x, y) = -\epsilon \ln\left[e^{-a(x-c)^2-b(y-d)^2} + e^{-a(x+c)^2-b(y+d)^2}\right], \tag{18}$$

where (x, y) are particle coordinates and a, b, c, and d are free parameters of the potential, respectively (reported in Table I). The particle evolves according to Langevin dynamics[45] with friction constant 0.001 fs^{-1}. The system was propagated using a Velocity Verlet[46] integrator with a step size of 5 fs for 2×10^6 steps (10 ns) for simulations in U_1 and 4×10^6 steps (20 ns) for simulations in U_2. The initial momenta were drawn randomly from the Maxwell–Boltzmann distribution at 300 K. Analytic reference PMFs were obtained by numerical integration of exact probability densities.

Sampling along x was enhanced using eABF,[15] WTM-eABF,[20] or GaWTM-eABF[23] with $\sigma = 2.0$ Å and $m_\lambda = 20$ a.u. The fictitious particle was confined to the range of interest by harmonic walls with a force constant of 500 kJ/(mol Å^2). The ABF force was scaled with a linear ramp and fully applied in bins with more than 100 samples. For the metadynamics potential, Gaussians with a standard deviation of 6.0 Å were added every 20 steps with an initial height of 1.0 kJ/mol and scaled down over the course of the simulation in the

TABLE I. Parameters of numerical potentials U_1 [Eq. (17)] and U_2 [Eq. (18)].

	U_1	U_2
a	8×10^{-6} $\frac{\text{kJ}}{\text{mol Å}^4}$	5×10^{-3} Å^{-2}
b	5×10^{-1} $\frac{\text{kJ}}{\text{mol Å}^2}$	4×10^{-2} Å^{-2}
c	80 Å	40 Å
d	160 Å	20 Å
ϵ	...	1 kJ mol^{-1}

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

well-tempered framework[47] applying an effective temperature of 4000 K. To obtain parameters for the GaMD potential, the system was equilibrated for 4×10^5 steps, including 5×10^4 initial steps where no boost potential is applied. σ_0 was set to 3.5 kJ/mol. PMFs along the x-axis and free energy difference were calculated with the MBAR and CZAR estimator as described above.

B. Computing equilibrium NMR shifts

To demonstrate the calculation of unbiased ensemble averages of observables in real chemical systems, we compute the equilibrium NMR shieldings of gaseous ammonia. For this purpose, *ab initio* simulations are carried out with the Python interface of our in-house quantum-chemical program suite FermiONs++[48–51] at the level of ωB97M-V/def2-TZVP.[52,53] The molecule was heated from 0.1 to 310 K over 3100 time steps with a step size of 0.1 fs. Initial momenta were randomly drawn from the Maxwell–Boltzmann distribution. Velocities were rescaled every 10 time steps to increase the temperature by 1 K. For the production run, the heated system was simulated for 30 ps with a time step of 0.5 fs. The temperature was controlled with the Langevin thermostat at 310 K with a friction coefficient of 0.001 fs^{-1}. The dynamics of the ammonia inversion was biased with the WTM-eABF algorithm.[19,20] The mean of the three improper torsional angles χ was chosen as collective variable,

$$\xi(\mathbf{x}) = \frac{1}{3}\big[\chi(\mathbf{r}_{H_3}, \mathbf{r}_{H_2}, \mathbf{r}_{H_1}, \mathbf{r}_N) + \chi(\mathbf{r}_{H_1}, \mathbf{r}_{H_3}, \mathbf{r}_{H_2}, \mathbf{r}_N) + \chi(\mathbf{r}_{H_2}, \mathbf{r}_{H_3}, \mathbf{r}_{H_1}, \mathbf{r}_N)\big], \tag{19}$$

where $\mathbf{r}_N$ denotes the position of the nitrogen atom and $\mathbf{r}_{H_1}$ to $\mathbf{r}_{H_3}$ the hydrogen atom positions, respectively. For the WTM potential, Gaussians with an initial height of 1 kJ/mol and variance of 4° were deposited every 10 fs. The effective temperature was set to 1000 K. The fictitious particle had a mass of 20 a.u. and was coupled to the CV with a thermal width of 2°. The WTM force and ABF were collected on a grid with bin size 2°. The ABF was scaled with a linear ramp and fully applied in bins with more than 500 samples. Absolute isotropic NMR shieldings were calculated for all MD frames and for the optimized geometry using B97-2/pcSseg-2.[54,55]

IV. RESULTS AND DISCUSSION

A. Numerical tests

The CZAR and MBAR estimators are tested for extended-system dynamics by simulations of a single particle on potential energy surfaces U_1 and U_2. In both potentials, two metastable states are separated by barriers of roughly 8 k_BT (20 kJ/mol). As CV, $\xi_1(x, y) = x$ is chosen for all test simulations. We start with potential U_1, which is shown in Fig. 1(a) together with data points of an eABF-biased trajectory of the particle. Figure 1(b) shows the trajectory of the CV, which frequently crosses the transition state and gives an overall uniform probability density of z and λ. The resulting PMFs as obtained from MBAR and CZAR are shown in Fig. 1(c), both are almost identical with the analytical PMF and the bootstrapped standard deviations of both estimates are negligible because of the large sample size along the x-axis. As both estimators are asymptotically unbiased, this is what we expected in the limit of large sample sizes. Figure 1(d) shows PMF RMSDs for increasing distance

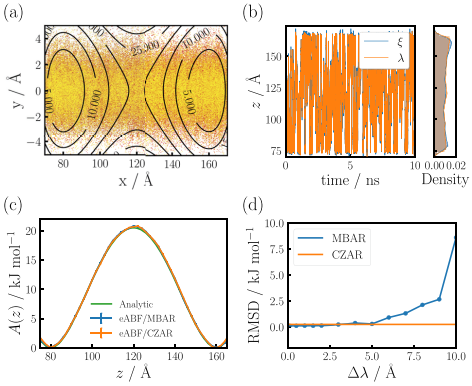


FIG. 1. (a) Every 10th data point of the eABF trajectory. The color indicates the time progression by gradually changing from yellow to red over the course of the simulation. Potential U_1 [Eq. (17)] is shown as contour plot. (b) Left: eABF trajectory along $\xi(x, y) = x$. Right: Probability densities of ξ and λ. (c) PMFs obtained with either analysis method, analytical reference in green. Error bars indicating the standard deviation of 100 bootstrap runs are smaller than the linewidth. (d) RMSD of PMFs with respect to the analytic reference. The MBAR PMF is computed for different distances between discrete λ-states. The PMF RMSD obtained with CZAR is shown as orange line. All PMFs are computed for a bin size of $\Delta z = 2$ Å.

between discrete λ-states $\Delta\lambda$, hereafter referred to as "window size." The corresponding PMF RMSD from CZAR is shown in orange. Interestingly, the PMF RMSD of the MBAR estimator is constant at about 0.1 kJ/mol if the window size is below or equal to the thermal coupling width σ of λ to the CV, which was 2 Å in this simulation. This can be rationalized by the fact that σ is the standard deviation of the CV for a given λ-state. Thus, it provides an intuitive and robust choice for $\Delta\lambda$. For larger windows of up to 5 Å, the PMF RMSD rises moderately to about 0.3 kJ/mol. Only for window sizes larger than 6 Å (3 σ), the error rises above 1 kJ/mol. The CZAR estimate has no $\Delta\lambda$-dependence and is, therefore, visualized as a line in Fig. 1(d).

We next turn to WTM-eABF and GaWTM-eABF simulations on potential U_2, where we again choose $\xi_1(x, y) = x$. Figure 2(a) shows the PMF obtained from WTM-eABF. As before, CZAR and MBAR estimates are almost identical and bootstrapped standard deviations are negligible. However, both computed PMFs are very different from the analytical result. The reason becomes clear when looking at the sampling of the potential $U_2(x, y)$, as depicted in Fig. 2(b). Because of the chosen collective variable, sampling is only accelerated along the x-axis, missing the y-component of the transition between both minima. Therefore, the system rarely crosses the transition state, but rather samples both parallel valleys. As shown in Figs. 2(c) and 2(d), PMFs estimated with GaWTM-eABF are much closer to the analytic reference because the dividing surface between both states is much better sampled. Here, the MBAR estimate slightly underestimates the analytical reference; an error in this region is expected as both valleys are still sampled more densely than the transition state. The CZAR estimate overestimates the PMF at

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

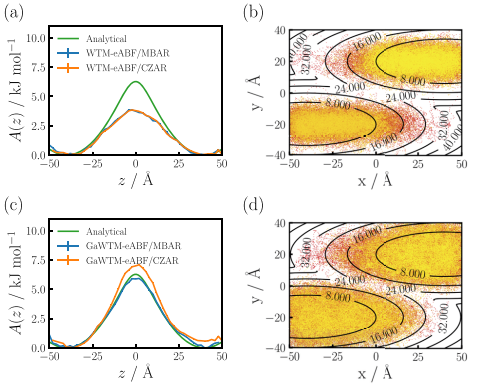


FIG. 2. [(a) and (c)] PMFs obtained from WTM-eABF and GaWTM-eABF simulations, respectively, using $\xi(x, y) = x$ and either analysis method; analytical reference in green. Error bars indicating the standard deviation of 100 bootstrap runs are smaller than the linewidth. All PMFs are computed with bin and window size 2 Å. [(b) and (d)] Every 20th data point of the WTM-eABF and GaWTM-eABF trajectory. The color indicates the time progression by gradually changing from yellow to red over the course of the simulation. Potential U_2 [Eq. (18)] is shown as contour plot.

the transition state. This can be attributed to the additional second order cumulant expansion of ΔU [Eq. (9)], which accounts for the contribution of GaMD to the PMF. The implied approximation only holds if ΔU obeys a near-Gaussian distribution (see also Sec. S2 of the supplementary material). For larger values of σ_0, this approximation breaks down because the distribution of ΔU is no longer Gaussian. For the MBAR estimator, no such approximations are made. Obtained PMFs are not affected by the harmonicity of ΔU and are, therefore, insensitive to the choice of GaMD parameters, making the analysis of GaWTM-eABF simulations much more robust.

Finally, to demonstrate how MBAR can be used to recover unbiased ensemble averages, we will reweight the PMF of WTM-eABF/MBAR and GaWTM-eABF/MBAR simulations shown in Fig. 2 to a new CV, $\xi_2(x, y) = 0.25x + y$, and we also compute the conditional average $\langle U_2 \rangle_z$. Figure 3(a) shows the reweighted PMF. Because ξ_2 is orthogonal to the dividing surface of both states, it constitutes the ideal CV for the given transition.[36] Because ξ_1 was chosen as CV during the simulations, using WTM-eABF, the transition state is barely sampled, again causing large deviations from the analytical PMF along ξ_2. GaWTM-eABF reproduces the analytic PMF remarkably well. For $\langle U_2 \rangle_z$, the difference between both sampling algorithms is less severe. Not only does the conditional average computed from GaWTM-eABF/MBAR match the analytic reference almost exactly, but also the result from WTM-eABF/MBAR is only slightly distorted. The reason is that with both methods, minima of the potential energy are sampled sufficiently, which carry the by far highest statistical weight for the mean potential energy along the x-axis. Overall, both examples demonstrate the ability of MBAR to

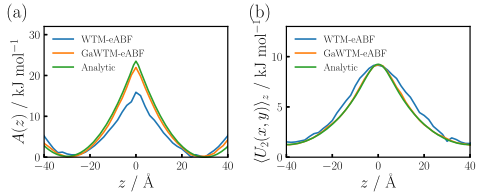


FIG. 3. (a) Reweighting of WTM-eABF and GaWTM-eABF simulations shown in Fig. 2: (a) to a different collective variable $\xi_2(x, y) = 0.25x + y$, (b) to the conditional average of the potential energy of U_2 along the x-axis.

produce accurate ensemble and conditional averages, if the underlying configuration space is sampled sufficiently. In Sec. IV B, we show how the same technique can be applied to a real chemical system, i.e., for the prediction of NMR shieldings.

B. Equilibrium NMR

To showcase the usefulness of computing equilibrium observables from biased trajectories of a real molecular system, we calculate NMR shieldings for ammonia. Figure 4(a) shows the WTM-eABF-biased trajectory of ammonia using a linear combination of the three improper torsional angles as CV [Eq. (19)]. The transition state at $z = 0°$, where the molecule is planar, is frequently crossed and the probability density is roughly uniform between $-50°$ and $50°$. Figure 4(b) shows the PMF and its error bars obtained from 100 bootstrapping runs with the MBAR or CZAR estimator, respectively. Using a grid size of $1°$, numerical errors of the CZAR estimate due to thermodynamic integration are negligible and both PMFs are almost identical with bootstrapped standard deviations of roughly 0.3 kJ/mol.

Absolute isotropic shieldings are calculated for all data frames of the WTM-eABF-biased trajectory and reweighted according to Eq. (13) in order to obtain unbiased averages. Figures 5(a) and 5(b) show probability densities of ^{15}N and ^{1}H shifts. Blue and orange histograms represent the biased and reweighted probability density, respectively. For ^{1}H, shieldings of all three protons are averaged for each data point. Mean chemical shifts are given as vertical lines.

Biased probability distributions are slightly skewed and broader due to enhanced sampling, which deliberately overestimates the

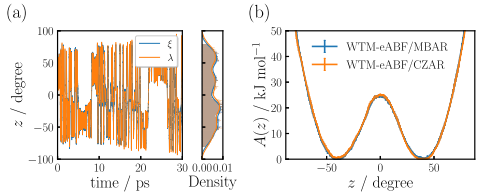


FIG. 4. (a) Left: WTM-eABF trajectory. Right: Probability densities of ξ and λ. (b) PMFs obtained from WTM-eABF/MBAR and WTM-eABF/CZAR. Error bars indicate the standard deviation from 100 bootstrapping runs.

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

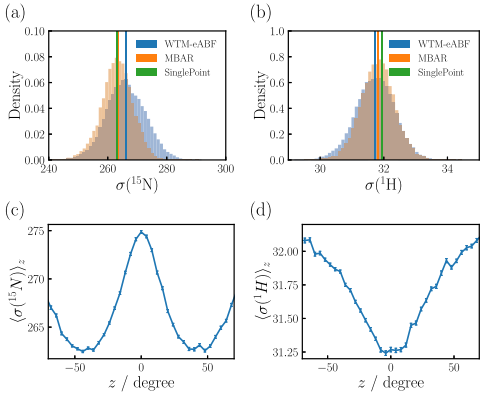


FIG. 5. [(a) and (b)] Absolute isotropic ^{15}N and ^{1}H shieldings, respectively. The distribution of chemical shifts as obtained from an WTM-eABF-biased trajectory is shown in blue and the MBAR reweighted distribution in orange. The vertical lines of the same color indicate mean values. The green line is positioned at the shielding value of the minimum energy geometry. [(c) and (d)] Conditional average of absolute isotropic ^{15}N and ^{1}H shieldings along the CV, respectively. Error bars indicate the standard error of the mean of the conditional average.

probability of configurations that are located in high energy (low probability) regions. Reweighting the probability of individual data frames with the unbiased probability as obtained from MBAR recovers the expected Gaussian-shaped Boltzmann distributions for the equilibrium NMR shieldings. Figures 5(c) and 5(d) show conditional averages of isotropic shieldings along the CV. For ^{15}N, $\langle\sigma\rangle_z$ has the same shape as that of the PMF, local minima and maxima of the chemical shielding curve occur at the same z values as those of the PMF. Therefore, the ^{15}N shielding for the optimized structure is almost identical to the unbiased mean. The biased mean is shifted by about 4 ppm, due to overestimation of the probability of transition state configurations. In contrast, the conditional average of ^{1}H shieldings has a V-like shape, where the minimum is located at the transition state. Since high energy geometries correspond to z values that can have both lower or higher isotropic shieldings, the biased and unbiased mean are much more similar than in the case of ^{15}N [compare Figs. 4(a) and 4(b)].

To summarize, we have shown how MBAR can be applied to post-process extended-system dynamics simulations of molecular transitions. Thus, having recovered the full unbiased statistical information of the given dataset, equilibrium properties such as NMR shieldings, are recovered seamlessly by reweighting the probability of each data frame.

V. CONCLUSION

We have shown that MBAR is a suitable and robust estimator of statistical information for the family of eABF-based enhanced sampling algorithms. While the CZAR estimator already yields fast and accurate on-the-fly estimates of PMFs, MBAR supplements it by offering a statistically optimal analysis in post-processing. By computing the unbiased statistical weight of each frame, it extends the application of eABF and its variants from the computation of PMFs to ensemble averages in general. The presented framework, therefore, enables the application of this highly efficient class of enhanced sampling algorithms without any loss of statistical information. We have shown how this can be useful for the prediction of equilibrium properties, such as NMR shieldings, which are highly sensitive to the molecular geometry and, therefore, have to be calculated from ensemble averages accounting for all contributing configurations. Additionally, we have shown how PMFs can be reweighted to other collective variables of interest, which might yield mechanistic insight into complex chemical processes.

We expect the recently published GaWTM-eABF to be particularly useful in this regard, as it reduces the dependence of sampling efficiency on the choice of collective variables. Here, MBAR additionally removes the requirement on the distribution of the GaMD boost potential to be Gaussian-shaped, making estimations of free energies significantly more accurate and robust against the choice of GaMD parameters.

SUPPLEMENTARY MATERIAL

The supplementary material contains a pdf file with a detailed discussion of the separation of the distribution of ξ into λ-dependent subsamples and the different dependence of the estimated PMF on GaMD parameters when using MBAR or a cumulant expansion. Additionally, a zip archive containing all scripts to perform the numerical simulations of Sec. IV A is given. A maintained and updated version of the code is available at https://github.com/ochsenfeld-lab/adaptive_sampling.

ACKNOWLEDGMENTS

The authors thank J. Kussmann (LMU Munich) for providing a development version of the FermiONs++ program package and Yannick Lemke for useful comments and discussions. Financial support was provided by the "Deutsche Forschungsgemeinschaft" (DFG, German Research Foundation) within the Cluster of Excellence "e-conversion" (Grant No. EXC 2089/1-390776260) and SFB 1309-325871075, "Chemical Biology of Epigenetic Modifications." C.O. acknowledges further support as a Max Planck Fellow at MPI-FKF Stuttgart. J.C.B.D. acknowledges support from the Leopoldina fellowship program, German National Academy of Sciences Leopoldina (Grant No. LPDS 2021-08).

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Andreas Hulm: Conceptualization (equal); Formal analysis (equal); Methodology (equal); Software (equal); Validation (equal); Writing – original draft (equal). **Johannes C. B. Dietschreit:** Conceptualization (equal); Formal analysis (equal); Methodology (equal); Writing – original draft (equal). **Christian Ochsenfeld:** Funding acquisition (equal); Supervision (equal); Writing – review & editing (equal).

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

DATA AVAILABILITY

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

REFERENCES

[1] C. Chipot and A. Pohorille, *Free Energy Calculations* (Springer, 2007).
[2] C. Chipot, Wiley Interdiscip. Rev.: Comput. Mol. Sci. **4**, 71 (2014).
[3] G. M. Torrie and J. P. Valleau, J. Comput. Phys. **23**, 187 (1977).
[4] E. Darve and A. Pohorille, J. Chem. Phys. **115**, 9169 (2001).
[5] A. Laio and M. Parrinello, Proc. Natl. Acad. Sci. U. S. A. **99**, 12562 (2002).
[6] C. Abrams and G. Bussi, Entropy **16**, 163 (2014).
[7] V. Spiwok, Z. Sucur, and P. Hosek, Biotechnol. Adv. **33**, 1130 (2015).
[8] O. Valsson, P. Tiwary, and M. Parrinello, Annu. Rev. Phys. Chem. **67**, 159 (2016).
[9] M. Köppen, in 5th Online World Conference on Soft Computing in Industrial Applications (WSC5), 2000, Vol. 1, pp. 4–8.
[10] E. Darve, D. Rodríguez-Gómez, and A. Pohorille, J. Chem. Phys. **128**, 144120 (2008).
[11] J. Comer, J. C. Gumbart, J. Hénin, T. Lelièvre, A. Pohorille, and C. Chipot, J. Phys. Chem. B **119**, 1129 (2015).
[12] J. G. Kirkwood, J. Chem. Phys. **3**, 300 (1935).
[13] G. Ciccotti, R. Kapral, and E. Vanden-Eijnden, ChemPhysChem **6**, 1809 (2005).
[14] G. Fiorin, M. L. Klein, and J. Hénin, Mol. Phys. **111**, 3345 (2013).
[15] A. Lesage, T. Lelièvre, G. Stoltz, and J. Hénin, J. Phys. Chem. B **121**, 3676 (2017).
[16] G. Bussi, A. Laio, and M. Parrinello, Phys. Rev. Lett. **96**, 090601 (2006).
[17] A. Laio and F. L. Gervasio, Rep. Prog. Phys. **71**, 126601 (2008).
[18] A. Barducci, M. Bonomi, and M. Parrinello, Wiley Interdiscip. Rev.: Comput. Mol. Sci. **1**, 826 (2011).
[19] H. Fu, H. Zhang, H. Chen, X. Shao, C. Chipot, and W. Cai, J. Phys. Chem. Lett. **9**, 4738 (2018).
[20] H. Fu, X. Shao, W. Cai, and C. Chipot, Acc. Chem. Res. **52**, 3254 (2019).
[21] H. Fu, H. Chen, X. a. Wang, H. Chai, X. Shao, W. Cai, and C. Chipot, J. Chem. Theory Comput. **60**, 5366 (2020).
[22] H. Fu, X. Shao, C. Chipot, and W. Cai, J. Chem. Theory Comput. **12**, 3506 (2016).
[23] H. Chen, H. Fu, C. Chipot, X. Shao, and W. Cai, J. Chem. Theory Comput. **17**, 3886 (2021).
[24] Y. Miao, V. A. Feher, and J. A. McCammon, J. Chem. Theory Comput. **11**, 3584 (2015).
[25] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, J. Comput. Chem. **13**, 1011 (1992).
[26] M. R. Shirts and J. D. Chodera, J. Chem. Phys. **129**, 124105 (2008).
[27] M. R. Shirts and A. L. Ferguson, J. Chem. Theory Comput. **16**, 4107 (2020).
[28] G. Tiana, Eur. Phys. J. B **63**, 235 (2008).
[29] M. Bonomi, A. Barducci, and M. Parrinello, J. Comput. Chem. **30**, 1615 (2009).
[30] P. Tiwary and M. Parrinello, J. Phys. Chem. B **119**, 736 (2015).
[31] T. M. Schäfer and G. Settanni, J. Chem. Theory Comput. **16**, 2042 (2020).
[32] H. Paliwal and M. R. Shirts, J. Chem. Theory Comput. **7**, 4115 (2011).
[33] M. Dracinsky, H. M. Möller, and T. E. Exner, J. Chem. Theory Comput. **9**, 3806 (2013).
[34] J. C. B. Dietschreit, A. Wagner, T. A. Le, P. Klein, H. Schindelin, T. Opatz, B. Engels, U. A. Hellmich, and C. Ochsenfeld, Angew. Chem., Int. Ed. **59**, 12669 (2020).
[35] C. Hartmann, J. C. Latorre, and G. Ciccotti, EPJ: Spec. Top. **200**, 73 (2011).
[36] J. C. B. Dietschreit, D. J. Diestler, and C. Ochsenfeld, J. Chem. Phys. **156**, 114105 (2022).
[37] D. Mendels, G. Piccini, and M. Parrinello, J. Phys. Chem. Lett. **9**, 2776 (2018).
[38] Y. Wang, J. M. L. Ribeiro, and P. Tiwary, Nat. Commun. **10**, 3573 (2019).
[39] L. Sun, J. Vandermause, S. Batzner, Y. Xie, D. Clark, W. Chen, and B. Kozinsky, J. Chem. Theory Comput. **18**, 1549 (2022).
[40] L. Bonati, V. Rizzi, and M. Parrinello, J. Phys. Chem. Lett. **11**, 2998 (2020).
[41] D. Wang and P. Tiwary, J. Chem. Phys. **154**, 134111 (2021).
[42] K. Minoukadeh, C. Chipot, and T. Lelièvre, J. Chem. Theory Comput. **6**, 1008 (2010).
[43] J. Comer, J. C. Phillips, K. Schulten, and C. Chipot, J. Chem. Theory Comput. **10**, 5276 (2014).
[44] B. Efron, Ann. Stat. **7**, 1 (1979).
[45] M. Kröger, *Models for Polymeric and Anisotropic Liquids* (Springer Science & Business Media, 2005), Vol. 675.
[46] W. C. Swope, H. C. Andersen, P. H. Berens, and K. R. Wilson, J. Chem. Phys. **76**, 637 (1982).
[47] A. Barducci, G. Bussi, and M. Parrinello, Phys. Rev. Lett. **100**, 020603 (2008).
[48] J. Kussmann and C. Ochsenfeld, J. Chem. Phys. **138**, 134114 (2013).
[49] J. Kussmann and C. Ochsenfeld, J. Chem. Theory Comput. **11**, 918 (2015).
[50] H. Laqua, T. H. Thompson, J. Kussmann, and C. Ochsenfeld, J. Chem. Theory Comput. **16**, 1456 (2020).
[51] H. Laqua, J. Kussmann, and C. Ochsenfeld, J. Chem. Phys. **154**, 214116 (2021).
[52] N. Mardirossian and M. Head-Gordon, J. Chem. Phys. **144**, 214110 (2016).
[53] F. Weigend and R. Ahlrichs, Phys. Chem. Chem. Phys. **7**, 3297 (2005).
[54] P. J. Wilson, T. J. Bradley, and D. J. Tozer, J. Chem. Phys. **115**, 9233 (2001).
[55] F. Jensen, J. Chem. Theory Comput. **11**, 132 (2015).

Supporting Information: Statistically Optimal Analysis of the Extended-system Adaptive Biasing Force (eABF) Method

Andreas Hulm,[1] Johannes C. B. Dietschreit,[1,2] Christian Ochsenfeld[*,1,3]

[1]Chair of Theoretical Chemistry, Department of Chemistry,
University of Munich (LMU), Butenandtstr. 7, D-81377 München, Germany
[2]Department of Materials Science and Engineering,
Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States
[3]Max Planck Institute for Solid State Research, Heisenbergstr. 1, D-70569 Stuttgart, Germany
[*]E-Mail: christian.ochsenfeld@uni-muenchen.de

Contents

1 Separation of the Distribution of ξ into λ-dependent Subsamples

To understand the nature of subsamples of eABF trajectories that are introduced in the application of MBAR to extended-system simulations, let us derive their λ-conditioned probability density. In the extended-system potential $U_{ext}(\mathbf{x}, \lambda, t)$, the physical system is coupled to fictitious particles λ with harmonic potentials.[1] The z-conditioned probability distribution of U_{ext} can be obtained by inserting eq. (5) into eq. (1) of the main manuscript to obtain

$$\begin{aligned}
\rho_{\text{ext}}(z, \lambda, t=\infty) &\propto \int \delta[\xi(\mathbf{x}) - z] \ \exp\big[-\beta U_{\text{ext}}(\mathbf{x}, \lambda, t=\infty)\big] \mathrm{d}\mathbf{x} \\
&= \int \delta[\xi(\mathbf{x}) - z] \ \exp\big[-\beta U(\mathbf{x})\big] \ \exp\big[-\frac{1}{2\sigma^2}(\xi(\mathbf{x}) - \lambda)^2\big] \ \exp\big[-\beta U_b(\lambda, t=\infty)\big] \ \mathrm{d}\mathbf{x} \\
&= \exp\big[-\frac{1}{2\sigma^2}(z-\lambda)^2\big] \ \exp\big[\beta A(\lambda)\big] \int \delta[\xi(\mathbf{x}) - z] \ \exp\big[-\beta U(\mathbf{x})\big] \ \mathrm{d}\mathbf{x} \\
&= Z \ \rho(z) \ \exp\big[-\frac{1}{2\sigma^2}(z-\lambda)^2\big] \ \exp\big[\beta A(\lambda)\big] \qquad \text{(S1)}
\end{aligned}$$

where we assume that $U_b(\lambda, t)$ converges to a potential $-A(\lambda)$ in the long time limit and Z is the configuration integral of the unbiased, non-extended ensemble. Finally, fixing λ to some value $\lambda = \lambda_i$ gives the λ-conditioned probability density of z

$$\rho^B(z|\lambda_i) \propto \rho(z) \ \exp\big[-\frac{1}{2\sigma^2}(z-\lambda_i)^2\big] \ , \qquad \text{(S2)}$$

which is proportional to $\rho(z)$ contracted with a Gaussian kernel of variance σ^2. In practice λ-windows are constructed on a grid with window size $\Delta\lambda$, which automatically yields M states with overlapping probability distributions, as shown in Fig. S1 for $\Delta\lambda = 2$ Å and $\Delta\lambda = 6$ Å.

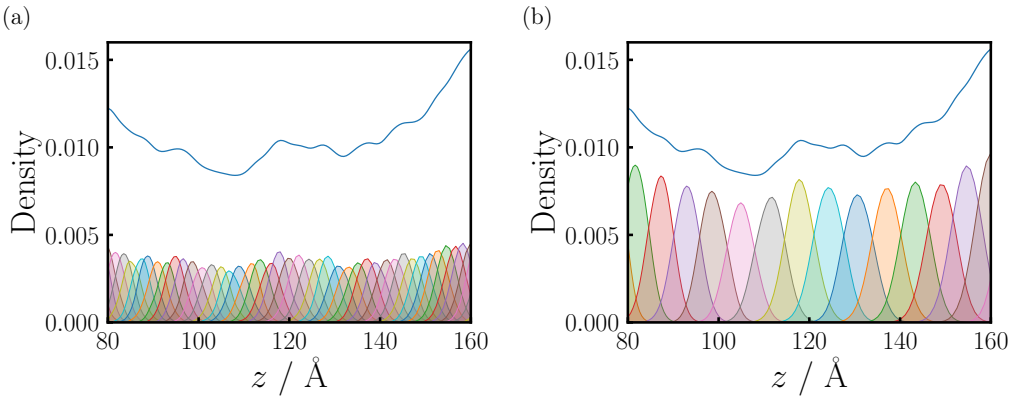


Figure S 1: Kernel density approximation of conditional probability densities $\rho^B(z|\lambda_i)$ in separate λ windows for simulations shown in Fig. 1 in the main manuscript. Probabilities densities are normalized to the original samples size. λ windows are located every (a) 2 Å and (b) 6 Å. The blue line denotes the mixture distribution of all combined λ windows, which corresponds to the original probability density $\rho^B(z)$ of the eABF simulation and is equal for both plots.

2 Different Dependence of CZAR and MBAR on the GaMD Parameters

Let us recall the principles of energetic reweighing using a cumulant expansion, which is commonly applied in GaMD simulations.[2] The biased probability distribution $\rho^B(z)$, where in the case of GaWTM-eABF the eABF bias has already been removed by means of CZAR,[3] can be reweighed with the boost potential denoted as $\Delta U(\mathbf{x})$

$$\rho(z) = \rho^B(z) \frac{\langle e^{\beta \Delta U(\mathbf{x})} \rangle_{\Delta U, \xi(\mathbf{x})=z}}{\langle e^{\beta \Delta U(\mathbf{x})} \rangle_{\Delta U}}, \tag{S3}$$

with the conditional average of Boltzmann factors of the boost potential in the GaMD biased ensemble $\langle e^{\beta \Delta U(\mathbf{x})} \rangle_{\Delta U, \xi(\mathbf{x})=z}$, which can be approximated by a cumulant-expansion[4]

$$\langle e^{\beta \Delta U(\mathbf{x})} \rangle_{\Delta U, \xi(\mathbf{x})=z} = \exp\left[\sum_{k=1}^{\infty} \frac{\beta^k}{k!} C_k \right], \tag{S4}$$

where C_k are the expansion coefficients. Usually, the cumulant expansion is truncated at the second order, assuming that the boost potential follows a near-Gaussian distribution. In this case it is more accurate than, e.g., the exponential average, as it reduces the energetic noise.

Therefore, in GaMD a Gaussian distribution of ΔU that is narrow enough for accurate reweighing is ensured by introducing the free parameter σ_0 as an upper bound to its standard deviation. However, this is obsolete using MBAR for analysis of GaWTM-eABF as no such approximation is made. The MBAR equations are solved for the full bias including ΔU, which directly gives unbiased probabilities of individual frames as for normal eABF. It is generally possible, however, to separate the GaMD-reweighting for MBAR just as is done for CZAR.

Figure S2 shows the convergence of eq. (S4) for the GaWTM-eABF simulation shown in Fig. (2) of the main manuscript. The analytical reference obtained by numeric integration of the exact probability density is shown in green. As discussed in the main manuscript the second order cumulant expansion, which is commonly applied in GaMD and GaWTM-eABF simulations, overestimates the PMF at the transition state significantly, as it fails to capture the anharmonicity of ΔU. The 3rd order of the cumulant expansion underestimates the PMF at the transition states but improves the PMF RMSD, shown in table S1, significantly. The 4th order expansion is closest to the analytical reference at the transition state, but increases the overall noise in the estimate which again raises the PMF RMSD. The MBAR estimate shown in blue is by far the most accurate, as it incorporates no such expansion.

Table S1: PMF RMSDs for the MBAR estimate as well as CZAR estimates with different orders of cumulant expansions to the analytical PMF for GaWTM-eABF simulation shown in Figure 2 of the main manuscript.

Estimator	PMF RMSD / kJ mol^{-1}
MBAR	0.16
CZAR, $k = 2$	0.64
CZAR, $k = 3$	0.34
CZAR, $k = 4$	0.42

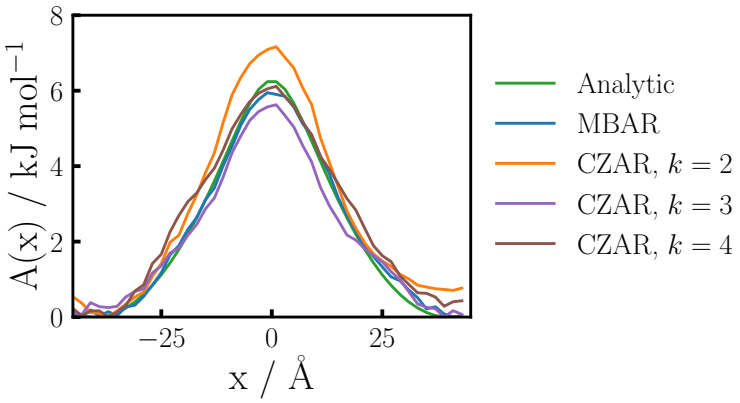


Figure S2: PMFs obtained from the GaWTM-eABF simulation shown in Figure 2 of the main manuscript using the MBAR and CZAR estimator; analytical reference in green. CZAR estimates are corrected with cumulant expansions of order 2, 3, and 4, respectively. All PMFs are computed with bin and window size 2 Å.

References

[1] A. Lesage, T. Lelievre, G. Stoltz, J. Henin, *J. Phys. Chem. B* **2017**, *121*, 3676–3685.

[2] Y. Miao, V. A. Feher, J. A. McCammon, *J. Chem. Theory Comput.* **2015**, *11*, 3584–3595.

[3] H. Chen, H. Fu, C. Chipot, X. Shao, W. Cai, *J. Chem. Theory Comput.* **2021**, *17*, 3886–3894.

[4] M. P. Eastwood, C. Hardin, Z. Luthey-Schulten, P. G. Wolynes, *J. Chem. Phys.* **2002**, *117*, 4602–4615.

3.2 Publication II: Improved Sampling of Adaptive Path Collective Variables by Stabilized Extended-System Dynamics

Abstract: Because of the complicated multistep nature of many biocatalytic reactions, an a priori definition of reaction coordinates is difficult. Therefore, we apply enhanced sampling algorithms along with adaptive path collective variables (PCVs), which converge to the minimum free energy path (MFEP) during the simulation. We show how PCVs can be combined with the highly efficient well-tempered metadynamics extended-system adaptive biasing force (WTM-eABF) hybrid sampling algorithm, offering dramatically increased sampling efficiency due to its fast adaptation to path updates. For this purpose, we address discontinuities of PCVs that can arise due to path shortcutting or path updates with a novel stabilization algorithm for extended-system methods. In addition, we show how the convergence of simulations can be further accelerated by utilizing the multistate Bennett's acceptance ratio (MBAR) estimator. These methods are applied to the first step of the enzymatic reaction mechanism of pseudouridine synthases, where the ability of path WTM-eABF to efficiently explore intricate molecular transitions is demonstrated.

Reprinted with permission from

A. Hulm; C. Ochsenfeld. "Improved Sampling of Adaptive Path Collective Variables by Stabilized Extended-System Dynamics" *J. Chem. Theory Comput.* **2023**, 19, 9202-9210. URL: `https://doi.org/10.1021/acs.jctc.3c00938`.

JCTC Journal of Chemical Theory and Computation

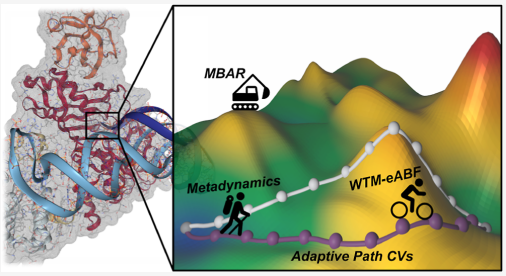
pubs.acs.org/JCTC Article

Improved Sampling of Adaptive Path Collective Variables by Stabilized Extended-System Dynamics

Andreas Hulm and Christian Ochsenfeld*

Cite This: J. Chem. Theory Comput. 2023, 19, 9202−9210 Read Online

ABSTRACT: Because of the complicated multistep nature of many biocatalytic reactions, an a priori definition of reaction coordinates is difficult. Therefore, we apply enhanced sampling algorithms along with adaptive path collective variables (PCVs), which converge to the minimum free energy path (MFEP) during the simulation. We show how PCVs can be combined with the highly efficient well-tempered metadynamics extended-system adaptive biasing force (WTM-eABF) hybrid sampling algorithm, offering dramatically increased sampling efficiency due to its fast adaptation to path updates. For this purpose, we address discontinuities of PCVs that can arise due to path shortcutting or path updates with a novel stabilization algorithm for extended-system methods. In addition, we show how the convergence of simulations can be further accelerated by utilizing the multistate Bennett's acceptance ratio (MBAR) estimator. These methods are applied to the first step of the enzymatic reaction mechanism of pseudouridine synthases, where the ability of path WTM-eABF to efficiently explore intricate molecular transitions is demonstrated.



INTRODUCTION

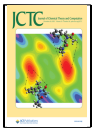
The computation of reliable reaction and activation-free energies of biocatalytic reactions requires extensive sampling of molecular transitions,[1,2] which is typically obtained using ab initio molecular dynamics (MD) simulations on composite quantum mechanics/molecular mechanics (QM/MM) level of theory.[3] Using these highly costly simulations, only time scales of up to several 100 ps can be reached, which means that reactive events that are separated by high free energy barriers can never be observed in conventional MD trajectories. It is therefore paramount to apply importance-sampling strategies that speed up the exploration of high-energy regions.[4]

Many of these methods rely on the definition of a low-dimensional set of collective variables (CVs) that discriminate between reactant and product states.[5−8] For example, metadynamics (MtD) accelerates the exploration of predefined CVs with an adaptive bias potential that builds up during the simulation.[7,9] However, a bad choice of CV results in hysteresis and poor convergence of the free energy estimate.[10] For simple transitions, it can be straightforward to find sufficient CVs based on chemical intuition, but for complicated enzymatic processes, the definition of good CVs that contain all the slow degrees of freedom of the given process can become exceedingly difficult. Additionally, CV space is typically limited to up to 3 dimensions because of the exponential growth of computational cost with the number of dimensions,[11] which is insufficient for complicated multistep transitions.

This problem motivates large research efforts to design one-dimensional CVs that can describe transitions with many slow degrees of freedom, for example, utilizing machine learning methods[12−15] or path CVs (PCVs).[16−19] For the latter, a path is defined by a string of discrete nodes that connect metastable states, and the CV is given by a progress parameter. This not only allows for a smooth, one-dimensional parametrization of complex transitions but also provides the opportunity for systematic on-the-fly improvement by iteratively moving a guess path closer to the minimum free energy path (MFEP).

In this contribution, we will build on path MtD as formulated by Ensing and co-workers.[18,19] We show how MtD can be replaced with the more efficient well-tempered metadynamics extended-system adaptive biasing force (WTM-eABF)[20,21] hybrid algorithm. Additionally, we show how postprocessing of path WTM-eABF with the multistate Bennett's acceptance ratio (MBAR)[22] estimator, which recovers the unbiased statistical weight of each simulation frame, can further improve the convergence of adaptive path simulations. To this end, after a short theoretical overview of the path WTM-eABF algorithm,

Received: August 25, 2023
Revised: November 8, 2023
Accepted: November 8, 2023
Published: December 11, 2023



https://doi.org/10.1021/acs.jctc.3c00938
J. Chem. Theory Comput. 2023, 19, 9202−9210

Journal of Chemical Theory and Computation pubs.acs.org/JCTC Article

we compare its performance to that of conventional path MtD on a numerical Müller−Brown (MB) potential. Afterward, we showcase the application of path WTM-eABF on the biocatalytic reaction mechanism of pseudouridine synthases (PUS), which involves a challenging rotating motion of an unbound uridine.

THEORY

Let $\xi(\mathbf{x})$ be some CV that represents the reaction coordinate and connects two metastable states. For a finite value z of the CV, it has the marginal probability distribution

$$\rho(z) = \int \delta[\xi(\mathbf{x}) - z]\rho(\mathbf{x})\mathrm{d}\mathbf{x} = \left\langle \delta[\xi(\mathbf{x}) - z] \right\rangle \tag{1}$$

where $\langle\rangle$ denotes the ensemble average and $\delta[x]$ denotes the Dirac delta distribution.[1] Our goal is to efficiently estimate $\rho(z)$, which defines the potential of mean force (PMF) (i.e., free energy profile), according to

$$A(z) = -\beta^{-1}\ln \rho(z) \tag{2}$$

where $\beta = (k_\mathrm{B}T)^{-1}$ and k_B is the Boltzmann constant.

For this purpose, MtD builds a repulsive potential

$$U_{\mathrm{bias}}^{\mathrm{MtD}}(\xi(\mathbf{x}), t) = \sum_{t=0,\tau_\mathrm{G},2\tau_\mathrm{G},\ldots} h_\mathrm{G} \mathrm{e}^{-(\xi(\mathbf{x})-\xi_t)^2/2\sigma_\mathrm{G}^2} \tag{3}$$

by adding Gaussian hills with height h_G and variance σ_G in regular time intervals τ_G, pushing the system away from already explored regions of CV space.[23] To ensure smooth convergence of $U_{\mathrm{bias}}^{\mathrm{MtD}}$, WTM adds an exponential scaling factor $\mathrm{e}^{-U_{\mathrm{bias}}^{\mathrm{MtD}}(\xi(\mathbf{x}),t)/k_\mathrm{B}\Delta T}$ with an effective bias temperature ΔT to decrease the height of new Gaussians over time.[9] Upon convergence, the PMF can be directly obtained from the inverse of the bias potentials.

In contrast, in the closely related WTM-eABF sampling algorithm,[20,21] a MD simulation of the extended system $(\mathbf{x}, \lambda)$ is performed in the potential

$$\begin{aligned} U(\mathbf{x},\lambda,t) &= U(\mathbf{x}) + U_{\mathrm{ext}}(\xi(\mathbf{x}),\lambda) + U_{\mathrm{bias}}(\lambda,t) \\ &= U(\mathbf{x}) + \frac{1}{2\beta\sigma^2}(\xi(\mathbf{x}) - \lambda)^2 + U_{\mathrm{bias}}(\lambda,t) \end{aligned} \tag{4}$$

where the molecular potential energy function $U(\mathbf{x})$ is coupled to an extended variable λ by a harmonic potential with thermal coupling width σ. Typically, small values of σ are chosen to ensure a tight coupling of λ to $\xi(\mathbf{x})$. This framework offers high flexibility and robustness against the choice of bias potential, $U_{\mathrm{bias}}(\lambda, t)$, which only acts on λ and has no direct impact on the physical system. In WTM-eABF, a combination of the WTM bias potential and an adaptive biasing force (ABF[8,24]), which is obtained as the average force acting on λ at a certain value of the CV, is applied to ensure fast exploration of the reaction coordinate.[21]

In practice, the biggest shortcoming of both MtD/WTM and WTM-eABF is their dependence on the choice of CVs, which are usually defined a priori. Bad choices of CVs lead to significant artifacts of sampling and free energy estimates.[10] In addition, it was shown that activation free energies and reaction rates are particularly vulnerable,[25] which are among the most important targets of mechanistic studies. One promising direction to mitigate this problem is the development of systematically improvable CVs like adaptive PCVs, which exist in various different flavors.[17−19,26] Here, we apply geometric PCVs as proposed by Ensing and co-workers,[18,19] although the presented framework is also valid for other formulations. A path is defined by M discrete, equidistant nodes that are placed along the reaction coordinate. The progress $s(\mathbf{z})$ along the path in the space of some selected CV space $\mathbf{z} = (\xi_1(\mathbf{x}), \xi_2(\mathbf{x}), \ldots)$ can be defined by

$$s(\mathbf{z}) = \frac{m}{M} \pm \frac{1}{2M} \left(\frac{\sqrt{(\mathbf{v}_1\cdot\mathbf{v}_3)^2 - |\mathbf{v}_3|^2(|\mathbf{v}_1|^2 - |\mathbf{v}_2|^2)} - (\mathbf{v}_1\cdot\mathbf{v}_3)}{|\mathbf{v}_3|^2} - 1 \right) \tag{5}$$

where m is the zero-based index of the closest node and vectors $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_3$ are defined by $\mathbf{v}_1 = \mathbf{z}_m - \mathbf{z}$, $\mathbf{v}_2 = \mathbf{z} - \mathbf{z}_{m-1}$, and $\mathbf{v}_3 = \mathbf{z}_{m+1} - \mathbf{z}_m$. The $\pm$ in eq 5 is positive if z is left of the closest path node and a negative sign otherwise.

Thus, in eq 4, λ is now coupled to the progress parameter $s(\mathbf{z})$ instead of a simple CV. We note that this definition does not completely eliminate the problem of manually selecting relevant CVs but rather replaces it with the more flexible choice of an appropriate CV space. To avoid the definition of a CV space, one could alternatively apply the arithmetic PCV as formulated by Branduardi and co-workers.[17] More details on our PCV implementation are given in Section S1 of the Supporting Information.

To converge the path to the MFEP, initial guess nodes are adapted to the average CV density perpendicular to the path by updating the node positions according to

$$\mathbf{z}_i^{t_j+1} = \mathbf{z}_i^{t_j} + \frac{\sum_k^N w_i^k(\mathbf{z}^{t_j} - \mathbf{P}(\mathbf{z}^{t_j}))}{\sum_k^N \mathrm{e}^{-\ln(2)/\tau} w_i^k} \tag{6}$$

where $\mathbf{z}_i^{t_j}$ denotes node positions after the last update and the weight of the update for node i in step k is given by

$$w_i^k = \max\left[0, \left(1 - \frac{\left\|\mathbf{z}_i^{t_j} - \mathbf{P}(\mathbf{z}^{t_j})\right\|}{\left\|\mathbf{z}_i^{t_j} - \mathbf{z}_{i+1}^{t_j}\right\|}\right)\right] \tag{7}$$

which is only nonzero for the two closest nodes.[19] The half-life of the weight of the original path can be chosen with the parameter τ. $\mathbf{P}(\mathbf{z})$ denotes the projection of $\mathbf{z}$ on the path. In practice, weights w_i^k and the average distance from the path are accumulated between updates, which are applied for every N-th step. The initial path might be obtained, for example, by linear interpolation between end points or with some zero-temperature path optimization method (e.g., nudged elastic band method[27]). The latter might additionally already provide some mechanistic information that supports the user in the manual selection of a suitable CV space. After every update, the path is reparametrized to ensure equidistant spacing of nodes, as described in the Supporting Information. To judge the convergence of the PCV, we monitor the quantity

$$D_\mathrm{S}(z) = \left\langle \left| \frac{\nabla s(\mathbf{z})}{|\nabla s(\mathbf{z})|} \cdot \frac{\nabla U(\mathbf{x})}{|\nabla U(\mathbf{x})|} \right| \right\rangle_z \tag{8}$$

which approaches exactly zero at the transition state (TS) for an ideal CV.[25]

In this framework, nonphysical jumps in $s(\mathbf{z})$ can occur at path updates or in regions of the path with high curvature when the

https://doi.org/10.1021/acs.jctc.3c00938
J. Chem. Theory Comput. 2023, 19, 9202−9210

system shortcuts the path. While in WTM simulations, this only causes mild heating due to the discontinuity of the bias potential, it can cause numerical instability with the WTM-eABF sampling algorithm due to the coupling to the extended variable. We solve this problem by correcting the position of λ at time step $t-1$ before integrating its position to time step t according to

$$\lambda_{\text{stable}}^{t-1} = \begin{cases} s(\mathbf{z}^t) + (\lambda^{t-1} - s(\mathbf{z}^{t-1})) & \text{if } |s(\mathbf{z}^t) - s(\mathbf{z}^{t-1})| > \sigma \\ \lambda^{t-1} & \text{otherwise} \end{cases} \quad (9)$$

We always use the thermal coupling width σ as a threshold for corrections of λ^t, which we find to be a very robust choice. A numerical example of the effect of eq 9 is given in Figure 1.

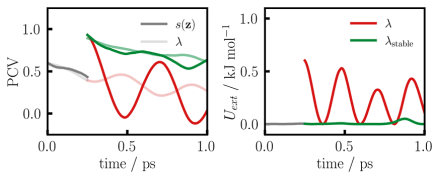


Figure 1. Numerical example for sampling a discontinuous CV with extended-system dynamics. On the left, the trajectories of $s(\mathbf{z})$ and λ are indicated by solid and transparent lines, respectively. After a step in the PCV, the conventional and stabilized trajectories are given in red and green, respectively. On the right, the corresponding coupling potential U_{ext} mediated by harmonic coupling of λ to $s(\mathbf{z})$ is shown.

Without the correction, large fluctuations of the extended variable, shown in red on the left, arise after the CV (gray) is stepped. In contrast, by correcting the position of the extended variable, its dynamics (green) are not affected, and the harmonic coupling potential, shown on the right, is continuous, which results in the superior stability of path WTM-eABF simulations.

Additionally, it was shown that a mild confinement of the distance $d(\mathbf{z})$ from the path, defined by

$$\mathrm{d}(\mathbf{z}) = \left| \mathbf{v}_1 + \frac{1}{2}\left(\frac{\sqrt{(\mathbf{v}_1\cdot\mathbf{v}_3)^2 - |\mathbf{v}_3|^2(|\mathbf{v}_1|^2 - |\mathbf{v}_2|^2)} - (\mathbf{v}_1\cdot\mathbf{v}_3)}{|\mathbf{v}_3|^2} - 1 \right)\mathbf{v}_4 \right| \quad (10)$$

where $\mathbf{v}_4 = \mathbf{z}_m - \mathbf{z}_{m-1}$ connects the closest to the second-closest path node, can reduce path shortcutting[28,29] and speed up the convergence of specific reaction channels.[18]

The central idea of path MtD/WTM is that once the path converges, the bias potential self-corrects as new Gaussians bury artifacts of sampling along the wrong path. Due to the complementary nature of both biasing strategies of the composite WTM-eABF method, simultaneously filling free energy basins and removing barriers,[20] we expect this self-correction to be much faster. In addition, to further accelerate the convergence of simulations along adaptive paths, we propose a reweighting procedure. For this purpose, the continuously sampled WTM-eABF trajectory is divided into N biased states with constant $\lambda = \lambda_i$ and a time-independent potential energy function

$$U_i(\mathbf{x}) = U(\mathbf{x}) + \frac{1}{2\beta\sigma^2}(s(\mathbf{z}) - \lambda_i)^2 + U_{\text{conf}}(\mathbf{x}) \quad (11)$$

where $U_{\text{conf}}(\mathbf{x})$ denotes some additional confinement potential, as for the distance from the path $\mathrm{d}(\mathbf{z})$.[30] This post hoc separation of the biased probability density $\rho^B(z)$ into overlapping λ-conditioned distributions $\rho^B(z|\lambda_i)$ allows for the application of popular estimators of the unbiased statistical weights of individual frames like the MBAR.[22] Note that technically the windows λ_i must be built separately for each intermediate path if it changes in updates. Besides removing potential artifacts that arise due to the application of confinements on $\mathrm{d}(\mathbf{z})$, this allows for reweighting of the PMF to any CV of choice. In the context of adaptive PCVs, we suggest applying this formalism to accelerate convergence of the PMF by mapping all data points to the final MFEP. Therefore, assuming that the underlying phase space is already sufficiently sampled, one instantly obtains the correct PMF for a new path without waiting for convergence of the WTM-eABF bias potential. Additionally, other properties like ensemble averages can be recovered independently of the CV, and path updates do not slow down their convergence.

COMPUTATIONAL DETAILS

Numerical Simulations. As a numerical test, we apply path WTM-eABF to the dynamics of a particle in a 2D MB potential, which is given by

$$U_{\text{MB}}(x,y) = B\sum_{i=1}^{4} A_i \exp[\alpha_i(x-x_i)^2 + \beta_i(x-x_i)(y-y_i) + \gamma_i(y-y_i)^2] \quad (12)$$

with $B = 1$ kJ/mol. Other numerical parameters are given in the Supporting Information. For each simulation, a single particle of mass 200 au was evolved in $U_{\text{MB}}(x, y)$ for 500 ps (1,000,000 steps) according to Langevin dynamics at 50 K with a friction constant of 0.001 fs^{-1}. MtD bias potentials were built from Gaussians with a standard deviation of 0.05 and an initial height of 0.01 kJ/mol, which were added every 25 steps. For WTM potentials, the height of new Gaussians is scaled down over the course of the simulation with an effective temperature of 5000 K. For path WTM-eABF, a fictitious particle was coupled to PCVs with $\sigma = 0.01$ and mass 25 au. The ABF was scaled with a linear ramp and only fully applied in bins with more than 50 samples. In all simulations, the bias force was accumulated on a grid with a bin width of 0.01. A guess path with 30 nodes was generated by linear interpolation between both minima. For adaptive path simulations, the path was updated every 10 ps according to eq 6. The minimum energy path (MEP) was obtained as a reference by optimizing the guess path with the nudged elastic band (NEB) method.[27] Scripts to repeat all numerical simulations are given in the Supporting Information.

Reaction Mechanism of PUS. The initial configuration of the enzyme−substrate complex was taken from the crystal structure of *Pyrococcus furiosus* box H/ACA PUS.[31,32] Two different crystal structures (PDB codes: 3HJW and 3HAY) were combined to minimize possible errors. The full enzymatic system contains 4 protein subunits (Cbf5, Gar1, Nop10, and L7Ae), as well as a guide and a substrate RNA. The uridine (U) unit in the active site was manually modified from 5-fluorouridine to native U. Charged amino acids were titrated to neutral pH using the H++ program[33] and placed in a cubic water box containing about 17,000 water molecules. To neutralize the system and set the physiological salt concentration, Mg^{2+} and Cl^- ions were added to the water box.

For classical MD simulations, the AMBER-ff19SB force field[34] was applied together with improved RNA parameters by Tan et al.[35] and the OPC 4-point water model.[36] MD simulations were performed with the OpenMM program package.[37] Electrostatic interactions were calculated by using periodic boundary conditions and particle mesh Ewald summation with a cutoff

https://doi.org/10.1021/acs.jctc.3c00938
J. Chem. Theory Comput. 2023, 19, 9202−9210

Journal of Chemical Theory and Computation pubs.acs.org/JCTC Article

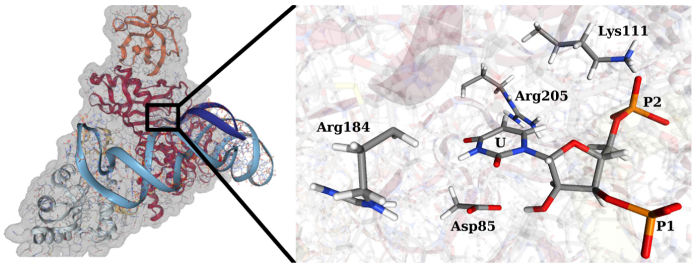


Figure 2. Fully functional *P. furiosus* box H/ACA PUS. The protein−substrate complex contains 4 protein subunits (red: catalytic Cbf5, orange: Gar1, yellow: Nop10, and gray: L7ae), as well as a H/ACA guide and substrate RNA, shown in light and dark blue, respectively (explicit water not shown). On the right, the active site is enlarged. All solid atoms are treated quantum mechanically in QM/MM simulations, while the rest of the protein−substrate complex is included in the MM region. Besides the substrate U, the QM region includes all charged protein residues of the active site, namely, Asp85, Lys111, Arg184, and Arg205.[55] To balance the charge of the QM region, two phosphates of the RNA backbone are included.

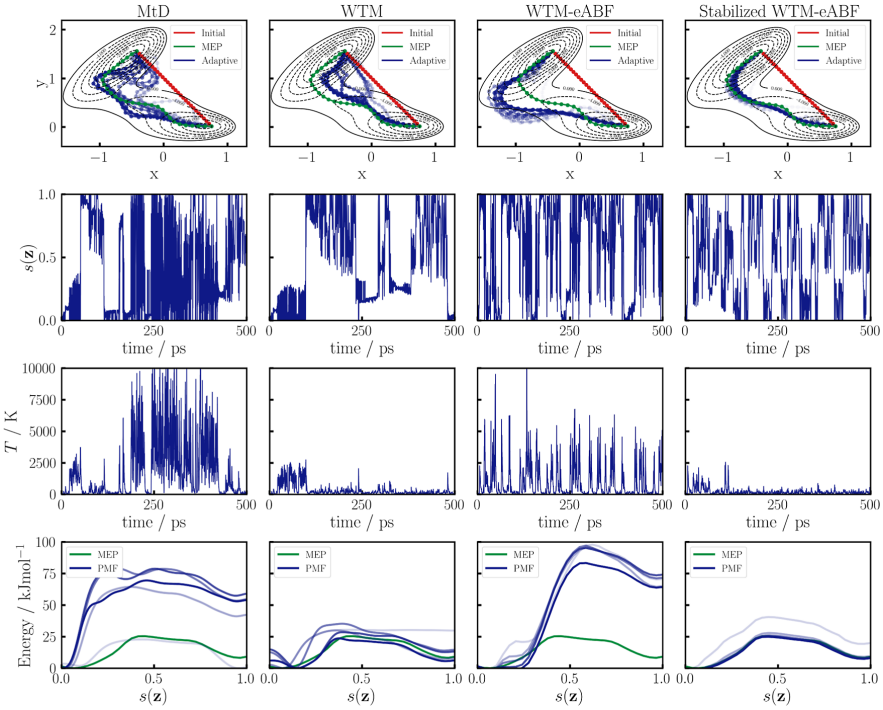


Figure 3. Sampling of a 2D MB potential with adaptive PCV and four different sampling algorithms, MtD and WTM in the first two columns and conventional WTM-eABF and stabilized WTM-eABF in the third and fourth column, respectively. In the top row, the MB potential is shown as a contour plot, with the initial path guess in red and the optimized MEP in green. Snapshots of the adaptive path in 100 ps intervals are shown in blue, losing transparency over time. In the second and third rows, the trajectory of the PCV and rolling temperature average taken over 1 ps are shown, respectively. The last row shows the current PMF estimation every 100 ps, earlier PMFs being more transparent, the potential energy along the MEP shown in green.

of 12 Å. Water molecules and H-bonds were constrained using the SETTLE[38] and SHAKE[39] algorithms, respectively. Time integration was performed at 300 K with a Langevin integrator using a time step of 2 fs and a friction constant of 1 ps^{-1}. Atmospheric pressure was set using a Monte Carlo Barostat.[40] The initial system was minimized to a tolerance of 10 kJ/mol and carefully heated from 5 to 300 K over 60 stages, 1000 steps each. Afterward, the system was simulated for 600 ns. In the period from 100 to 300 ns, the temperature was increased to 320 K to enhance sampling and enable penetration of water into the active site. Further details are provided in the Supporting Information.

QM/MM simulations were performed in our in-house program package FermiONs++.[41−45] The QM system was

https://doi.org/10.1021/acs.jctc.3c00938
J. Chem. Theory Comput. 2023, 19, 9202−9210

centered inside the simulation box, and interactions between the QM and MM subsystems were treated with electrostatic embedding using a cutoff of 10 Å.[46] For technical reasons, the TIP3P water model[47] was applied instead of the OPC. For efficient evaluation of the QM/MM Hamiltonian Grimme's PBEh-3c DFT functional[48] was applied. Significant further speed-ups were achieved by fast evaluation of seminumerical exact exchange with the sn-LinK method[45,49] and evaluation of the Coulomb energy in the RI-J approximation.[50] Additionally, SCF convergence acceleration was achieved by using accurate guess densities obtained from the previous nine density matrices according to the extended-Lagrangian extrapolation method.[51,52] The final QM/MM system, which is shown in Figure 2, contained a total of 99 QM and more than 100,000 MM atoms. Geometry optimizations of the QM/MM system were performed in a PyChemshell[53,54] interface to FermiONs++.

Unconstrained ab initio MD simulations were performed with a time step of 0.5 fs. The temperature was controlled with a Langevin thermostat at 300 K. All atoms farther than 20 Å from the QM region have been frozen. A benchmark of the influence of the PBEh-3c functional, the QM size, and the electrostatic cutoff on the reaction energy of C1′−N1 bond cleavage and proton transfer from H2′ to Asp85-O is given in the Supporting Information.

For the calculation of PMFs, our own Python-based implementation of the path WTM-eABF-enhanced sampling algorithm and MBAR was applied. The full source code is available in the adaptive-sampling package[30] at https://github.com/ochsenfeld-lab/adaptive_sampling. PMFs were calculated from 10 independent walkers starting from different protein conformations that were picked from the last 100 ns of the MM MD trajectory in an equidistant manner. Trajectories of simulations of the rebound and glycal schemes extend to a combined total of >600 ps and >500 ps, respectively. Reaction and activation-free energies were obtained from the PMF as proposed by Dietschreit and co-workers.[25,56] More details are given in the Supporting Information.

RESULTS AND DISCUSSION

Path WTM-eABF on a Numerical Potential. We first demonstrate the benefits of the path WTM-eABF algorithm compared to conventional MtD/WTM for a 2D MB potential. We simulate the dynamics of a single particle at 50 K for 500 ps with MtD, WTM, WTM-eABF, and the stabilized variant of WTM-eABF. The MB potential energy surface is shown as a contour plot in the first row of Figure 3. We apply an adaptive PCV as described in Section 2. The initial path, colored red, is a linear interpolation between both minima. We note that for the MB potential the MEP and MFEP are identical, except for small thermal fluctuations of the latter, since the MB energy surface is harmonic in orthogonal direction to the MEP. For the same reason, the analytical probability density is sharply peaked along the MEP and the PMF along the MFEP is approximated well by the potential energy curve along the MEP. Therefore, we use the MEP as a reference, which is shown in green. In blue, snapshots of the adaptive path indicate its evolution over the course of the simulation. The lower rows of Figure 3 show trajectories of the PCV, running temperature averages, and the evolution of the PMF in 100 ps intervals with reference to the MEP potential energy curve (green). The system is initialized in the left minimum at $s(\mathbf{z}) \approx 0$. Note that the initial path guess is orthogonal to the MEP. Therefore, in MtD and WTM simulations, a steep bias potential is built initially until this orthogonal barrier can be crossed. The system escapes the first minimum at about 50/100 ps for MtD/WTM, respectively. With MtD, large temperature fluctuations are observed throughout the simulation because of the constant addition of repulsive Gaussian hills that drive the system out of equilibrium. In contrast, the WTM simulation is stable after the first 100 ps, as new Gaussians are scaled down. However, the artificial bias potential that builds up in the initial 100 ps hinders the back reaction to the first minimum, which occurs only shortly before the end of the simulation. Therefore, the adaptive path and also the PMF, which are estimated directly from the bias potential, are both not fully converged after 500 ps. Without stabilization (third column of Figure 3), the WTM-eABF simulation shows large temperature fluctuations as well that cause significant artifacts to the adaptive path and hinder the convergence of the PMF. As discussed in Section 2, these fluctuations are caused by the heating of the extended variable due to the discontinuous nature of the PCV at the path updates. In contrast, with the proposed WTM-eABF stabilization (eq 9), the adaptive path as well as PMF converge rapidly and reproduce the reference almost exactly after less than 200 ps. The dramatically increased performance compared to MtD/WTM can be attributed to the faster adaptation of the WTM-eABF algorithm to path updates due to the combination of two complementary biasing strategies. We note that it is still advisable to perform frequent path updates to avoid the accumulation of large bias potentials along a bad initial guess path. Overall, this shows that the (stabilized) path WTM-eABF is able to significantly outperform MtD/WTM both in terms of robustness and sampling efficiency.

In the following, we discuss additional benefits that arise from the combination of WTM-eABF sampling with postprocessing using the MBAR. To this end, we sample along a static, linear PCV using WTM-eABF. On the left of Figure 4, the MB potential is shown along with the PCV and the reference MFEP. Additionally, data points of the trajectory are given in blue. The PMF along the PCV in red on the right of Figure 4, shows large deviations from the reference PMF along the MFEP. However, MBAR allows for the mapping of data points to any reaction coordinate of choice. Therefore, by reweighting the simulation to the MFEP, the reference PMF can be largely recovered. Only at $s(\mathbf{z}) \approx 0.2$ an artifact arises, where the sampling has no overlap with the MFEP. In general, this shows that, as the MBAR returns the unbiased weight of each data frame, properties like reaction free energies or ensemble averages are calculated from data frames alone. Therefore, there is no strict need to converge the

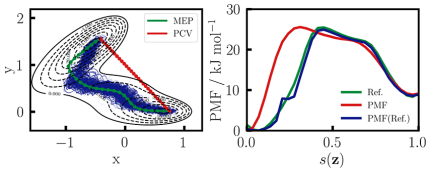


Figure 4. WTM-eABF sampling along a static linear path. On the left, the MB potential energy surface is shown as a contour plot, with PCV nodes in red and the converged MFEP in green. Data points of the WTM-eABF simulation are given in blue. On the right, a reference PMF is shown in green, together with the PMF along the linear guess path in red. The PMF along the MFEP, obtained by reweighting the simulation with MBAR, is given in blue.

https://doi.org/10.1021/acs.jctc.3c00938
J. Chem. Theory Comput. 2023, 19, 9202−9210

bias potential after path updates, which, in turn, does not reduce the convergence rate of WTM-eABF/MBAR. In the next section, we will show the beneficial properties of the path WTM-eABF sampling algorithm on a real biocatalytic reaction mechanism.

Reaction Mechanism of PUS. To show how the path WTM-eABF method can be applied to explore enzymatic reaction mechanisms, we investigate the first step of the catalytic mechanism of PUS, which enables the site-specific modification of U to pseudouridine (Ψ) in various types of RNA. Because of the highly conserved active site of this family of proteins, which always includes an essential aspartate (Asp85), it is assumed that all PUS enzymes operate by one uniform reaction mechanism.[55] Recently, in products of PUS, besides the major *ribo* sugar, an *arabino* isomer, which differs in the stereochemistry at C2′, was observed.[57] Additionally, a large kinetic isotope effect for the proton at C2′ was reported.[58] Both observations could be explained by the reaction over a glycal intermediate.[58,59] However, it was also suggested that the reaction instead proceeds over a direct rebound scheme and that the catalytic role of Asp85 is merely to provide conformational strain to the ribo sugar.[60] An overview of both reaction mechanisms is given in Figure 5.

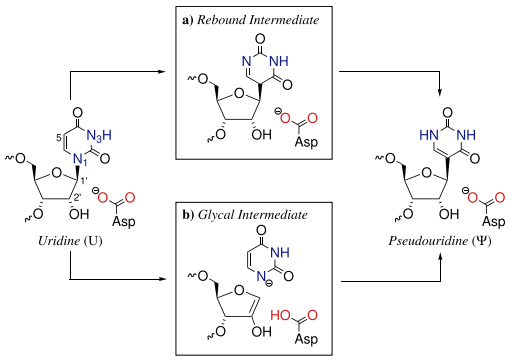


Figure 5. Suggested reaction mechanisms for the conversion of *U* to Ψ catalyzed by PUS. The reaction might run over (a) a rebound intermediate, where the C1′−C5 bond forms directly after C1′−N1 bond cleavage, and (b) a glycal intermediate involving deprotonation of C2′ by Asp85.

We apply path WTM-eABF together with QM/MM MD to explore the formation of both intermediates using the relatively cost-effective PBEh-3c DFT functional and 99 QM atoms. This setup is chosen to reach total simulation times of hundreds of picoseconds, which are necessary to converge PMFs for such a large system. However, we note that due to the inherent inaccuracy of the PBEh-3c functional and the limited QM region size, absolute reaction barriers tend to be overestimated. CV spaces for the calculation of PCVs contain 3 or 5 bond distances, respectively. The breaking U C1′−N1 and forming Ψ C1′−C5 bonds are always included. Additionally, we add the C1′−N3 bond to gain optimal control over the rotation of U. To describe the mechanism over a glycal intermediate, additional slow degrees of freedom that account for proton transfer from C2′ to Asp85 are taken into account. A schematic representation of the CV spaces is given in Figure 6.

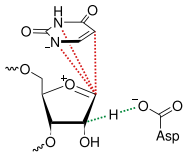


Figure 6. Illustration of the CV space for the calculation of PCVs for the PUS reaction mechanism. For the rebound mechanism, three bond distances marked in red are used. The glycal mechanism additionally involves proton transfer from C2′ to Asp85, with corresponding bond distances marked in green.

Rebound Mechanism. First, we consider the rebound mechanism, as proposed by Kiss et al.,[60] where C1′−N1 bond cleavage is followed by direct formation of the characteristic Ψ C1′−C5 bond. In Figure 7, on the left, the final path is projected

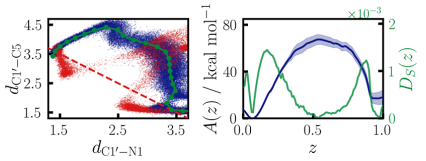


Figure 7. On the left, a 2D projection of path nodes on $d_{C1'-N1}$ and $d_{C1'-C5}$ is shown in green, with every 20th data point of all combined path WTM-eABF trajectories in blue. Red points denote data points obtained for sampling along a CV of the form $\xi(\mathbf{x}) = d_{C1'-N1} - d_{C1'-C5}$ (indicated by the dashed red line). On the right, the PMF obtained from all combined simulations is shown in blue; in green, the orthogonality measure $D_S(z)$ (eq 8). Light blue area denotes the 95% confidence interval.

on the C1′−N1 and C1′−C5 bond distances, with trajectory data points shown in blue. On the right are the obtained PMF together with the CV criterion $D_S(z)$ (eq 8). The obtained reaction mechanism can be divided into three steps. First, both the C1′−N1 and C1′−C5 bonds elongate as U unbinds from the glycase backbone building an uridilate ion. For this step, we obtain a high activation-free energy of about 69 kcal/mol, which can be rationalized by the involved charge separation. At a C1′−N1 bond distance of about 2.5 Å, the uridilate ion begins to rotate under shortening of the C1′−C5 bond distance. Finally, the C1′−C5 bond forms, building the stable rebound intermediate. The CV criterion has three clear minima at the reactant, product, and TS, respectively. That it is zero at the TS indicates the good quality of the obtained PCV and confirms that the TS is successfully located. We note that this rotating motion cannot be fully captured by a simple linear combination of the form $d_{C1'-N1} - d_{C1'-C5}$, which we mark in red in Figure 7. On the contrary, the application of this very common choice of CV leads to sampling defects. Due to the wrong projection of the bias force, simultaneous $d_{C1'-N1}$ elongation and $d_{C1'-C5}$ shortening are enforced, which leads to a bending motion until rapid breaking of the C1′−N1 bond occurs. Also, various irrelevant side reactions are observed, like the formation of C1′−O2 bonds, which make simulations unstable. With a mild confinement on the distance from the path such side reactions are fully suppressed in PCV simulations.

Glycal Mechanism. The large activation-free energy obtained for the formation of the rebound intermediate indicates that in this mechanism, the catalytic role of Asp85 is not correctly

https://doi.org/10.1021/acs.jctc.3c00938
J. Chem. Theory Comput. 2023, 19, 9202−9210

captured. Recently, a large kinetic isotope effect was shown for exchanging H2′ with deuterium, an observation that might be explained by H2′ proton transfer to Asp85, forming a glycal intermediate. Therefore, we build a new path guess, where parts of the final path of the rebound mechanism (without C1′−C5 bond formation) are coupled to proton transfer from C2′ to Asp85. A 2D projection of the new CV space is shown in Figure 8 on the left, with every 20th data point of the simulation shown

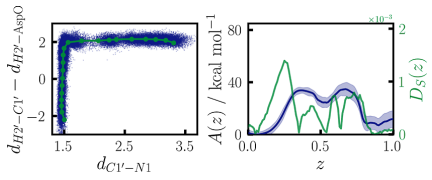


Figure 8. On the left, a 2D projection of path nodes on $d_{C1'-N1}$ and $d_{H2'-C1'} - d_{H2'-AspO}$ is shown in green, with every 20th data point of all combined trajectories in blue. On the right, the obtained PMF is shown in blue, and in green, the orthogonality measure $D_S(z)$ (eq 8). Light blue area denotes the 95% confidence interval.

in blue and the final path in green. Clearly, a sequential mechanism is observed, where proton transfer from C2′ to Asp85 (up direction) is followed by C1′−N1 bond cleavage (left to right). The new PMF and corresponding CV criterion are given in Figure 8, on the right. Two distinct TSs and one intermediate minimum, resembling the deprotonated ribo sugar before C1′−N1 bond cleavage, are observed. Each of them is confirmed by the clear minima of $D_S(z)$. With an activation-free energy of about 35 kcal/mol, the initial proton transfer is the rate-limiting step and significantly activates sequential C1′−N1 bond cleavage, for which a small remaining activation-free energy of around 10 kcal/mol is obtained. The significant relative reduction of the activation-free energy compared to the rebound mechanism by over 30 kcal/mol displays the large catalytic effect on C1′−N1 bond breaking. At the same time, we concede that the absolute activation barrier obtained is still too high to be crossed on relevant time scales under biological conditions. However, we also expect it to be overestimated due to the inherent inaccuracy of the PBEh-3c functional and the limited size of the QM subsystem (see Supporting Information).

Overall, this shows how adaptive PCVs enable the detailed study of nonlinear, multistep molecular transitions. The above example depicts how their application, together with the highly efficient path WTM-eABF algorithm, facilitates the systematic exploration of reaction mechanisms, even using costly QM/MM simulations. In addition, we show how the adaptive path can directly yield mechanistic information, like the activation of C1′−N1 bond breaking by deprotonation of C2′ in PUS enzymes.

CONCLUSIONS

We have shown the benefits of combining the highly efficient WTM-eABF sampling algorithm with adaptive PCVs. To this end, we provide an implementation of path WTM-eABF with a new stabilization algorithm that ensures the temperature stability of simulations even if path shortcutting or path updates cause sudden jumps in the PCV. Additionally, we show how reweighting data to an updated path with the MBAR can be used to speed up the convergence of PMFs along adaptive paths. Overall, we argue that WTM-eABF, PCVs, and the MBAR estimator elegantly complement each other and together offer a highly competitive approach to the systematic investigation of reaction mechanisms in complicated biochemical systems.

We apply these methods to investigate the reaction mechanism of PUS and show how path WTM-eABF enables the exploration of challenging nonlinear molecular motions like uridine rotation in the rebound mechanism. Furthermore, we obtain a significantly reduced activation free energy for a glycal mechanism where proton transfer of H2′ to the essential Asp85 activates C1′−N1 bond breaking, a result that is in line with the experimental observation of a large deuterium kinetic isotope effect for H2′.[58] In a future study, we plan to use the presented framework to provide a full mechanistic picture of PUS.

ASSOCIATED CONTENT

Supporting Information

The data that supports the findings of this study are available from the corresponding author upon reasonable request. The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.3c00938.

Additional information on the implementation of PCVs, numerical parameters of the MB potential, initial MM simulation of PUS, accuracy benchmarks for the QM/MM setup, and details on QM/MM free energy simulations (PDF)

Python scripts to repeat and analyze MtD, WTM, and WTM-eABF simulations along static and adaptive PCVs on a numerical MB potential (ZIP)

AUTHOR INFORMATION

Corresponding Author

Christian Ochsenfeld − *Chair of Theoretical Chemistry, Department of Chemistry, LMU Munich, München D-81377, Germany; Max Planck Institute for Solid State Research, Stuttgart D-70569, Germany;* orcid.org/0000-0002-4189-6558; Email: christian.ochsenfeld@uni-muenchen.de

Author

Andreas Hulm − *Chair of Theoretical Chemistry, Department of Chemistry, LMU Munich, München D-81377, Germany;* orcid.org/0000-0003-1268-7578

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.3c00938

Funding

Open access funded by Max Planck Society.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank J. Kussmann (LMU Munich) for providing a development version of the FermiONs++ program package and Alexandra Stan-Bernhardt (LMU Munich) for useful comments and discussions. Financial support was provided by the "Deutsche Forschungsgemeinschaft" (DFG, German Research Foundation) within SFB 1309-325871075 "Chemical Biology of Epigenetic Modifications". C.O. acknowledges further support as Max-Planck-Fellow at the MPI-FKF Stuttgart.

https://doi.org/10.1021/acs.jctc.3c00938
J. Chem. Theory Comput. 2023, 19, 9202−9210

REFERENCES

(1) Chipot, C.; Pohorille, A. *Free Energy Calculations: Theory and Applications in Chemistry and Biology*; Springer: Berlin Heidelberg, 2007; Vol. *86*.

(2) Dietschreit, J. C.; von der Esch, B.; Ochsenfeld, C. Exponential averaging versus umbrella sampling for computing the QM/MM free energy barrier of the initial step of the desuccinylation reaction catalyzed by sirtuin 5. *Phys. Chem. Chem. Phys.* **2022**, *24*, 7723−7731.

(3) Senn, H. M.; Thiel, W. QM/MM methods for biomolecular systems. *Angew. Chem., Int. Ed.* **2009**, *48*, 1198−1229.

(4) Chipot, C. Frontiers in Free-Energy Calculations of Biological Systems. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2014**, *4*, 71−89.

(5) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Chem. Phys.* **1977**, *23*, 187−199.

(6) Darve, E.; Rodriguez-Gomez, D.; Pohorille, A. Adaptive biasing force method for scalar and vector free energy calculations. *J. Chem. Phys.* **2008**, *128*, 144120.

(7) Bussi, G.; Laio, A.; Parrinello, M. Equilibrium free energies from nonequilibrium metadynamics. *Phys. Rev. Lett.* **2006**, *96*, 090601.

(8) Lesage, A.; Lelievre, T.; Stoltz, G.; Henin, J. Smoothed biasing forces yield unbiased free energies with the extended-system adaptive biasing force method. *J. Phys. Chem. B* **2017**, *121*, 3676−3685.

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

(10) Barducci, A.; Bonomi, M.; Parrinello, M. Metadynamics. *Rev. Comput. Mol. Sci.* **2011**, *1*, 826−843.

(11) Köppen, M. The curse of dimensionality. *5th Online World Conference on Soft Computing in Industrial Applications (WSCS)*, 2000; pp 4−8.

(12) Mendels, D.; Piccini, G.; Parrinello, M. Collective Variables from Local Fluctuations. *J. Phys. Chem. Lett.* **2018**, *9*, 2776−2781.

(13) Wang, Y.; Ribeiro, J. M. L.; Tiwary, P. Past−future information bottleneck for sampling molecular reaction coordinate simultaneously with thermodynamics and kinetics. *Nat. Commun.* **2019**, *10*, 3573.

(14) Sun, L.; Vandermause, J.; Batzner, S.; Xie, Y.; Clark, D.; Chen, W.; Kozinsky, B. Multitask Machine Learning of Collective Variables for Enhanced Sampling of Rare Events. *J. Chem. Theory Comput.* **2022**, *18*, 2341−2353.

(15) Wang, D.; Tiwary, P. State predictive information bottleneck. *J. Chem. Phys.* **2021**, *154*, 134111.

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

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

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

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

(20) Fu, H.; Zhang, H.; Chen, H.; Shao, X.; Chipot, C.; Cai, W. Zooming across the free-energy landscape: shaving barriers, and flooding valleys. *J. Phys. Chem. Lett.* **2018**, *9*, 4738−4745.

(21) Fu, H.; Shao, X.; Cai, W.; Chipot, C. Taming rugged free energy landscapes using an average force. *Acc. Chem. Res.* **2019**, *52*, 3254−3264.

(22) Shirts, M. R.; Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* **2008**, *129*, 124105.

(23) Laio, A.; Gervasio, F. L. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Rep. Prog. Phys.* **2008**, *71*, 126601.

(24) Comer, J.; Gumbart, J. C.; Henin, J.; Lelievre, T.; Pohorille, A.; Chipot, C. The adaptive biasing force method: Everything you always wanted to know but were afraid to ask. *J. Phys. Chem. B* **2015**, *119*, 1129−1151.

(25) Dietschreit, J. C.; Diestler, D. J.; Hulm, A.; Ochsenfeld, C.; Gómez-Bombarelli, R. From free-energy profiles to activation free energies. *J. Chem. Phys.* **2022**, *157*, 084113.

(26) Rogal, J.; Schneider, E.; Tuckerman, M. E. Neural-network-based path collective variables for enhanced sampling of phase transformations. *Phys. Rev. Lett.* **2019**, *123*, 245701.

(27) Jónsson, H.; Mills, G.; Jacobsen, K. W. *Classical and Quantum Dynamics in Condensed Phase Simulations*; World Scientific: Singapore, 1998; pp 385−404.

(28) Chen, H.; Ogden, D.; Pant, S.; Cai, W.; Tajkhorshid, E.; Moradi, M.; Roux, B.; Chipot, C. A companion guide to the string method with swarms of trajectories: Characterization, performance, and pitfalls. *J. Chem. Theory Comput.* **2022**, *18*, 1406−1422.

(29) Kolossváry, I.; Sherman, W. Comprehensive Approach to Simulating Large Scale Conformational Changes in Biological Systems Utilizing a Path Collective Variable and New Barrier Restraint. *J. Phys. Chem. B* **2023**, *127*, 5214.

(30) Hulm, A.; Dietschreit, J. C.; Ochsenfeld, C. Statistically optimal analysis of the extended-system adaptive biasing force (eABF) method. *J. Chem. Phys.* **2022**, *157*, 024110.

(31) Liang, B.; Zhou, J.; Kahen, E.; Terns, R. M.; Terns, M. P.; Li, H. Structure of a functional ribonucleoprotein pseudouridine synthase bound to a substrate RNA. *Nat. Struct. Mol. Biol.* **2009**, *16*, 740−746.

(32) Duan, J.; Li, L.; Lu, J.; Wang, W.; Ye, K. Structural mechanism of substrate RNA recruitment in H/ACA RNA-guided pseudouridine synthase. *Mol. Cell* **2009**, *34*, 427−439.

(33) Anandakrishnan, R.; Aguilar, B.; Onufriev, A. V. H++ 3.0: automating p K prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Res.* **2012**, *40*, W537−W541.

(34) Tian, C.; Kasavajhala, K.; Belfon, K. A. A.; Raguette, L.; Huang, H.; Migues, A. N.; Bickel, J.; Wang, Y.; Pincay, J.; Wu, Q.; Simmerling, C. ff19SB: Amino-acid-specific protein backbone parameters trained against quantum mechanics energy surfaces in solution. *J. Chem. Theory Comput.* **2020**, *16*, 528−552.

(35) Tan, D.; Piana, S.; Dirks, R. M.; Shaw, D. E. RNA force field with accuracy comparable to state-of-the-art protein force fields. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, *115*, E1346−E1355.

(36) Izadi, S.; Anandakrishnan, R.; Onufriev, A. V. Building water models: a different approach. *J. Phys. Chem. Lett.* **2014**, *5*, 3863−3871.

(37) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* **2017**, *13*, No. e1005659.

(38) Miyamoto, S.; Kollman, P. A. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* **1992**, *13*, 952−962.

(39) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977**, *23*, 327−341.

(40) Åqvist, J.; Wennerström, P.; Nervall, M.; Bjelic, S.; Brandsdal, B. O. Molecular dynamics simulations of water and biomolecules with a Monte Carlo constant pressure algorithm. *Chem. Phys. Lett.* **2004**, *384*, 288−294.

(41) Kussmann, J.; Ochsenfeld, C. Pre-selective screening for matrix elements in linear-scaling exact exchange calculations. *J. Chem. Phys.* **2013**, *138*, 134114.

(42) Kussmann, J.; Ochsenfeld, C. Preselective screening for linear-scaling exact exchange-gradient calculations for graphics processing units and general strong-scaling massively parallel calculations. *J. Chem. Theory Comput.* **2015**, *11*, 918−922.

(43) Kussmann, J.; Ochsenfeld, C. Hybrid CPU/GPU integral engine for strong-scaling ab initio methods. *J. Chem. Theory Comput.* **2017**, *13*, 3153−3159.

(44) Laqua, H.; Thompson, T. H.; Kussmann, J.; Ochsenfeld, C. Highly efficient, linear-scaling seminumerical exact-exchange method

https://doi.org/10.1021/acs.jctc.3c00938
J. Chem. Theory Comput. 2023, 19, 9202−9210

for graphic processing units. *J. Chem. Theory Comput.* **2020**, *16*, 1456−1468.
(45) Laqua, H.; Kussmann, J.; Ochsenfeld, C. Accelerating semi-numerical Fock-exchange calculations using mixed single- and double-precision arithmethic. *J. Chem. Phys.* **2021**, *154*, 214116.
(46) Field, M. J.; Bash, P. A.; Karplus, M. A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations. *J. Comput. Chem.* **1990**, *11*, 700−733.
(47) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926−935.
(48) Grimme, S.; Brandenburg, J. G.; Bannwarth, C.; Hansen, A. Consistent structures and interactions by density functional theory with small atomic orbital basis sets. *J. Chem. Phys.* **2015**, *143*, 054107.
(49) Laqua, H.; Dietschreit, J. C.; Kussmann, J.; Ochsenfeld, C. Accelerating Hybrid Density Functional Theory Molecular Dynamics Simulations by Seminumerical Integration, Resolution-of-the-Identity Approximation, and Graphics Processing Units. *J. Chem. Theory Comput.* **2022**, *18*, 6010−6020.
(50) Kussmann, J.; Laqua, H.; Ochsenfeld, C. Highly efficient resolution-of-identity density functional theory calculations on central and graphics processing units. *J. Chem. Theory Comput.* **2021**, *17*, 1512−1521.
(51) Niklasson, A. M.; Tymczak, C.; Challacombe, M. Time-reversible Born-Oppenheimer molecular dynamics. *Phys. Rev. Lett.* **2006**, *97*, 123001.
(52) Peters, L. D.; Kussmann, J.; Ochsenfeld, C. Efficient and Accurate Born−Oppenheimer Molecular Dynamics for Large Molecular Systems. *J. Chem. Theory Comput.* **2017**, *13*, 5479−5485.
(53) Kästner, J.; Carr, J. M.; Keal, T. W.; Thiel, W.; Wander, A.; Sherwood, P. DL-FIND: An Open-Source Geometry Optimizer for Atomistic Simulations. *J. Phys. Chem. A* **2009**, *113*, 11856−11865.
(54) Lu, Y.; Farrow, M. R.; Fayon, P.; Logsdail, A. J.; Sokol, A. A.; Catlow, C. R. A.; Sherwood, P.; Keal, T. W. Open-Source, python-based redevelopment of the ChemShell multiscale QM/MM environment. *J. Chem. Theory Comput.* **2019**, *15*, 1317−1328.
(55) Hamma, T.; Ferré-D'Amaré, A. R. Pseudouridine synthases. *Chem. Biol.* **2006**, *13*, 1125−1135.
(56) Dietschreit, J. C.; Diestler, D. J.; Ochsenfeld, C. How to obtain reaction free energies from free-energy profiles. *J. Chem. Phys.* **2022**, *156*, 114105.
(57) Miracco, E. J.; Mueller, E. G. The products of 5-fluorouridine by the action of the pseudouridine synthase TruB disfavor one mechanism and suggest another. *J. Am. Chem. Soc.* **2011**, *133*, 11826−11829.
(58) Veerareddygari, G. R.; Singh, S. K.; Mueller, E. G. The pseudouridine synthases proceed through a glycal intermediate. *J. Am. Chem. Soc.* **2016**, *138*, 7852−7855.
(59) Kiss, D. J.; Oláh, J.; Tóth, G.; Menyhárd, D. K.; Ferenczy, G. G. Quantum chemical calculations support pseudouridine synthase reaction through a glycal intermediate and provide details of the mechanism. *Theor. Chem. Acc.* **2018**, *137*, 162.
(60) Kiss, D. J.; Oláh, J.; Tóth, G.; Varga, M.; Stirling, A.; Menyhárd, D. K.; Ferenczy, G. G. The structure-derived mechanism of box H/ACA Pseudouridine synthase offers a plausible paradigm for programmable RNA editing. *ACS Catal.* **2022**, *12*, 2756−2769.

https://doi.org/10.1021/acs.jctc.3c00938
J. Chem. Theory Comput. 2023, 19, 9202−9210

Supporting Information: Improved Sampling of Adaptive Path Collective Variables by Stabilized Extended-System Dynamics

Andreas Hulm,[1] Christian Ochsenfeld[1,2,*]

[1]Chair of Theoretical Chemistry, Department of Chemistry,
University of Munich (LMU), Butenandtstr. 7, D-81377 München, Germany
[2]Max Planck Institute for Solid State Research, Heisenbergstr. 1, D-70569 Stuttgart, Germany
*E-Mail: christian.ochsenfeld@uni-muenchen.de

Contents

1 Implementation of Path Collective Variables

In practice, to avoid numerical problems if the system leaves the path ($s(\mathbf{z}) < 0$ or $s(\mathbf{z}) > 1$) we add one boundary node at each side by linear extrapolation of the outer two nodes. Therefore, eq. 5 of the main manuscript is calculated including the boundary nodes and m/M is replaced by $(m-1)/M$ to ensure that $s(\mathbf{z}) = 0$ if $\mathbf{z}$ is equal to the first original node vector and $s(\mathbf{z}) = 1$ if $\mathbf{z}$ is equal to the last original node vector. Additionally, to keep the system in the range of interest in path WTM-eABF simulations, the fictitious particle λ is always confined to the range $0 <= s(\mathbf{z}) <= 1$ with harmonic wall potentials.

One inherent problem of PCVs is there dependence on the definition of path nodes. If the nodes along the path are ill-defined, for example due to large kinks in the path or cluttering of parts of the path with multiple nodes, the PCV will be ill-defined as well. Therefore, it is necessary to perform a reparametrization step after every path update [1]. First, the path is smoothed according to

$$\mathbf{z}_i^* = (1-s)\mathbf{z}_i + \frac{s}{2}(\mathbf{z}_{i-1} + \mathbf{z}_{i+1}) \tag{S1}$$

where s denotes a damping factor for the smoothing which ranges from 0 (no smoothing) to 1 (linear interpolation between neighbor nodes). Second, equidistant spacing of nodes is ensured via

$$\mathbf{z}_i^{**} = \mathbf{z}_{j-1}^* + (s(i) - L(j-1))\frac{\mathbf{z}_j^* - \mathbf{z}_{j-1}^*}{|\mathbf{z}_j^* - \mathbf{z}_{j-1}^*|}. \tag{S2}$$

Here $s(m) = (m-1)\frac{L(N)}{N-1}$, with total length of path $L(N)$, denotes the position of node m on the path under equidistant spacing and $L(m)$ is actual position of node m along the path. Index k is such that $L(k-1) < s(m) \leq L(k)$. Eq. S2 is iterated until the change in the total path length drops below some tolerance (e.g. 0.001).

2 Numerical Parameters for the Müller-Brown Potential

The 2D Müller-Brown potential energy surface is defined by

$$U_{MB}(x,y) = B\sum_{i=1}^{4} A_i \exp\left[\alpha_i(x-x_i)^2 + \beta_i(x-x_i)(y-y_i) + \gamma_i(y-y_i)^2\right] \tag{S3}$$

with $B = 1\,\mathrm{kJ/mol}$ and other numerical parameters given in Table S1.

Table S1: Applied parameters of MB potential.

i	A_i	α_i	β_i	γ_i	x_i	y_i
1	-40.0	-1.0	0.0	-10.0	1.0	0.0
2	-10.0	-1.0	0.0	-10.0	0.0	0.5
3	-34.0	-6.5	11.0	-6.5	-0.5	1.5
4	3.0	0.7	0.6	0.7	-1.0	1.0

3 Initial MM Simulation of the Enzyme-Substrate Complex

It has been shown experimentally that deprotonation or reprotonation of C2' is partially rate limiting for the reaction mechanisms of PUS. Furthermore, experimental evidence points towards direct deprotonation by the nearby catalytic Asp85 [2]. Therefore, we consider the AspO-C2'H distance as a first indicator of potential reactivity. Specifically, only configurations where this distance is smaller than 4 Å are considered as suitable starting points for further QM/MM investigations of the reaction mechanism.

The top left panel of figure S1 shows the rolling average of the temperature over the course of a 600 ns MM-MD simulation. Between 100 and 300 ns the temperature is increased to enhance sampling and enable penetration of the water into the active region. The backbone RMSDs of individual protein subunits shown in the bottom left panel are not affected by the heating period and always below 2 Å. Finally, the top right panel shows the AspO-C2'H distance. During the first 100 ns the system is trapped in a metastable unreactive state because of electrostatic attraction of the negatively charged Asp85 to the nearby positively charged Arg184. After 100 ns upon heating of the system to 320 K the mobility of Asp85 increases, and it enters a second metastable configuration. The electrostatic attraction of Asp85 to Arg184 is reduced by water that has penetrated the active site in the first few hundred ns of the simulation, while a stabilizing H-bond to O2'H of the nucleoside backbone forms. Thus, 10 equidistant snapshots are taken from the last 100 ns of the MM MD simulation for further QM/MM simulations. Overall, the final structure agrees well with the X-Ray structure of the active site, which confirms that these configurations can be regarded as realistic starting point for the mechanistic study.

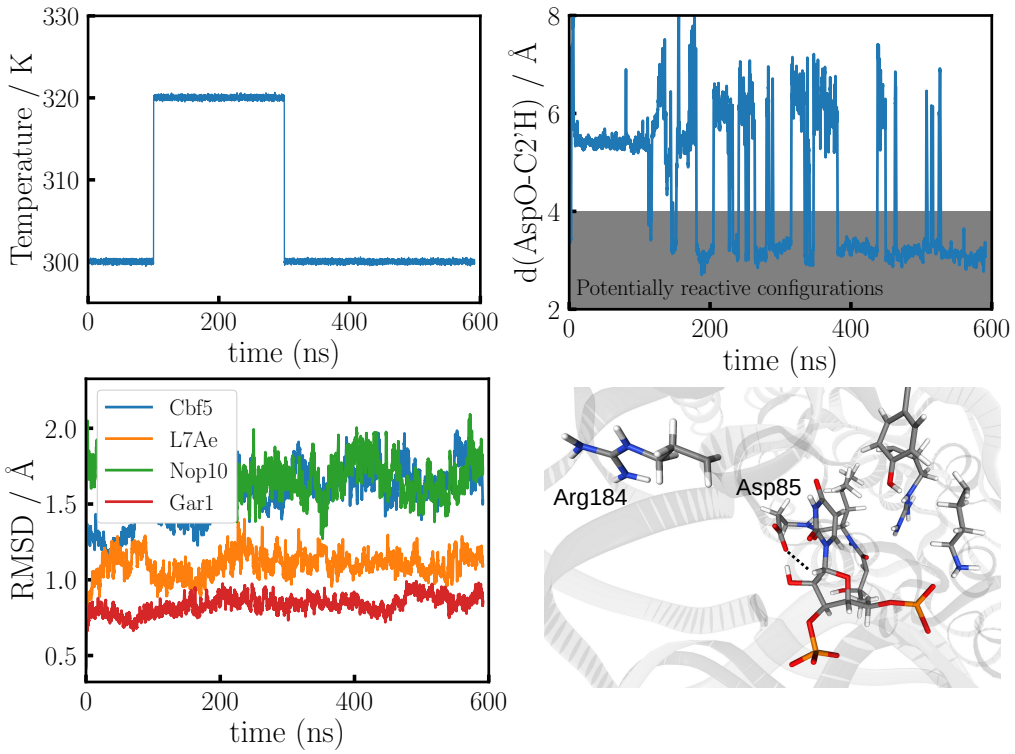


Figure S1: On the top left the rolling average of the temperature during 600 ns MM-MD simulation is shown. In the time period from 100 to 300 ns the system is heated from 300 to 320 K to enhance sampling. The bottom left panel shows the rolling average of the individual backbone RMSDs of all 4 protein subunits. Cbf5 in blue, L7Ae in orange, Nop10 in green and Gar1 in red. On the top right the rolling average of the O-H distance between Asp85-O1/O2 and U-C2'H during 600 ns MM-MD simulation. To take into account the rotation of the carboxyl group always the distance to the closer carboxylic oxygen is plotted. The active site of the protein is shown on the bottom right, with the target U, the negatively charged catalytic Asp85 and the nearby positively charged Arg184. The dashed line shows the AspO-C2'H distance.

4 QM/MM Accuracy

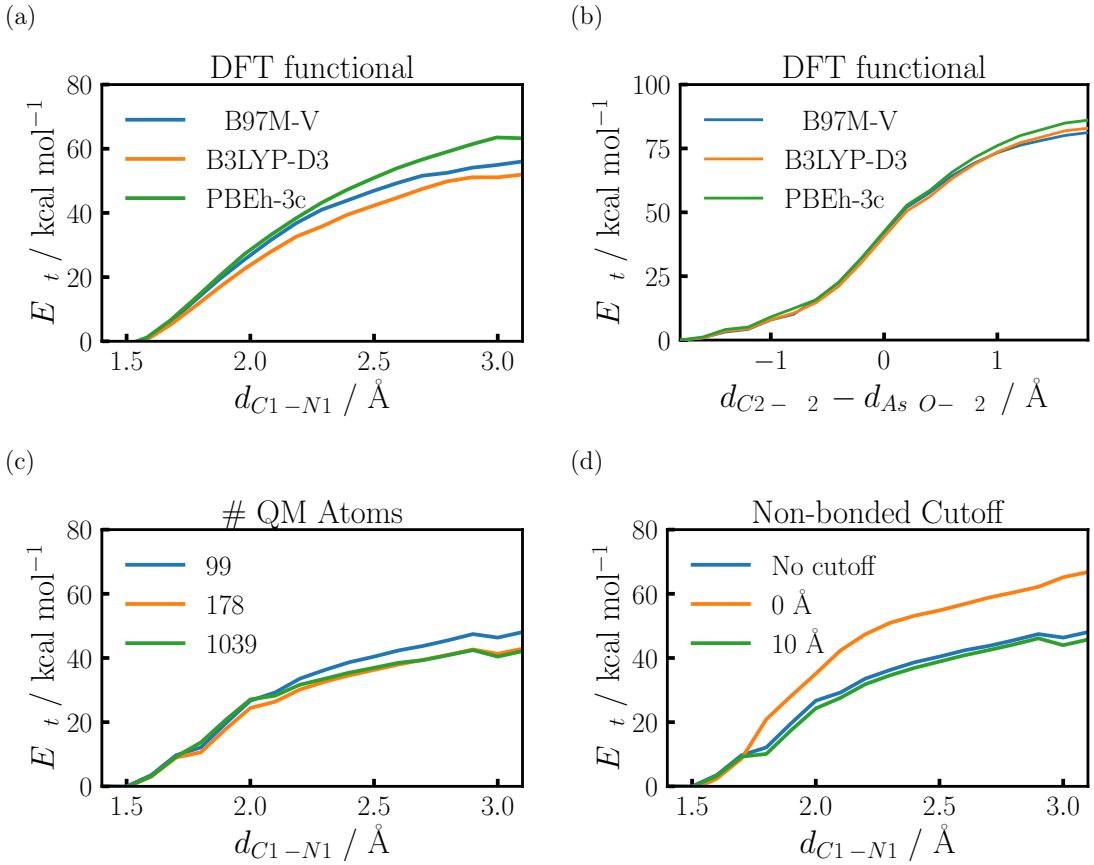


Figure S 2: Benchmark calculations on minimum energy paths (MEPs) of the C1'-N1 bond cleavage and proton transfer reaction of H2' to AspO. (a,b) For different DFT functionals on a single ribo-nucleotide implicitly solvated in water (COSMO [3]). For (b) also all charges residues in 5 Å proximity and two phosphates of the RNA backbone are added. Shown are MEPs computed using the highly accurate ωB97M-V functional [4, 5], the popular B3LYP-D3 functional [6, 7] (both with the triple-ζ basis set def2-TZVP [8]), and Grimme's cost-effective PBEh-3c functional [9]. Due to its much higher cost efficiency we will use PBEh-3c for *ab-initio* molecular dynamics simulations, which tends to overestimate both reaction barriers. By this choice resulting reaction (free) energies might safely be regarded as upper bound to the true energy. (c) For the number of QM atoms used in QM/MM simulations. The smallest region with 99 atoms contains the ribo-nucleotide in the active site together with all charged residues in 5 Å proximity, while the largest QM region with 1039 contains all atoms within 5 Å. The medium sized QM region with 178 atoms is designed to contain all important interactions in the active site and reproduce the energy of the largest QM region. No cutoff is used for electrostatic interactions with the MM system. (d) For the influence of the cutoff of electrostatic interactions with the MM subsystem. With a cutoff of 0 Å the electrostatic interaction of the QM region with the environment is switched off, while using no cutoff corresponds to inclusion of electrostatic interaction with all MM atoms. Using a cutoff of 10 Å preserves high accuracy while offering about 4-fold speedup of the calculation with reference to no cutoff. All QM/MM calculations are performed on one snapshot of the full system taken from the last 100 ns of an MM-MD using the PBEh-3c functional. For (d) the medium-sized QM region with 178 atoms is applied.

5 Details on Calculations of PUS Reaction Mechanisms

Reaction free energy profiles are calculated using QM/MM-MD simulations. A time step of 0.5 fs is used, and the temperature is controlled at 300 K using Langevin dynamics with a friction constant of 0.001 fs^{-1}. The exploration of reaction coordinates is accelerated with the Well-Tempered Metadynamics extended-system Adaptive Biasing Force (WTM-eABF) hybrid algorithm [10]. For the WTM potential, Gaussian's with an initial height of 1 kJ/mol and variance of 0.03 were deposited every 10 fs. The effective temperature was set to 4000 K. The fictitious particle had a mass of 40 a.u. and was coupled to the CV with a thermal coupling width of 0.01. The WTM and ABF forces were collected on a grid with bin width 0.01. The ABF force was scaled up with a linear ramp and fully applied in bins with more than 200 samples. Our own implementation of adaptive PCVs [11, 12] is used to describe transitions associated with different reaction mechanisms. The distance from the path is confined with a harmonic potential with a force constant of 100 kJ/mol$Å^2$ to suppress potential side reactions. For the glycal mechanisms still side reactions are observed due to the high mobility of H2', which are filtered out in post-processing. Observed side reactions include protonation of the uridine N1 with H2' and exchange of the hydroxy O2'H (which is hydrogen bonded to the catalytic Asp85) with H2'. Additionally, frames with a confinement force to the path larger than 60 kcal/mol are removed to only consider conformations that are reasonably close to the path. This leads to a slightly reduced total initial simulation time of around 530 ps compared to the rebound mechanism with over 600 ps.

Table S 2: Bond distance thresholds for filtering of frames in post-processing of simulations of the glycal mechanism.

	Threshold
$d_{H2'-N1}$	< 2.00
$d_{O2'-O2H'}$	> 1.25

To get optimal control over path convergence 10 walkers are simulated for 5-20 ps each at a time and the update is calculated manually from the full data. PCVs are defined in the space of a small set of bond distances, that are suitable to describe the slow degrees of freedom of the given process (see Table S3). Here, the distances of H2' to Asp85 oxygens is defined as

$$d_{\mathrm{AspO-H2'}} = \min\left[d_{\mathrm{Asp_{85}O1-H2'}}, d_{\mathrm{Asp_{85}O2-H2'}}\right] \tag{S4}$$

Note, that $d_{\mathrm{AspO-H2'}}$ is smooth even if Asp85 rotates during the simulation and the proton acceptor oxygen changes.

Table S3: Bond distances that build the CV space for calculations of PCVs.

	Rebound	Glycal
C1'-N1	√	√
C1'-C5	√	√
C1'-N3	√	√
C2'-H2'	×	√
AspO-H2'	×	√

Final free energy profiles are obtained from WTM-eABF biased trajectories using the MBAR estimator [13], as proposed in Ref. [14]. The MBAR equations are solved self-consistently. The starting guess for reduced free energies $\beta\hat{f}_i$ is zero and $\hat{f}_1$ is set to zero after every cycle. Convergence is reached when the largest change of $\beta\hat{f}_i$ compared to the last cycle drops under 10^{-6}. Reaction and

activation free energies are estimated from the PMF as proposed in Refs. [15, 16]. Confidence intervals are calculated from the standard deviation between independent simulations. The full source code is available on Github under `https://github.com/ochsenfeld-lab/adaptive_sampling`.

References

[1] Maragliano, L., Fischer, A., Vanden-Eijnden, E., Ciccotti, G., *J. Chem. Phys.* **2006**, *125*, 024106.

[2] Veerareddygari, G. R., Singh, S. K., Mueller, E. G., *J. Am. Chem. Soc.* **2016**, *138*, 7852–7855.

[3] Klamt, A., *Wiley Interdiscip. Rev.-Comput. Mol. Sci.* **2011**, *1*, 699–709.

[4] Mardirossian, N., Head-Gordon, M., *J. Chem. Phys.* **2016**, *144*, 214110.

[5] Mardirossian, N., Head-Gordon, M., *Mol. Phys.* **2017**, *115*, 2315–2372.

[6] Stephens, P. J., Devlin, F. J., Chabalowski, C. F., Frisch, M. J., *J. Phys. Chem.* **1994**, *98*, 11623–11627.

[7] Grimme, S., Antony, J., Ehrlich, S., Krieg, H., *J. Chem. Phys.* **2010**, *132*, 154104.

[8] Schäfer, A., Horn, H., Ahlrichs, R., *J. Chem. Phys.* **1992**, *97*, 2571–2577.

[9] Grimme, S., Brandenburg, J. G., Bannwarth, C., Hansen, A., *J. Chem. Phys.* **2015**, *143*, 054107.

[10] Fu, H., Shao, X., Cai, W., Chipot, C., *Acc. Chem. Res.* **2019**, *52*, 3254–3264.

[11] Leines, G. D., Ensing, B., *Phys. Ref. Lett.* **2012**, *109*, 020601.

[12] Pérez de Alba Ortız, A., Tiwari, A., Puthenkalathil, R., Ensing, B., *J. Chem. Phys.* **2018**, *149*, 072320.

[13] Shirts, M. R., Chodera, J. D., *J. Chem. Phys.* **2008**, *129*, 124105.

[14] Hulm, A., Dietschreit, J. C., Ochsenfeld, C., *J. Chem. Phys.* **2022**, *157*, 024110.

[15] Dietschreit, J. C., Diestler, D. J., Ochsenfeld, C., *J. Chem. Phys.* **2022**, *156*, 114105.

[16] Dietschreit, J. C., Diestler, D. J., Hulm, A., Ochsenfeld, C., Gómez-Bombarelli, R., *J. Chem. Phys.* **2022**, *157*, 084113.

3.3 Publication III: On the Molecular Mechanism of ATP Hydrolysis Catalyzed by p97: a QM/MM Study

Abstract: A computational study of p97/VCP ATPase using hybrid quantum mechanics/molecular mechanics (QM/MM) simulations is presented that explores the conformational landscape of the active site and hydrolysis-competent states of the crystallographic water molecules. Our investigation focuses on the reaction mechanism, particularly the events of the rate-determining first reaction step, which we study using extensive sampling with the path well-tempered metadynamics extended-system adaptive biasing force (WTM-eABF) enhanced sampling method. We identify the highly conserved glutamate (Glu305) from the Walker B motif as a catalytic base that activates the lytic water molecule for nucleophilic attack on the γ-phosphate in the first reaction step, while the final product is formed in a second step that involves proton transfer and rearrangements in the Mg^{2+} coordination sphere. We show that phosphate bond formation and breakage occur concertedly in the first reaction step. The findings gained through versatile QM/MM approaches are validated against recent cryo-EM and NMR data for the post-hydrolysis protein state, elucidating the role of amino acids from conserved motifs across the AAA+ protein family. To the best of our knowledge, this is the first in silico exploration of ATP hydrolysis in p97/VCP or any other AAA+ protein.

Reprinted with permission from

J. K. Szántó; A. Hulm; C. Ochsenfeld. "Molecular Mechanism of ATP Hydrolysis Catalyzed by p97: a QM/MM Study" *J. Chem. Theory Comput.* **2025**, 21, 19, 9459–9469.
URL: `https://doi.org/10.1021/acs.jctc.5c00928`.

JCTC Journal of Chemical Theory and Computation

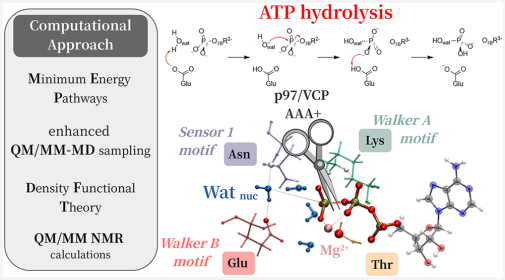
pubs.acs.org/JCTC
Article

Molecular Mechanism of ATP Hydrolysis Catalyzed by p97: A QM/MM Study

Judit Katalin Szántó, Andreas Hulm, and Christian Ochsenfeld*

Cite This: J. Chem. Theory Comput. 2025, 21, 9459−9469
Read Online

ABSTRACT: A computational study of p97/VCP ATPase using hybrid quantum mechanics/molecular mechanics (QM/MM) simulations is presented that explores the conformational landscape of the active site and hydrolysis-competent states of the crystallographic water molecules. Our investigation focuses on the reaction mechanism, particularly the events of the rate-determining first reaction step, which we study using extensive sampling with the path well-tempered metadynamics extended-system adaptive biasing force (WTM-eABF) enhanced sampling method. We identify the highly conserved glutamate (Glu305) from the Walker B motif as a catalytic base that activates the lytic water molecule for nucleophilic attack on the γ-phosphate in the first reaction step, while the final product is formed in a second step that involves proton transfer and rearrangements in the Mg^{2+} coordination sphere. We show that phosphate bond formation and breakage occur concertedly in the first reaction step. The findings gained through versatile QM/MM approaches are validated against recent cryo-EM and NMR data for the post-hydrolysis protein state, elucidating the role of amino acids from conserved motifs across the AAA+ protein family. To the best of our knowledge, this is the first *in silico* exploration of ATP hydrolysis in p97/VCP or any other AAA+ protein.



1. INTRODUCTION

The conversion of chemical energy in the form of ATP to exert mechanical force is one of the fundamental riddles of biochemistry. An example is p97, also known as valosin-containing protein (VCP), a hexameric motor complex and member of the AAA+ (ATPases associated with diverse cellular activities) protein superfamily that binds, hydrolyzes, and releases ATP to regulate various cellular pathways.[1] Extensive research has been carried out on the conformational changes of the global p97 protein structure during the ATPase cycle,[2−8] but no previous study has elucidated the molecular mechanism of ATP hydrolysis catalyzed by p97. ATP hydrolysis in solution can proceed through multiple pathways,[9,10] and the conformational landscape becomes even more complex at the active site of a protein. Despite their functional diversity, nucleoside triphosphate (NTP) hydrolyzing enzymes (NTPase proteins) often share a common nucleotide binding fold. For instance, P-loop NTPases use a highly conserved loop to bind and efficiently hydrolyze nucleotides.[11] Here, computer simulations that capture protein structure and dynamics using a hybrid QM/MM framework can provide full atomic details at high temporal resolution and have, in several studies, elucidated the catalytic mechanism of P-loop NTPases to which p97 belongs. For example, recent QM/MM studies on Ras-GTPases—whose malfunction drives many cancers—have revealed how oncogenic mutations alter the catalytic activity[12] and uncovered how the structural complexity of the active site, involving different side-chain tautomers, facilitates phosphate hydrolysis.[13]

Given the complexity of enzyme-catalyzed reactions, a comprehensive understanding of factors such as the roles of amino acids, water molecules, and metal ions at the active site is essential and guides the computational exploration of the reaction mechanism. Therefore, we build on insights gained from the existing literature on ATPases and GTPases,[14] as well as kinetic and mutational studies on the AAA+ protein family and other P-loop NTPases to which p97 belongs. For our theoretical study, amino acids that are close to the substrate and whose mutations affect nucleotide binding or ATP hydrolysis rates are crucially important. These include glutamate (Glu305) from the Walker B motif, a highly conserved residue in the nucleotide binding pocket of several AAA+ proteins.[1,15−17] Upon mutation to glutamine, ATP binding is preserved, but hydrolysis is hindered[15,18−20] in

Received: June 6, 2025
Revised: August 15, 2025
Accepted: September 5, 2025
Published: September 19, 2025

https://doi.org/10.1021/acs.jctc.5c00928
J. Chem. Theory Comput. 2025, 21, 9459−9469

Journal of Chemical Theory and Computation pubs.acs.org/JCTC Article

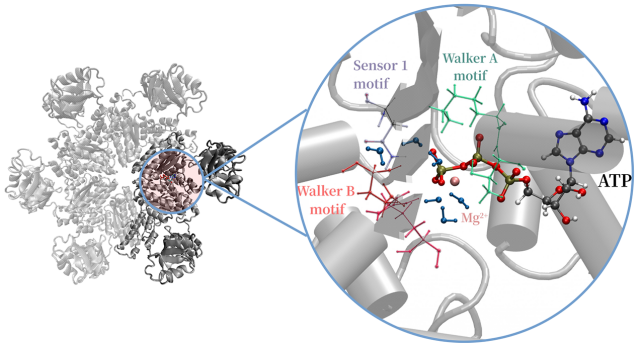


Figure 1. Left: Structure of the hexameric p97 featuring the N-terminal domain and the D1 nucleotide binding domain (PDB 4KO8[31]). Two adjacent subunits are highlighted in dark gray, which are selected for our computations. Right: schematic representation of the binding site with ATP, the Mg^{2+} ion, and crystallographic water molecules. Walker A, B, and the Sensor 1 motif, as highly conserved regions across AAA+ proteins, are highlighted with colors.

various members of this protein family. More precisely, experimental findings in p97 show that this mutation leads to a 20-fold decrease in enzymatic activity,[20] suggesting that it must play a crucial role in the reaction mechanism of ATP hydrolysis.

Another important feature of the AAA+ family is the Sensor 1 motif (typically asparagine, serine, threonine, or aspartate), which was hypothesized to help orient the water molecule for the nucleophilic attack.[15,19,21−24] Experimental evidence from kinetic studies on AAA+ Sensor 1 mutants revealed decreased catalytic activity after mutating the Sensor 1 unit,[25,26] the typical mutation being asparagine to alanine.[1] Other amino acids of high relevance are the conserved arginine residues,[24,27] which, as positively charged residues, are thought to stabilize the transition state[28] and help in the intersubunit communication.[3] R359 and R362 residues are also called *trans*-acting arginine fingers, as they are located at the subunit interface, extending from one subunit into the active site of the neighboring one. Furthermore, a recent high-resolution cryo-EM study[8] showed that threonine (Thr252) from the P-loop of p97 plays a key role in coordinating the Mg^{2+} ion, which neutralizes negative charges of the nucleotide's phosphate groups. The Mg^{2+} ion is also an important protagonist in ATP catalysis, as it has previously been shown that the presence of metal ions at the active site can alter the reaction mechanism of phosphate ester hydrolysis.[10,29,30] The full hexameric p97 complex and key residues of the active site are shown in Figure 1.

In addition to amino acids and metal ions, buried water molecules contribute to the mechanistic complexity of ATP hydrolysis. Previous QM/MM studies of other P-loop NTPases have shown that multiple water molecules can participate in the reaction, posing challenges for the computational exploration of reaction pathways. In ABC transporters, a single water molecule was found to be sufficient for ATP hydrolysis.[32] In contrast, myosin, kinesin, and F1-ATPase require multiple water molecules. In myosin, two catalytic water molecules were found: an attacking and a helping water, which are positioned by a dense hydrogen bonding network.[33] ATP hydrolysis in F1-ATPase occurs with the help of three water molecules, which are directly involved in the reaction process.[34] This raises the important question of how many water molecules are involved in the catalytic mechanism of p97.

Our interest lies not only in gaining mechanistic insight but also in understanding how ^{31}P chemical shifts evolve during hydrolysis. This is motivated by our previous study,[35] where we observe a drastic change in the chemical shifts of the P_β nucleus for pre- and post-hydrolysis protein states, which can only fully be explained by direct investigation of the mechanism of p97.

In this work, we present a detailed investigation of the p97 reaction mechanism, using extended QM/MM calculations. For this purpose, we followed a three-step computational workflow. We start by exploring possible reaction mechanisms in the active site by testing reactions of nearby water molecules. Second, the reaction paths that lead to stable intermediates are optimized to obtain minimum energy pathways. Lastly, after a thorough benchmark of the reliability of the QM/MM setup, enhanced sampling MD simulations are performed to obtain accurate reaction and activation free energies of the rate-limiting step. We discuss the structural rearrangements at the active site during product formation and finally predict NMR chemical shifts along the reaction pathway, always thoroughly relating our findings to the experimental data. In this way, we aim to provide a complete picture of ATP hydrolysis in p97 and similar members of the AAA+ protein family.

2. METHODS

For the QM/MM study on the reaction mechanism, we build on insights gained from our recent computational study[35] on p97 ATPase, as well as experimental NMR,[20] cryo-EM, and MM-MD studies[8] on this protein. The crystal structure (PDB 4KO8[31]) of p97 originally contains the hydrolysis-resistant ATPγS at the active site, which was transformed into ATP, followed by QM/MM structure optimizations, and served as educt structure and starting point for exploring the reactivity. The conversion of ATPγS to ATP, system preparation, protonation state assignment of titratable groups, and equilibration of the ATP-bound p97 protein were performed by Shein et al.,[8] who provided us with the resulting equilibrated educt structure. Compared to the hydrolysis-resistant substrate analogue (ATPγS), the presence of the true

https://doi.org/10.1021/acs.jctc.5c00928
J. Chem. Theory Comput. 2025, 21, 9459−9469

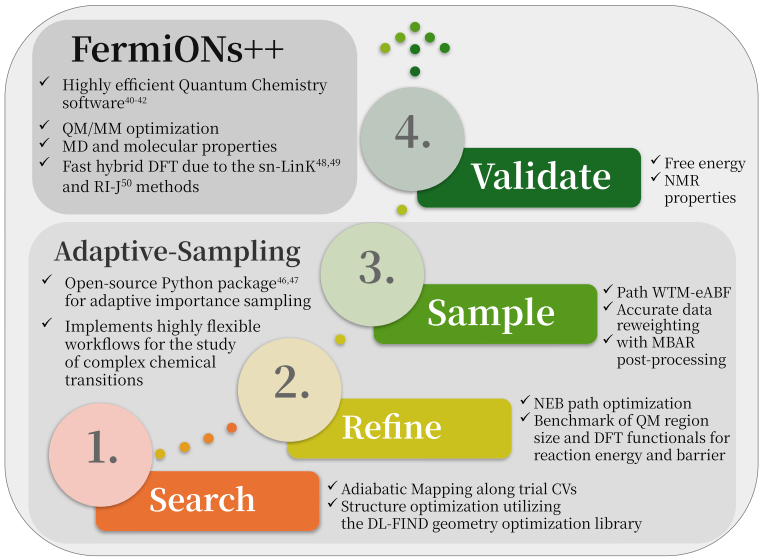


Figure 2. Workflow overview, which builds on the FermiONs++[40−42] quantum chemistry software and the Adaptive-Sampling Python package.[46,47]

substrate (ATP) at the active site prompts the reorientation of key protein residues such as the *trans*-acting R359 and R362 arginine fingers and Glu305 (see Figure S2 in the Supporting Information (SI)). The functional p97 machine assembles from six identical subunits, also called protomers, from which we chose two neighboring protein subunits as subsystem with ca. 30000 atoms for our simulations (see Figure 1). Events such as substrate binding, inorganic phosphate release, and ADP unbinding involve complex conformational changes in the global structure of p97,[20] and for the *in silico* study of these steps, intersubunit communication must be explicitly modeled, requiring analysis of the entire hexamer. However, our focus is on the chemical mechanism of ATP hydrolysis; therefore, only a subsystem comprising two neighboring subunits was selected for our study.

All QM/MM calculations were performed using the PBEh-3c[36] DFT functional on 120−160 QM atoms, the rest of the system being described by the MM part using the Amber ff14Sb.[37] Before choosing the DFT functional used for the QM/MM calculations, we performed a DFT functional benchmark study on the minimum energy pathways. Here, we observed a good agreement within 1−2 kcal/mol between the efficient PBEh-3c/def2-mSVP method and the DLPNO−CCSD(T)/def2-QZVP[38,39] coupled-cluster approximation. Therefore, we decided to use the PBEh-3c hybrid DFT functional, which was developed for efficient geometry optimizations and reaction energy evaluations in large molecular systems using a smaller DZ basis set to balance accuracy and cost. The influence of different DFT functionals and basis sets is presented in Section 4 of the SI.

For the QM/MM calculations, the FermiONs++[40−42] quantum chemistry program package in combination with the LibXC[43] library of exchange-correlation functionals, the OpenMM[44,45] library, and the Adaptive-Sampling Python package[46,47] were employed.

FermiONs++ enables hybrid DFT applications on extended biomolecular systems, due to its efficient implementations of the sn-LinK[48,49] and RI-J[50] methods, and also supports the linear-scaling computation of NMR chemical shieldings.[51] QM/MM interactions were treated with an additive scheme using electrostatic embedding,[52] and QM/MM NMR calculations were conducted at the B97−2/pcSseg-2[53,54] level of theory. Complete computational details on the used methods and benchmark studies are provided in the Supporting Information, while in the following, a brief summary of the employed workflow is given.

Our methodology is based on the steps, as summarized in the overview above (Figure 2), followed by the final relation of the results to experimental studies. We begin the initial exploration using adiabatic mapping (AM) to test the nucleophilic attack of crystallographic water molecules close to the P_γ and P_β atoms of the ATP molecule. AM pathways were obtained from a sequence of energy minimizations along selected reaction coordinates. It is important to note that reaction coordinates involve only the forming of the O_{Wat}-P_γ and/or dissociating P_γ-O_{3B} distances, while water protons migrate to the best suited acceptor in an unbiased fashion. In our workflow, AM was used solely to identify the next local minimum on the potential energy surface (PES). Once a minimum was found, the Nudged Elastic Band (NEB) method[55] was applied to determine the minimum energy path (MEP) between the optimized metastable states. Next, we carried out a thorough benchmark study to determine the appropriate QM region size, as shown in Sections S5−S6 of the SI.

To obtain accurate activation and reaction free energies, we explore the potential of mean force (PMF) (i.e., free energy surface) of the rate-determining step by sampling the underlying PES. This is the most computationally demanding step of our approach, as the time step is 0.5 fs, and the total sampling time exceeds 2 ns. This extensive sampling was

https://doi.org/10.1021/acs.jctc.5c00928
J. Chem. Theory Comput. 2025, 21, 9459−9469

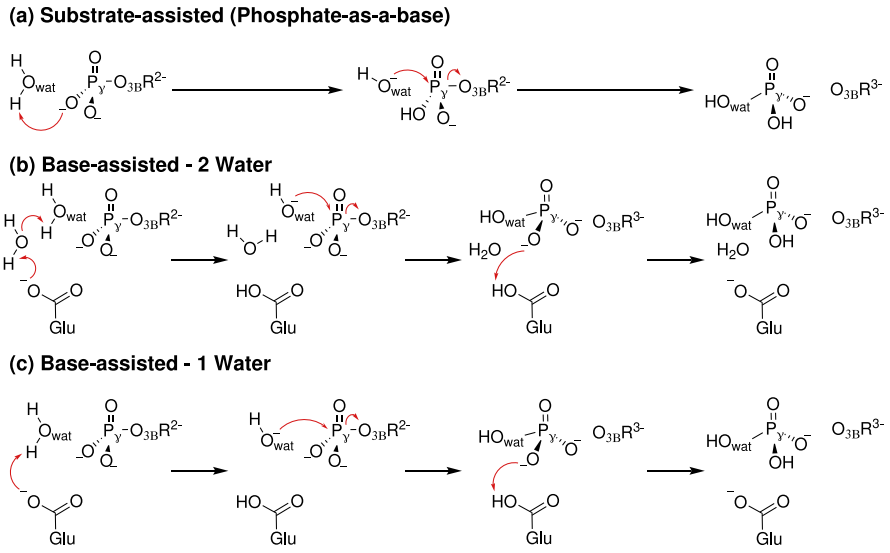


Figure 3. Observed reaction mechanisms using adiabatic mapping.

achieved through the high efficiency of our in-house FermiONs++[40−42] quantum chemistry code, enabling the computation of statistically robust free energy profiles of the rate-limiting reaction steps.

The choice of the collective variable (CV), which provides a measure of the reaction progress, is crucial for successful importance sampling simulations, as a poor selection of the CV can drastically influence the obtained activation free energy.[56] Therefore, we used the NEB optimized reaction pathway as a path collective variable (PCV)[57] for sampling with the well-tempered metadynamics extended-system adaptive biasing force method (WTM-eABF).[58−60] We choose WTM-eABF over static sampling methods like Umbrella Sampling (US), because it facilitates the free diffusion of the system along CVs, resulting in the broad and reliable sampling of transition pathways.[61] The CV space for the PCV was defined *a priori* by selecting breaking and forming bond distances involved in the corresponding step, ensuring that the CVs clearly differentiate key states and remain minimal in number. For further details on the parameters used in the QM/MM-MD sampling, see Section S7 of the SI. Finally, we validate our results by comparison with recent cryo-EM[8] and experimental NMR data,[20] which capture the ADP.P_i post-hydrolysis protein state prior to the release of inorganic phosphate (P_i).

3. RESULTS AND DISCUSSION

The X-ray structure with PDB 4KO8[31] reveals buried water molecules at the active site (see Figure S3 of the SI), which form a highly conserved and integral part of the structure of the p97 protein. In the computational modeling of the reaction mechanism, the first challenge is identifying the catalytic water molecule responsible for cleaving the phosphate group as well as the role of key catalytic amino acids such as conserved Glu305 and Sensor 1 Asn348 in the process, as discussed in Sections 3.1−3.3, respectively. After the initial nucleophilic attack, the final product is built in a second reaction step that involves proton transfer and rearrangement of the Mg^{2+} coordination sphere (Section 3.4). Finally, both reaction pathways are compared to solid-state NMR measurements (Section 3.5).

3.1. Walker B Glu305 Acts as a Catalytic Base. We started by testing the influence of differently oriented crystallographic water molecules on the activation barrier of the first step in ATP hydrolysis. Close to the reaction center, we observe the same number of water molecules as in the crystal structure, occupying well-defined locations throughout the long-timescale MD simulation of the educt structure reported by Shein et al.,[8] as shown in Figures S10 and S11 of the SI. We identified three water molecules as likely candidates for the nucleophilic attack, based on their proximity to the P_γ atom or to a suitable proton acceptor. In the first step of phosphate hydrolysis, the H^+ of the lytic water molecule is transferred to a nearby proton acceptor, OH^- attacks the P_γ, and the P_γ-O_{3B} bond breaks. Our exploration using adiabatic mappings (see Section S2 of the SI) shows that the proton can be accepted by either a protein group acting as a base (base-assisted) or by the phosphate itself (substrate-assisted). Additionally, one or two water molecules can be involved. An overview of the obtained reaction mechanisms is given in Figure 3.

From the MEPs presented in Figure 4, which are obtained by NEB optimization of AM reaction pathways, we conclude that the base-assisted (Glu305) mechanisms with the participation of a single water molecule are more favorable than the two water or the substrate-assisted (phosphate-as-a-base) mechanisms, whose activation energy barrier is more than 20 kcal/mol higher. For the Glu305-assisted, single water path, we identify two distinct reaction channels (channel A and B), as further discussed below. Furthermore, we can conclude that a single water molecule can hydrolyze ATP in p97. However, for enzymes where the catalytic base is positioned farther away from the P_γ or the nucleophilic water molecule, the proton transfer may require a longer pathway, potentially

https://doi.org/10.1021/acs.jctc.5c00928
J. Chem. Theory Comput. 2025, 21, 9459−9469

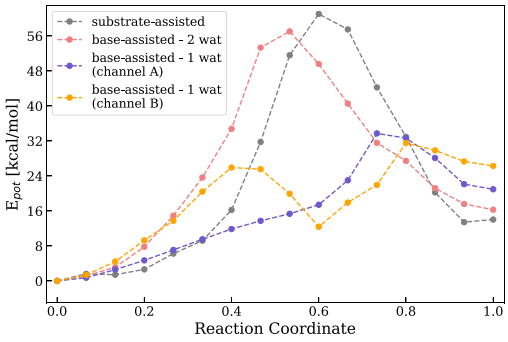


Figure 4. First step of ATP hydrolysis. Minimum energy profiles (MEPs) were obtained for the substrate- and base-assisted reaction mechanisms.

involving one or more bridging water molecules.[33,62,63] That the substrate-assisted mechanism is not feasible is supported by previous QM/MM studies[32,64] on P-loop NTPases, which likewise identify the base-assisted pathway as the more favorable route—reported to be 10 kcal/mol[32] or 26 kcal/mol lower in energy.[64]

The observation that glutamate plays a direct role in ATP hydrolysis aligns well with experimental findings in p97, which show that mutating this conserved glutamate to glutamine alters the catalytic rate constant and abolishes ATP hydrolysis,[31,65] more specifically resulting in a 20-fold decrease in enzymatic activity.[20] Additional support for the proposed glutamate-assisted mechanism comes from experimental and computational studies on nucleotide triphosphate (NTP) hydrolysis in other NTPases,[34,63,64,66,67] which, like AAA+ proteins, share the Walker B motif and a highly conserved glutamate in this region.

3.2. Orientation of the Catalytic Water by Sensor 1. For the glutamate-assisted mechanism, we identify two reaction channels. Channel A, where a water molecule attacks that is stabilized and oriented by hydrogen bonds to Sensor 1 Asn348, and channel B, which involves a different water molecule. In Figure 5, both water molecules are shown in blue and orange, respectively, together with the key interatomic

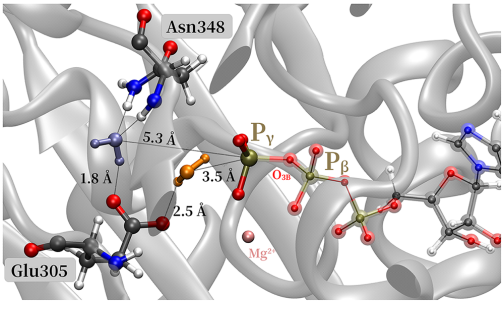


Figure 5. Binding pocket: two water molecules close to Glu305 and the P_γ atom. The water molecule marked by violet blue is oriented and stabilized by hydrogen bonds with the Sensor 1 Asn348 (channel A). The water molecule marked by orange is the closest to the P_γ atom (channel B).

distances to P_γ and Glu305. The channel A water molecule (blue) is further away from P_γ than the channel B water, but already well positioned for nucleophilic attack, resulting in a smooth reaction energy profile (blue curve in Figure 4). Additionally, the channel A water molecule is stabilized by a strong H-bond formed with the amide of the peptide bond between Asn348 and Thr347 and a weaker H-bond, as well (see also Figure S11). This buried water molecule is also part of the crystal structure, located within 3 Å from the H-donor N of the asparagine and 5.3 Å away from the P_γ atom (see Figures S3, S10, and S11).

The channel B water molecule (orange) is closer to P_γ, but it is not well positioned for hydrogen transfer to Glu305 (see also Figures S9−S11 in the SI). Hence, the corresponding MEP of Figure 4 (orange) shows two reaction barriers, where the first corresponds to the reorientation of the water molecule to a hydrolysis-competent state, where the attacking angle is more optimal. Active site configurations for the tested water positions and the resulting adiabatic mapping pathways are shown in Section S2 of the SI.

Figure 6 shows the minimum energy pathways as obtained from NEB calculations together with key interatomic distances (O_{Wat}-P_γ in red, O_{3B}-P_γ in green, and proton distance to Glu305 in blue). The first nine images of the channel B NEB path capture the reorientation of the water molecule to a position where it is closer to the proton acceptor, while the O_{Wat}-P_γ distance does not change. At the same time, the H^+ approaches Glu305 to 1.8 Å. This distance of the O_{Glu305}-H_{Wat} does not change for channel A as the preoriented water molecule only needs to get closer to the P_γ. Another important structural characteristic that distinguishes the two channels is the attack angle (Figure S9) that needs to reach a nearly collinear state before hydrolysis occurs. After reorientation of the channel B water, for both channels, the O_{3B}-P_γ bond cleavage (green) and O_{Wat}-P_γ bond formation (red) occur concertedly in an S_N2-like reaction mechanism and with a similar reaction energy barrier.

Therefore, we conclude that the reaction requires a water molecule that is well positioned for the nucleophilic attack as well as hydrogen-bonded to the proton-accepting Glu305. The Sensor 1 Asn348 provides perfectly preoriented water molecules, whereas the nucleophilic attack of other waters involves an additional reorientation step, reaching an activated intermediate configuration. In line with our observation of the Asn348 residue's role in p97, a recent QM/MM study on helicase-catalyzed ATP hydrolysis[64] similarly reported that a hydrogen bond involving the backbone of a glycine residue at the active site helps to orient the nucleophilic water molecule for the attack.

3.3. QM/MM-MD Conformational Sampling Lowers the Activation Barrier. In the next step, the optimized MEP serves as PCV for free energy simulations. Because of the high cost of these simulations, we select only the most promising MEP, the one corresponding to the attack of the Sensor 1-oriented water molecule in a base-assisted (Glu305) mechanism. MD simulations are initiated from every image of the optimized NEB path and a reaction free energy profile is computed from QM/MM-MD simulations using the path Well-Tempered Metadynamics extended-system Adaptive Biasing Force (WTM-eABF) algorithm as implemented in our Adaptive-Sampling package.[60]

In Figure 8, the resulting PMF (right), as well as trajectories (left) and histograms (middle) of the PCV for the two

https://doi.org/10.1021/acs.jctc.5c00928
J. Chem. Theory Comput. 2025, 21, 9459−9469

Journal of Chemical Theory and Computation pubs.acs.org/JCTC Article

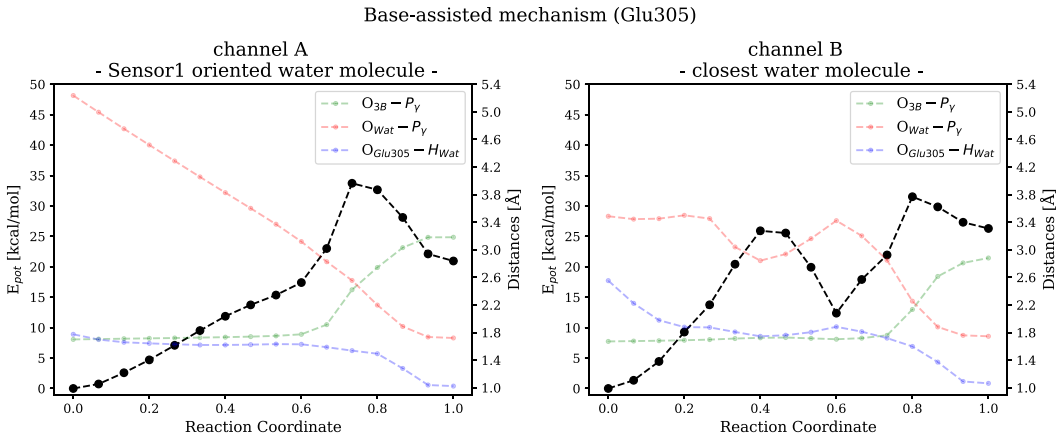


Figure 6. MEPs for channel A and channel B reactions are shown in black. Colored lines denote key interatomic distances for the base-assisted mechanism, with the corresponding axis given on the right.

simulations that start from both minima, are shown. The trajectories of 14 additional simulations are shown in Section S10 of the SI, reaching a total sampling time of 2 ns. A PMF from the data of all combined simulations is shown in gray on the right side of Figure 8. The trajectories (left) show how the system is reversibly driven from one basin to another. The first transition in the forward direction occurs after 23−30 ps as the system has to overcome a high energy barrier in this direction, while the first transition in the backward direction takes place in less than 25 ps. The histograms of the PCV show a sufficiently uniform distribution over the sampled 200 ps, during which four full transitions occur between the educt and the intermediate structures. This suggests that the employed WTM-eABF algorithm effectively enhances sampling. It drives the system across the free energy landscape, allowing even exploration of the reactant, transition state, and product regions. For further details, see Figures S24−S25, which illustrate the evolution of key interatomic distances during sampling. Efficient enhancement of the QM/MM-MD sampling was essential for two reasons: the large size of the QM region (Figure 7) and the extended simulation time. The QM region consists of 164 atoms, selected around the reaction center defined by the attacking water molecule, the P_γ atom, and the catalytic base E305, as illustrated in Figure S16 of the Supporting Information.

The free energy barrier (25 kcal/mol) obtained from PMF is significantly lower than the static NEB result (35 kcal/mol). Furthermore, the intermediate is also strongly stabilized, as evidenced by the local minimum corresponding to the ADP + HPO_4^{2-} state at about 10 kcal/mol in contrast to the 20 kcal/mol located on the MEP (see Figure 6).

From the experimentally measured reaction rate constant for ATP hydrolysis in p97,[8,20] we estimate the activation free energy using the Eyrings equation $k = \frac{\kappa k_B T}{h} \exp\left[-\frac{\Delta G^\ddagger}{RT}\right]$, where κ is assumed to be 1, k_B is the Boltzmann constant, T is the temperature used in the experiment, h is Planck's constant, and $\Delta G^\ddagger$ is the activation free energy. According to this, a free energy barrier of at least 19.67 kcal/mol corresponds to the experimentally measured $k_{hydrolysis}$ = 20 min^{-1} (50 °C) rate

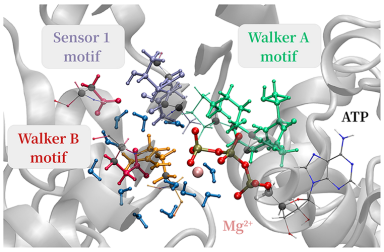


Figure 7. First step: the QM region consist of 164 atoms: the phosphate backbone of the ATP molecule, the 12 closest water molecules (blue), Glu305 and Asp307 from the Walker B motif (red), Ala346, Thr347, and Asn348 from the Sensor 1 motif (purple), Gly248, Thr249, Gly250, and Lys251 from the Walker A motif (green), as well as Arg359 from the adjacent protein subunit (orange). Gray spheres indicate carbon atoms at the QM/MM boundaries, where hydrogen link atoms were introduced; only nonpolar C−C bonds were cut to define the QM region.

constant,[20] which is in reasonable agreement with the obtained results. It is important to note that while a 2−5 kcal/mol error can be standard for hybrid DFT applications,[68] the slight overestimation of the barrier in our case can be attributed to the confinement of the path CV. When including the harmonic confinement potential into the MBAR analysis, the free energy barrier is reduced and closer to the experimentally derived value (see Figure S22 and the discussion in Section S7).

3.4. Forming the Product under the Rearrangement of the Mg^{2+} Coordination Shell. After the first step, we reach a high-energy intermediate at the active site, which corresponds to ADP, HPO_4^{2-}, and the protonated Glu305. The last step consists of a H^+ transfer from the Glu305 to the O atom of the P_γ atom, forming $H_2PO_4^-$, the inorganic phosphate (P_i). An optimized NEB path together with key interatomic distances is shown in Figure 9. In black, the MEP for this reaction is shown, which has a low activation barrier of 7 kcal/mol and is strongly exothermic by about 19 kcal/mol. The reaction is characterized by two structural rearrangements: first, the proton from the catalytic base is transferred to the

https://doi.org/10.1021/acs.jctc.5c00928
J. Chem. Theory Comput. 2025, 21, 9459−9469

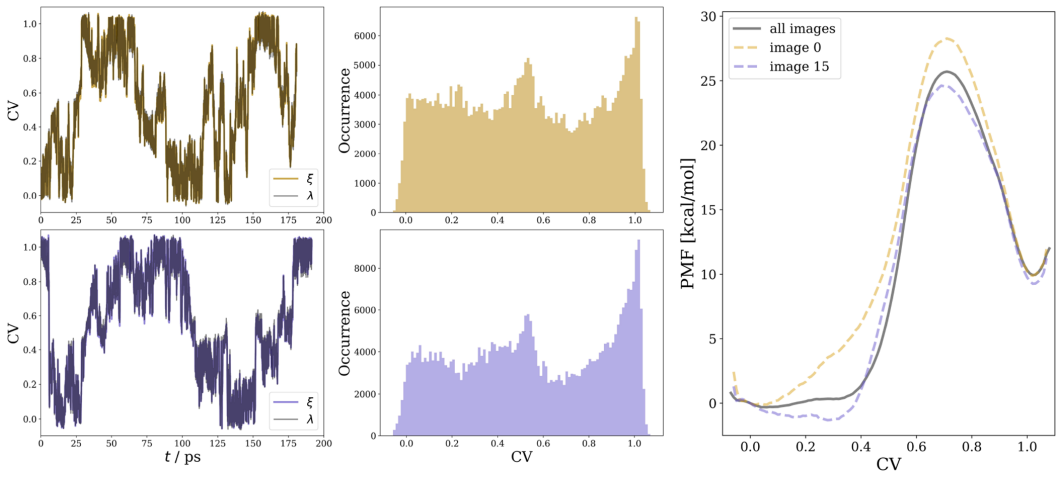


Figure 8. Left: Trajectories and histograms of the PCVs started from the educt (gold) and intermediate structure (purple). Right: PMF profiles were computed for QM/MM-MD trajectories initiated from the respective NEB images. The gray free energy profile was computed from the cumulative data of all 16 trajectories started from different NEB images. The PCV represents the hydrolysis reaction progress along the minimum free energy path, i.e., "0" corresponds to ATP and "1" to the ADP + HPO_4^{2-} intermediate.

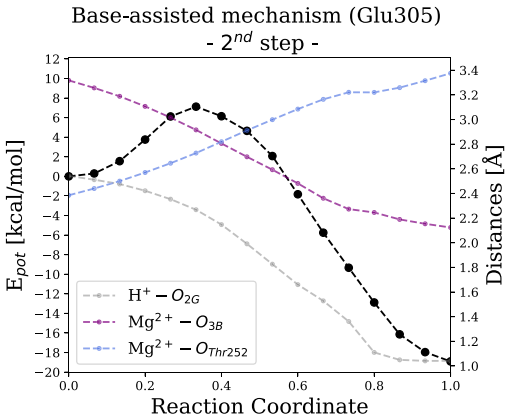


Figure 9. MEP of the second reaction step is shown in black (left axis) and key interatomic distances in gray, purple, and blue (right axis).

O_{2G} oxygen atom of P_γ (gray line of Figure 9), which coordinates Mg^{2+} (see Figure 10), likely due to steric proximity or the high electron density of this oxygen atom. This assumption is supported by a ^{19}F NMR study on GTP hydrolysis,[69] showing that the oxygen atom coordinated to Mg^{2+} has the highest electron density among the oxygens of P_γ and is the proton acceptor.[32] Second, the proton transfer step also requires a rearrangement of the divalent ion coordination shell, as shown by the purple and blue lines that denote the evolution of the Mg^{2+}-O_{3B} and Mg^{2+}-O_{Thr252} distances.

As shown in Figure 10, both in the intermediate and product states, Mg^{2+} is tightly coordinated by six ligands, forming an octahedral geometry. A key event in the second step is the reorganization of the Mg^{2+} coordination shell. In both the educt and intermediate states, Thr252—a highly conserved residue across many P-loop NTPases[11]—strongly coordinates to Mg^{2+}. However, upon $H_2PO_4^-$ formation, the threonine leaves the ion's coordination shell and O_{3B} becomes part of it. This observation aligns with the high-resolution cryo-EM structure,[8] which captures the ADP·P_i state of human ATPase p97 just before the release of the inorganic phosphate from the binding site. The cryo-EM density of this study also indicates that compared to the ATP-bound state in the product, the Mg^{2+} ion has already dissociated from the threonine. The rearrangement in the coordination shell leads to a 3-fold connection between Mg^{2+}, ADP, and P_i which remains stable for 10 ps in an unbiased QM/MM-MD simulation of the product state (see Section S8 of the SI), after which a water molecule enters the coordination shell.

As shown in Figure 11, in addition to the coordination shell, a strong H-bonding network stabilizes the product state. The Walker A residue Lys251—the immediate neighbor of Thr252—acts as a tridentate, contacts O_{3G} and O_{2B}, and thus bridges ADP and P_i. The inorganic phosphate is further stabilized by forming H-bonds with the H^+ donor Glu305 and Arg359, which are stronger than those of the educt state (see Section 8 of the SI).

We conclude that the minimum energy pathway of the proton transfer step leads to a product state, where the strongest electrostatic interactions between the substrate and protein residues occur with the Mg^{2+} ion and the Lys251 side chain, both of which were identified as key stabilizers of the post-hydrolysis ADP·P_i state.[8] The unbinding of the inorganic phosphate has a high kinetic barrier[20] and will be anticipated by a rearrangement in the Mg^{2+} coordination shell, where the inorganic phosphate detaches and the vacant places are occupied by neighboring water molecules. However, studying the detachment process of the inorganic phosphate is outside the scope of this paper.

3.5. Comparison to Solid-State NMR Measurements.

In our previous work,[35] we have discussed computed NMR shifts as ensemble properties predicted from microseconds-

https://doi.org/10.1021/acs.jctc.5c00928
J. Chem. Theory Comput. 2025, 21, 9459−9469

Journal of Chemical Theory and Computation pubs.acs.org/JCTC Article

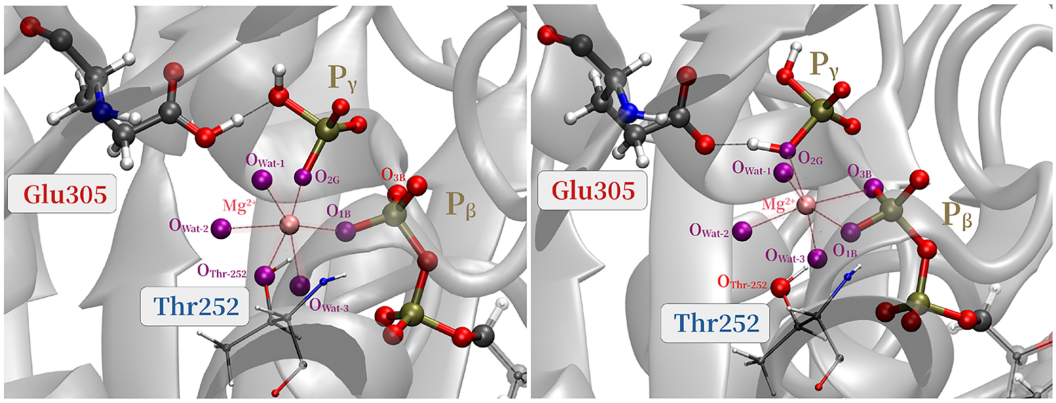


Figure 10. Second step: Mg^{2+} coordination shell in the intermediate (left) and product (right) states.

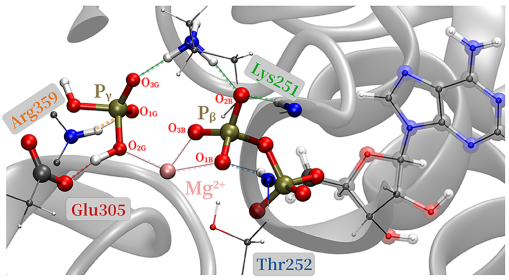


Figure 11. Key protein residues stabilize the product state.

long MM-MD trajectories of the educt and the product state without studying the reaction mechanism of ATP hydrolysis. The product state in this case featured ADP bound to p97 without the cleaved inorganic phosphate, which had already detached from the binding site. Here, we have observed that pre- and post-hydrolysis chemical shifts of the nucleoside differ the most for the P_β nucleus.[35] The chemical shift computed for this nucleus undergoes a drastic downfield shift of more than 12 ppm, showing a good agreement with experimental NMR measurements.[20] After the reaction mechanism of ATP hydrolysis is studied, NMR calculations are performed using a QM/MM DFT framework, allowing the computation of NMR chemical shifts along the reaction pathway as one transitions from the educt to the product structure. Compared to NMR measurements, which capture averaged chemical shifts over the acquisition time, this enables direct observation of the influence of cleavage of the phosphate bond in the first step, as well as proton transfer and Mg^{2+} coordination in the second step on chemical shifts.

The results of this analysis are shown in Figure 12 in combination with key interatomic distances. There are more factors that contribute to these changes in the chemical shifts; here, however, we restrict our discussion to the key events that are observed during the first and second reaction steps. For the

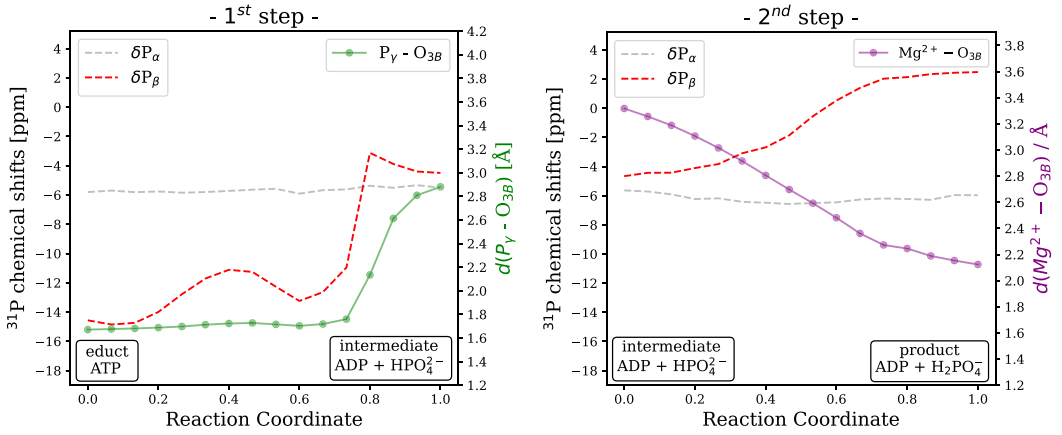


Figure 12. Monitoring ^{31}P NMR chemical shifts along the two-step reaction progress. The dashed lines show the chemical shifts of the P_α (gray) and the P_β nucleus (red). Continuous lines with points represent the measured interatomic distances during the 1st and 2nd reaction steps (P_γ-O_{3B} in green and Mg^{2+}-O_{3B} in purple).

https://doi.org/10.1021/acs.jctc.5c00928
J. Chem. Theory Comput. 2025, 21, 9459−9469

P_α nucleus, which is not directly involved in the reaction, chemical shifts barely change during the two reaction steps. In contrast, for the P_β nucleus, a 16 ppm downfield shift is observed upon hydrolysis, showing an excellent agreement with the experimentally observed difference between the shift measured before (ATP: −20 ppm) and after hydrolysis (ADP: −4 ppm) (see Figure 2 of the experimental NMR study[20]). The contribution of the second reaction step to the overall downfield shift of the P_β nucleus is significant (see the right panel of Figure 12). This is not surprising as we have already observed how the Mg^{2+} coordination sphere in the immediate proximity of the P_β nucleus changes as the reaction progresses toward the product state (see Figure 10). The O_{3B} atom enters the coordination shell, while Thr252 leaves it, as captured by the MEP.

Further, as the scissile P_γ-O_{3B} bond elongates (green line in Figure 12), the P_β shift immediately moves toward the downfield region, eventually reaching −5 ppm, such that the intermediate state already has a drastically changed P_β shift. The wild-type ADP-bound spectra show a downfield-shifted P_β signal, and based on the computed NMR chemical shifts along the reaction progress, we conclude that this change happens already upon hydrolysis, rather than only during the inorganic phosphate (P_i) release. Therefore, the experimental observation of a P_β shift in the upfield region near −15 ppm in the post-hydrolysis ADP.P_i state[20] remains elusive, likely due to the use of the E305Q mutation needed for the ADP.P_i NMR measurements, which could lead to a mixture of ATP-bound and hydrolyzed species.

4. CONCLUSIONS

In this work, we applied a hybrid QM/MM approach that combines minimum energy pathway optimizations and enhanced sampling methods to provide mechanistic insights into the protein-mediated ATP hydrolysis catalyzed by p97. Our findings clarify the role of various amino acids at the binding site:

- Glu305 from the Walker B motif, a highly conserved feature in many AAA+ proteins, catalyzes the reaction by accepting a proton from the catalytic water, which is later transferred back to the inorganic phosphate to form the final product.
- The Asn348 residue of the Sensor 1 motif, also present in all AAA+ proteins, orients the attacking water molecule. A crystallographic water molecule is positioned near a strong proton acceptor (Glu305), stabilized by hydrogen bonds, and aligned almost collinearly with the P−O bond that is about to be cleaved.
- The coordination shell of the Mg^{2+} ion stabilizes the transition state and product together with positively charged protein residues such as Lys251 of the Walker A motif and Arg359 of the Walker B motif. To form the product, the Walker A Thr252 leaves the Mg^{2+} coordination shell in favor of the O_{3B} of P_β.

Furthermore, we show that a single water molecule can hydrolyze ATP and that phosphate bond breaking and bond formation occur concertedly in a single reaction step. For studying the rate-limiting step of hydrolysis, we apply extensive sampling and obtain an activation free energy barrier that is in reasonable agreement with the experimental catalytic turnover rates. Starting from the pre-hydrolysis crystal structure, our computational exploration of the reaction mechanism finally leads to a conformation that is consistent with the cryo-EM structure of the post-hydrolysis protein state. Finally, we complement our mechanistic insights by exploring how ^{31}P NMR chemical shifts evolve along the proposed ATP hydrolysis pathway, linking their drastic changes, which are also observed in experimental NMR studies,[20] to key structural rearrangements. Overall, our contribution provides a full picture of the chemical step of ATP hydrolysis catalyzed by p97.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.5c00928.

Computational details about the adiabatic mappings and NEB minimum energy pathway calculations; method and basis set benchmark calculations for the minimum energy pathways as well as a QM region size convergence study; details about the QM/MM-MD enhanced sampling simulations, MBAR reweighting, uncertainties associated with the computed free energy profiles, details about the DFT NMR calculations; and illustrations of the H-bonds between the substrate and key active site residues in the educt and product states (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Christian Ochsenfeld − *Chair of Theoretical Chemistry, Department of Chemistry, University of Munich (LMU), D-81377 München, Germany; Max Planck Institute for Solid State Research, D-70569 Stuttgart, Germany;* orcid.org/0000-0002-4189-6558; Email: christian.ochsenfeld@uni-muenchen.de

Authors

Judit Katalin Szántó − *Chair of Theoretical Chemistry, Department of Chemistry, University of Munich (LMU), D-81377 München, Germany;* orcid.org/0000-0003-4767-0987

Andreas Hulm − *Chair of Theoretical Chemistry, Department of Chemistry, University of Munich (LMU), D-81377 München, Germany;* orcid.org/0000-0003-1268-7578

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.5c00928

Funding

Open access funded by Max Planck Society.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Dr. Jörg Kussmann (LMU Munich) for providing a development version of the FermiONs++ program package, and Prof. Dr. Anne Schütz (LMU Munich) for fruitful discussions. Financial support was provided by the "Deutsche Forschungsgemeinschaft" (DFG, German Research Foundation) within SFB 1309-325871075 "Chemical Biology of Epigenetic Modifications". C.O. acknowledges further support as Max-Planck-Fellow at the MPI-FKF Stuttgart.

https://doi.org/10.1021/acs.jctc.5c00928
J. Chem. Theory Comput. 2025, 21, 9459−9469

REFERENCES

(1) Hanson, P. I.; Whiteheart, S. W. AAA+ proteins: have engine, will work. *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 519−529.
(2) Rouiller, I.; DeLaBarre, B.; May, A. P.; Weis, W. I.; Brunger, A. T.; Milligan, R. A.; Wilson-Kubalek, E. M. Conformational changes of the multifunction p97 AAA ATPase during its ATPase cycle. *Nat. Struct. Biol.* **2002**, *9*, 950−957.
(3) DeLaBarre, B.; Brunger, A. T. Nucleotide dependent motion and mechanism of action of p97/VCP. *J. Mol. Biol.* **2005**, *347*, 437−452.
(4) Tang, W. K.; Li, D.; Li, C.-c.; Esser, L.; Dai, R.; Guo, L.; Xia, D. A novel ATP-dependent conformation in p97 N-D1 fragment revealed by crystal structures of disease-related mutants. *EMBO J.* **2010**, *29*, 2217−2229.
(5) Tonddast-Navaei, S.; Stan, G. Mechanism of transient binding and release of substrate protein during the allosteric cycle of the p97 nanomachine. *J. Am. Chem. Soc.* **2013**, *135*, 14627−14636.
(6) Schuller, J. M.; Beck, F.; Lössl, P.; Heck, A. J.; Förster, F. Nucleotide-dependent conformational changes of the AAA+ ATPase p97 revisited. *FEBS Lett.* **2016**, *590*, 595−604.
(7) Schütz, A. K.; Rennella, E.; Kay, L. E. Exploiting conformational plasticity in the AAA+ protein VCP/p97 to modify function. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, E6822−E6829.
(8) Shein, M.; Hitzenberger, M.; Cheng, T. C.; Rout, S. R.; Leitl, K. D.; Sato, Y.; Zacharias, M.; Sakata, E.; Schütz, A. K. Characterizing ATP processing by the AAA+ protein p97 at the atomic level. *Nat. Chem.* **2024**, *16*, 363−372.
(9) Prasad, B. R.; Plotnikov, N. V.; Warshel, A. Addressing open questions about phosphate hydrolysis pathways by careful free energy mapping. *J. Phys. Chem. B* **2013**, *117*, 153−163.
(10) Klähn, M.; Rosta, E.; Warshel, A. On the mechanism of hydrolysis of phosphate monoesters dianions in solutions and proteins. *J. Am. Chem. Soc.* **2006**, *128*, 15310−15323.
(11) Kozlova, M. I.; Shalaeva, D. N.; Dibrova, D. V.; Mulkidjanian, A. Y. Common mechanism of activated catalysis in P-loop fold nucleoside triphosphatases—United in diversity. *Biomolecules* **2022**, *12*, 1346.
(12) Berta, D.; Gehrke, S.; Nyíri, K.; Vértessy, B. G.; Rosta, E. Mechanism-based redesign of GAP to activate oncogenic Ras. *J. Am. Chem. Soc.* **2023**, *145*, 20302−20310.
(13) Pardos, J.; García-Martínez, A.; Ruiz-Pernía, J. J.; Tuñón, I. Mechanistic insights into GTP hydrolysis by the RhoA protein: Catalytic impact of glutamine tautomerism. *ACS Catal.* **2025**, *15*, 4415−4428.
(14) Kamerlin, S. C.; Sharma, P. K.; Prasad, R. B.; Warshel, A. Why nature really chose phosphate. *Q. Rev. Biophys.* **2013**, *46*, 1−132.
(15) Puchades, C.; Sandate, C. R.; Lander, G. C. The molecular principles governing the activity and functional diversity of AAA+ proteins. *Nat. Rev. Mol. Cell Biol.* **2020**, *21*, 43−58.
(16) Leipe, D. D.; Koonin, E. V.; Aravind, L. Evolution and classification of P-loop kinases and related proteins. *J. Mol. Biol.* **2003**, *333*, 781−815.
(17) Iyer, L. M.; Leipe, D. D.; Koonin, E. V.; Aravind, L. Evolutionary history and higher order classification of AAA+ ATPases. *J. Struct. Biol.* **2004**, *146*, 11−31.
(18) Steel, G. J.; Harley, C.; Boyd, A.; Morgan, A. A screen for dominant negative mutants of SEC18 reveals a role for the AAA protein consensus sequence in ATP hydrolysis. *Mol. Biol. Cell* **2000**, *11*, 1345−1356.
(19) Wendler, P.; Ciniawsky, S.; Kock, M.; Kube, S. Structure and function of the AAA+ nucleotide binding pocket. *Biochim. Biophys. Acta, Mol. Cell Res.* **2012**, *1823*, 2−14.
(20) Rydzek, S.; Shein, M.; Bielytskyi, P.; Schütz, A. K. Observation of a transient reaction intermediate illuminates the mechanochemical cycle of the AAA-ATPase p97. *J. Am. Chem. Soc.* **2020**, *142*, 14472−14480.
(21) Schmidt, H.; Gleave, E. S.; Carter, A. P. Insights into dynein motor domain function from a 3.3-Å crystal structure. *Nat. Struct. Mol. Biol.* **2012**, *19*, 492−497.
(22) Schmidt, H.; Carter, A. P. Review: Structure and mechanism of the dynein motor ATPase. *Biopolymers* **2016**, *105*, 557−567.
(23) Afanasyeva, A.; Hirtreiter, A.; Schreiber, A.; Grohmann, D.; Pobegalov, G.; McKay, A. R.; Tsaneva, I.; Petukhov, M.; Käs, E.; Grigoriev, M.; Werner, F. Lytic water dynamics reveal evolutionarily conserved mechanisms of ATP hydrolysis by TIP49 AAA+ ATPases. *Structure* **2014**, *22*, 549−559.
(24) Hänzelmann, P.; Schindelin, H. Structural basis of ATP hydrolysis and intersubunit signaling in the AAA+ ATPase p97. *Structure* **2016**, *24*, 127−139.
(25) Karata, K.; Inagawa, T.; Wilkinson, A. J.; Tatsuta, T.; Ogura, T. Dissecting the role of a conserved motif (the second region of homology) in the AAA family of ATPases: site-directed mutagenesis of the ATP-dependent protease FtsH. *J. Biol. Chem.* **1999**, *274*, 26225−26232.
(26) Hattendorf, D. A.; Lindquist, S. L. Cooperative kinetics of both Hsp104 ATPase domains and interdomain communication revealed by AAA sensor-1 mutants. *EMBO J.* **2002**, *21*, 12−21, DOI: 10.1093/emboj/21.1.12.
(27) Wang, Q.; Song, C.; Irizarry, L.; Dai, R.; Zhang, X.; Li, C.-C. H. Multifunctional roles of the conserved Arg residues in the second region of homology of p97/valosin-containing protein. *J. Biol. Chem.* **2005**, *280*, 40515−40523.
(28) Ogura, T.; Whiteheart, S. W.; Wilkinson, A. J. Conserved arginine residues implicated in ATP hydrolysis, nucleotide-sensing, and inter-subunit interactions in AAA and AAA+ ATPases. *J. Struct. Biol.* **2004**, *146*, 106−112.
(29) Kamerlin, S. C. L.; Wilkie, J. The role of metal ions in phosphate ester hydrolysis. *Org. Biomol. Chem.* **2007**, *5*, 2098−2108.
(30) Mateeva, T.; Klähn, M.; Rosta, E. Structural dynamics and catalytic mechanism of ATP13A2 (PARK9) from simulations. *J. Phys. Chem. B* **2021**, *125*, 11835−11847.
(31) Tang, W. K.; Xia, D. Altered intersubunit communication is the molecular basis for functional defects of pathogenic p97 mutants. *J. Biol. Chem.* **2013**, *288*, 36624−36635.
(32) Prieß, M.; Göddeke, H.; Groenhof, G.; Schaäfer, L. V. Molecular mechanism of ATP hydrolysis in an ABC transporter. *ACS Cent. Sci.* **2018**, *4*, 1334−1343.
(33) Kiani, F. A.; Fischer, S. Stabilization of the ADP/metaphosphate intermediate during ATP hydrolysis in pre-power stroke myosin: quantitative anatomy of an enzyme. *J. Biol. Chem.* **2013**, *288*, 35569−35580.
(34) Hayashi, S.; Ueno, H.; Shaikh, A. R.; Umemura, M.; Kamiya, M.; Ito, Y.; Ikeguchi, M.; Komoriya, Y.; Iino, R.; Noji, H. Molecular mechanism of ATP hydrolysis in F1-ATPase revealed by molecular simulations and single-molecule observations. *J. Am. Chem. Soc.* **2012**, *134*, 8447−8454.
(35) Szántó, J. K.; Dietschreit, J. C.; Shein, M.; Schütz, A. K.; Ochsenfeld, C. Systematic QM/MM Study for Predicting 31P NMR Chemical Shifts of Adenosine Nucleotides in Solution and Stages of ATP Hydrolysis in a Protein Environment. *J. Chem. Theory Comput.* **2024**, *20*, 2433−2444.
(36) Grimme, S.; Brandenburg, J. G.; Bannwarth, C.; Hansen, A. Consistent structures and interactions by density functional theory with small atomic orbital basis sets. *J. Chem. Phys.* **2015**, *143*, No. 054107, DOI: 10.1063/1.4927476.
(37) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696−3713.
(38) Riplinger, C.; Sandhoefer, B.; Hansen, A.; Neese, F. Natural triple excitations in local coupled cluster calculations with pair natural orbitals. *J. Chem. Phys.* **2013**, *139*, No. 134101, DOI: 10.1063/1.4821834.
(39) Riplinger, C.; Pinski, P.; Becker, U.; Valeev, E. F.; Neese, F. Sparse maps—A systematic infrastructure for reduced-scaling electronic structure methods. II. Linear scaling domain based pair natural orbital coupled cluster theory. *J. Chem. Phys.* **2016**, *144*, No. 024109, DOI: 10.1063/1.4939030.

https://doi.org/10.1021/acs.jctc.5c00928
J. Chem. Theory Comput. 2025, 21, 9459−9469

(40) Kussmann, J.; Ochsenfeld, C. Pre-selective screening for matrix elements in linear-scaling exact exchange calculations. *J. Chem. Phys.* **2013**, *138*, No. 134114.
(41) Kussmann, J.; Ochsenfeld, C. Preselective screening for linear-scaling exact exchange-gradient calculations for graphics processing units and general strong-scaling massively parallel calculations. *J. Chem. Theory Comput.* **2015**, *11*, 918−922.
(42) Kussmann, J.; Ochsenfeld, C. Hybrid CPU/GPU integral engine for strong-scaling ab initio methods. *J. Chem. Theory Comput.* **2017**, *13*, 3153−3159.
(43) Lehtola, S.; Steigemann, C.; Oliveira, M. J. T.; Marques, M. A. L. Recent developments in libxc—A comprehensive library of functionals for density functional theory. *SoftwareX* **2018**, *7*, 1−5.
(44) Eastman, P.; Pande, V. OpenMM: A hardware-independent framework for molecular simulations. *Comput. Sci. Eng.* **2010**, *12*, 34−39.
(45) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D. e. a.; et al. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* **2017**, *13*, No. e1005659.
(46) Hulm, A.; Dietschreit, J. C.; Ochsenfeld, C. Statistically optimal analysis of the extended-system adaptive biasing force (eABF) method. *J. Chem. Phys.* **2022**, *157*, No. 024110, DOI: 10.1063/5.0095554.
(47) Hulm, A.; Lemke, Y.; Johannes, D.; Glinkina, L.; Stan-Bernhardt, A. *adaptive_sampling*. https://github.com/ochsenfeld-lab/adaptive_sampling.
(48) Laqua, H.; Thompson, T. H.; Kussmann, J.; Ochsenfeld, C. Highly efficient, linear-scaling seminumerical exact-exchange method for graphic processing units. *J. Chem. Theory Comput.* **2020**, *16*, 1456−1468.
(49) Laqua, H.; Dietschreit, J. C.; Kussmann, J.; Ochsenfeld, C. Accelerating Hybrid Density Functional Theory Molecular Dynamics Simulations by Seminumerical Integration, Resolution-of-the-Identity Approximation, and Graphics Processing Units. *J. Chem. Theory Comput.* **2022**, *18*, 6010−6020.
(50) Kussmann, J.; Laqua, H.; Ochsenfeld, C. Highly efficient resolution-of-identity density functional theory calculations on central and graphics processing units. *J. Chem. Theory Comput.* **2021**, *17*, 1512−1521.
(51) Ochsenfeld, C.; Kussmann, J.; Koziol, F. Ab initio NMR spectra for molecular systems with a thousand and more atoms: a linear-scaling method. *Angew. Chem.* **2004**, *116*, 4585−4589.
(52) Field, M. J.; Bash, P. A.; Karplus, M. A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations. *J. Comput. Chem.* **1990**, *11*, 700−733.
(53) Wilson, P. J.; Bradley, T. J.; Tozer, D. J. Hybrid exchange-correlation functional determined from thermochemical data and ab initio potentials. *J. Chem. Phys.* **2001**, *115*, 9233−9242.
(54) Jensen, F. Segmented contracted basis sets optimized for nuclear magnetic shielding. *J. Chem. Theory Comput.* **2015**, *11*, 132−138.
(55) Henkelman, G.; Jónsson, H. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *J. Chem. Phys.* **2000**, *113*, 9978−9985.
(56) Dietschreit, J. C. B.; Diestler, D. J.; Hulm, A.; Ochsenfeld, C.; Gómez-Bombarelli, R. From free-energy profiles to activation free energies. *J. Chem. Phys.* **2022**, *157*, No. 084113, DOI: 10.1063/5.0102075.
(57) Díaz Leines, G.; Ensing, B. Path finding on high-dimensional free energy landscapes. *Phys. Rev. Lett.* **2012**, *109*, No. 020601.
(58) Fu, H.; Zhang, H.; Chen, H.; Shao, X.; Chipot, C.; Cai, W. Zooming across the free-energy landscape: shaving barriers, and flooding valleys. *J. Phys. Chem. Lett.* **2018**, *9*, 4738−4745.
(59) Fu, H.; Shao, X.; Cai, W.; Chipot, C. Taming rugged free energy landscapes using an average force. *Acc. Chem. Res.* **2019**, *52*, 3254−3264.
(60) Hulm, A.; Ochsenfeld, C. Improved Sampling of Adaptive Path Collective Variables by Stabilized Extended-System Dynamics. *J. Chem. Theory Comput.* **2023**, *19*, 9202−9210.
(61) Aho, N.; Groenhof, G.; Buslaev, P. Do all paths lead to Rome? How reliable is umbrella sampling along a single path? *J. Chem. Theory Comput.* **2024**, *20*, 6674−6686.
(62) Kiani, F. A.; Fischer, S. Comparing the catalytic strategy of ATP hydrolysis in biomolecular motors. *Phys. Chem. Chem. Phys.* **2016**, *18*, 20219−20233.
(63) McGrath, M. J.; Kuo, I.-F. W.; Hayashi, S.; Takada, S. Adenosine triphosphate hydrolysis mechanism in kinesin studied by combined quantum-mechanical/molecular-mechanical metadynamics simulations. *J. Am. Chem. Soc.* **2013**, *135*, 8908−8919.
(64) García-Martínez, A.; Zinovjev, K.; Ruiz-Pernía, J. J.; Tuñón, I. Conformational changes and ATP hydrolysis in zika helicase: the molecular basis of a biomolecular motor unveiled by multiscale simulations. *J. Am. Chem. Soc.* **2023**, *145*, 24809−24819.
(65) Ye, Y.; Meyer, H. H.; Rapoport, T. A. The AAA ATPase Cdc48/p97 and its partners transport proteins from the ER into the cytosol. *Nature* **2001**, *414*, 652−656.
(66) Mader, S. L.; Lopez, A.; Lawatscheck, J.; Luo, Q.; Rutz, D. A.; Gamiz-Hernandez, A. P.; Sattler, M.; Buchner, J.; Kaila, V. R. Conformational dynamics modulate the catalytic activity of the molecular chaperone Hsp90. *Nat. Commun.* **2020**, *11*, No. 1410.
(67) Crampton, D. J.; Mukherjee, S.; Richardson, C. C. DNA-induced switch from independent to sequential dTTP hydrolysis in the bacteriophage T7 DNA helicase. *Mol. Cell* **2006**, *21*, 165−174.
(68) Mardirossian, N.; Head-Gordon, M. Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. *Mol. Phys.* **2017**, *115*, 2315−2372.
(69) Jin, Y.; Molt, R. W., Jr.; Waltho, J. P.; Richards, N. G.; Blackburn, G. M. 19F NMR and DFT Analysis Reveal Structural and Electronic Transition State Features for RhoA-Catalyzed GTP Hydrolysis. *Angew. Chem., Int. Ed.* **2016**, *55*, 3318−3322.

https://doi.org/10.1021/acs.jctc.5c00928
J. Chem. Theory Comput. 2025, 21, 9459−9469

Supplementary Material: On the Molecular Mechanism of ATP Hydrolysis Catalyzed by p97: a QM/MM Study

Judit Katalin Szántó,[1] Andreas Hulm,[1] Christian Ochsenfeld*,[1,2]

[1]Chair of Theoretical Chemistry, Department of Chemistry,
University of Munich (LMU), Butenandtstr. 7, D-81377 München, Germany,
[2]Max Planck Institute for Solid State Research,
Heisenbergstr. 1, D-70569 Stuttgart, Germany
*E-Mail: christian.ochsenfeld@uni-muenchen.de

Contents

1 QM/MM calculations

QM/MM simulations were performed in our in-house program package FERMIONS++[1–3]. Unless otherwise noted, for the QM part Grimme's PBEh-3c DFT functional was used [4]. Significant speed-ups were achieved using the sn-LinK method [5–7] and the RI-J approximation [8] for fast evaluation of exact exchange and Coulomb energy terms, respectively. An automatic workflow was employed to place H-atoms as links between the MM and QM region. These link atoms were introduced between C_β and C_α atoms if single amino acids were selected into the QM region (Figure S1 a) and between C_α and C atoms in case of amino acid sequences (Figure S1 b). The QM/MM boundaries were chosen such that peptide bonds along the protein backbone and polar bonds were not cut.

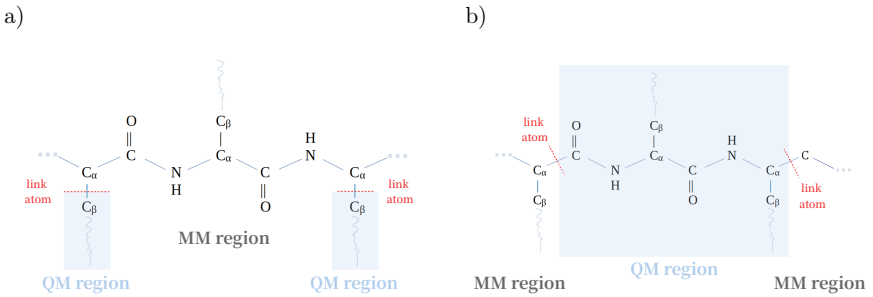


Figure S1: Representation of QM/MM partitioning scheme: a) Bonds between the C_α and C_β atoms were cut when including single amino acids in the QM region and b) Bonds between the C_α and C atoms were cut when including a series of amino acids.

QM/MM calculations were carried out with the equilibrated educt structure, which in contrast to the static X-Ray structure contains amino acids in a catalysis-ready conformation.

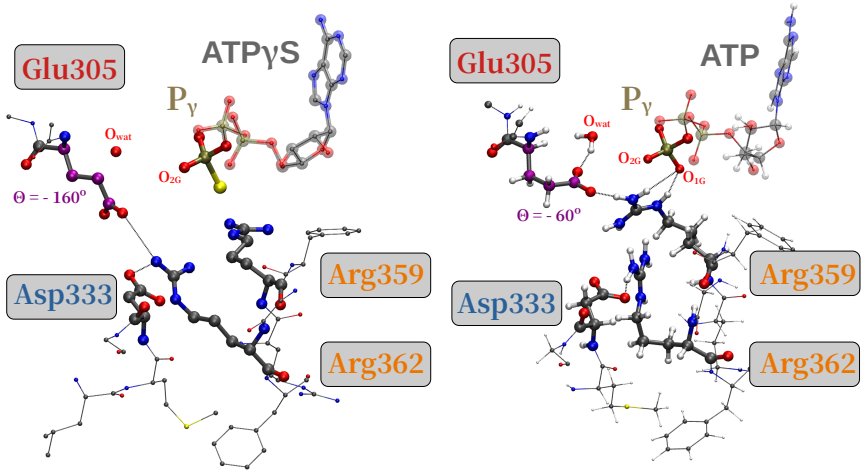


Figure S2: Comparison of the active site conformations in p97. Right: X-ray structure (PDB: 4KO8) with the non-hydrolyzable ATP analog ATPγS. Left: ATP-bound protein state. Key residues involved in nucleotide sensing and catalysis (E305, R359, R362, D333) are shown to illustrate conformational differences.

2 Structure optimizations and adiabatic mappings

For the adiabatic mappings (AMs), snapshots were extracted from the MM-MD simulations of the ATP-bound state of p97 [9, 10], followed by QM/MM structure optimizations. The DL-Find library [11] implemented in PyChemShell [12] was used for structure optimizations connected to the Python-interface of FermiONs++. Convergence criteria for DFT QM/MM structure optimizations were set as follows:

The optimized structures were used as starting points for the adiabatic mapping pathways, where reactants were stepwise optimized while pulling along a predefined collective variable (CV) using harmonic restraints with a force constant of 1 kJ/molÅ^2. The used collective variable is the $d(O_{wat} - P_\gamma)$ distance (see Fig. S7 and Fig. S6) or the linear combination of the bond formation ($d(O_{wat} - P_\gamma)$) and bond cleavage ($d(P_\gamma - O_{3B})$) (see Fig. S5), as implemented in the *colvars* module of the adaptive-sampling python package[13]. The proton transfer from the nucleophilic water is not explicitly biased,

Table S1: Convergence criteria for structure optimization of ATP in p97

Criteria	Threshold
Energy	4×10^{-4} E_h
RMS gradient	1×10^{-3} E_h
Max. gradient	8×10^{-3} E_h
RMS step	5×10^{-3} E_h
Max. step	5×10^{-1} E_h

allowing the proton to migrate freely without being constrained by the reaction coordinate. Constrained optimizations were performed along the chosen reaction coordinate. In each step, the CV value is, changed and then fixed, while the system was minimized. All residues in a radius of 10 Å around the ATP molecule were relaxed. The QM region contained 126 atoms from amino acids which were found within 3.5 Angstroms around the P_γ, the two O atoms in the GluE305 side group and the attacking water molecule: Gly248, Thr249, Gly250, Lys251, Glu305, Asn348, Arg359, the Mg^{2+} ion at the binding site, and the closest water molecules. From the ATP molecule only the phosphate backbone is included in the QM region, the adenosine remains in the MM subsystem.

Within 6 Å around the P_γ atom, the X-Ray structure contains 8 stable water molecules, without the resolved positions of the protons. Therefore, the first question to answer is which water molecule is a good candidate for the nucleophilic attack.

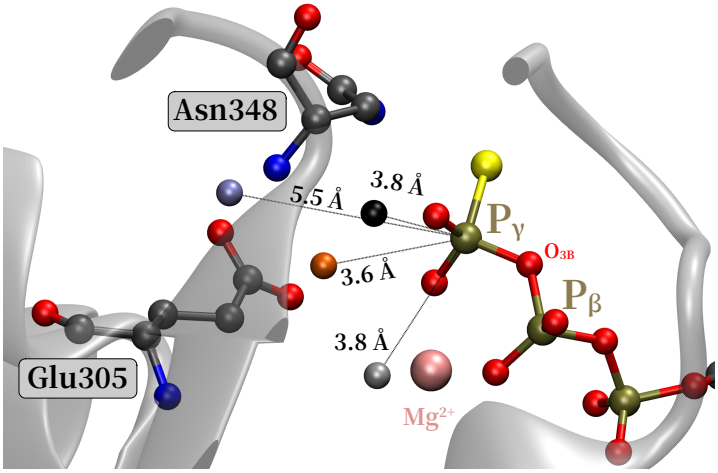


Figure S3: Oxygen atoms of buried water molecules resolved in the X-Ray structure of ATPγS-bound p97 (PDB 4KO8 [14]) and the phosphate backbone of the substrate. The colors of the water molecules correspond to the color coding used in the main manuscript (see Fig. 4-5 in the main text).

In the first approach, we have tested the nucleophilic attack of the closest water molecules to the P_γ and P_β atoms.

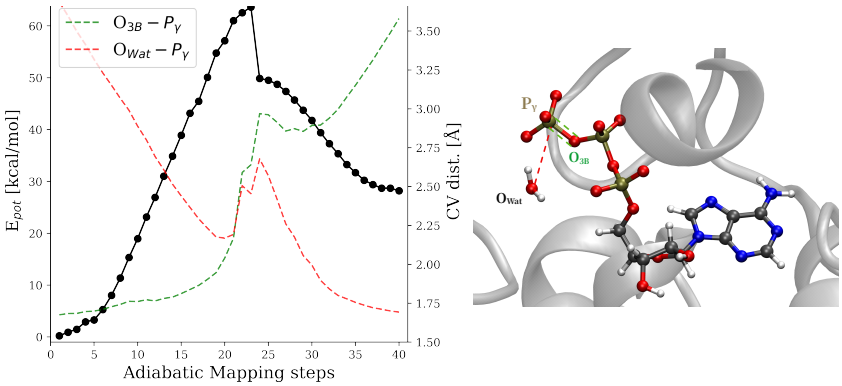


Figure S4: Adiabatic mapping pathway - substrate-assisted mechanism

In the second approach, the two water molecules closest to the Asn348 and the two O atoms of the GluE305 were considered as candidates for the nucleophilic attack. In this simulation, the lytic water molecule is the one closest to the Pg atom and a second, assisting water molecule is the proton donor to the GluE305. A linear combination with decreasing $d(O_{Wat-attack} - P_\gamma)$ and $d(H_{Wat-attack} - O_{Wat-assist})$ distances was used as collective variable.

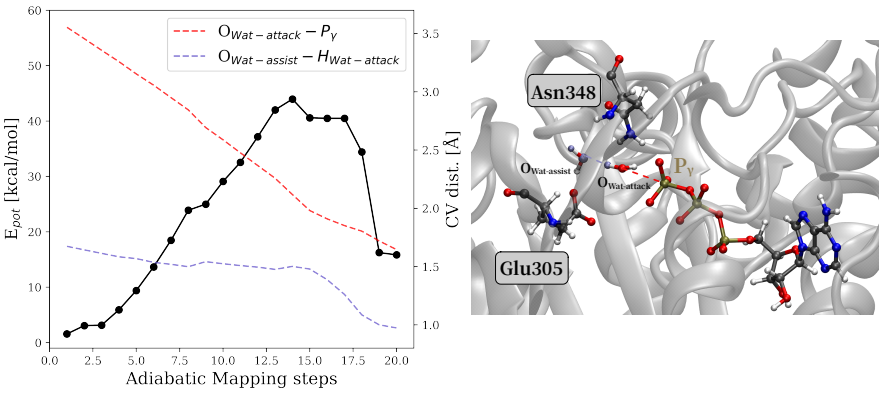


Figure S5: Adiabatic mapping pathway - Base-assisted mechanism (Glu305) - 2 water mechanism

For the single water mechanism the $d(O_{wat} - P_\gamma)$ distance was used as collective variable.

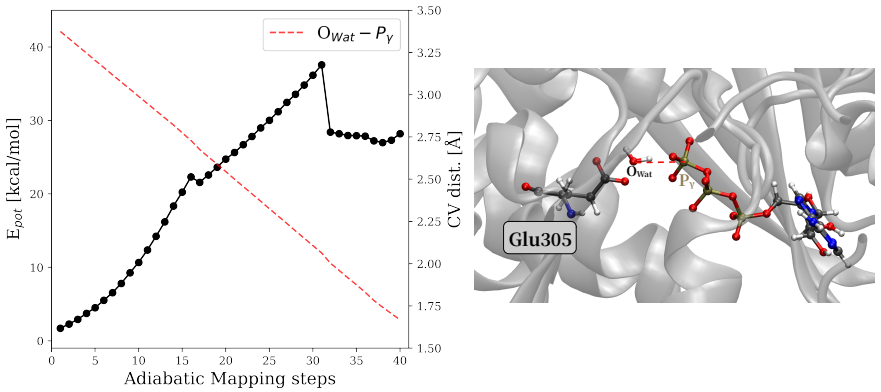


Figure S6: Adiabatic mapping pathway - Base-assisted mechanism (Glu305) - 1 water mechanism (channel B)

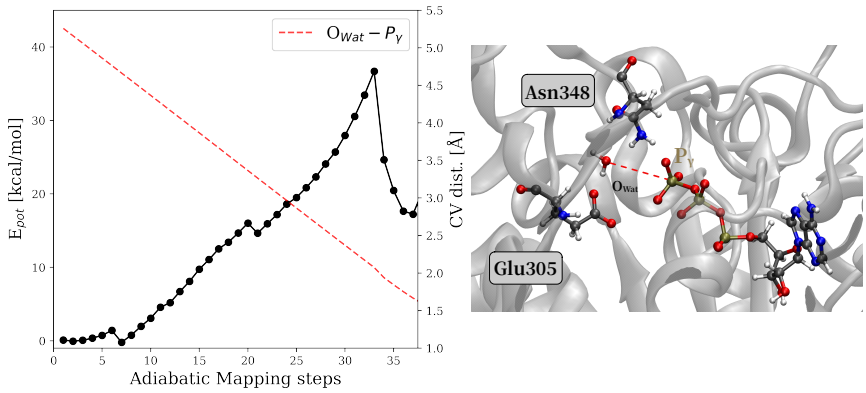


Figure S7: Adiabatic mapping pathway - Base-assisted mechanism (Glu305) - 1 water mechanism (channel A)

After finding adiabatic mapping pathways for the first reaction step, NEB simulations were performed and the glutamate-as-base mechanism was selected as a plausible mechanism. To explore the last step of the reaction and reach the product state, the adiabatic mapping was started from the intermediate reached in the first step. The collective variable for the proton transfer step was defined as the distance between the H^+ and the closest O of the P_γ atom. The CV was decreased by 0.04 Å in each step, and then fixed, while the energy of the system was minimized.

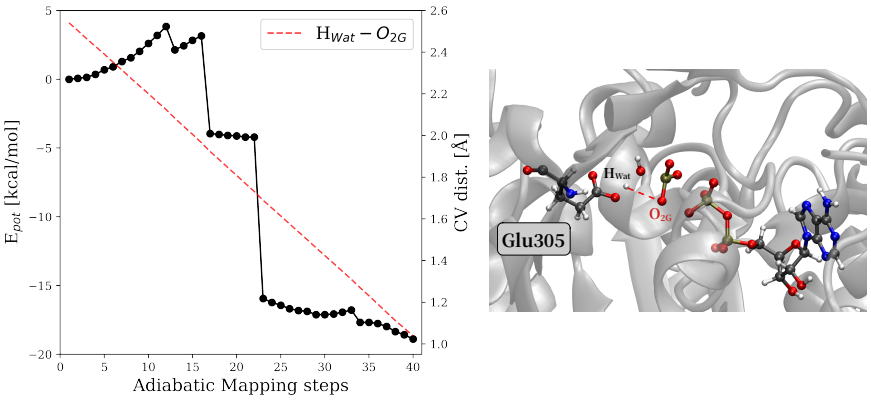


Figure S8: Adiabatic mapping pathway of the second reaction step

There are several key features that influence reactivity, but for this initial exploration,

we used a single distance as the collective variable (CV) or the linear combination of the breaking and forming bond. In all adiabating mapping simulations, the proton migrates freely to GluE305 or to the substrate without being constrained by the CV. AM is highly sensitive to the CV and the initial configuration[15], the pathways we have found using adiabatic mapping show discontinuities, which are very frequent and typical for this approach [16].

3 Finding minimum energy pathways using the nudged elastic band method

NEB simulations were performed using FERMIONS++[1–3] together with the FENEB module of the adaptive-sampling package[13]. For each MEP 16 equidistant NEB images were created using linear interpolation between the two optimized endpoints of the adiabatic mapping. 1000 steps of steepest descent optimization were carried out for all NEB images, including the two extreme points corresponding to the reactant and the educt structure. Adjacent images on this path are connected by springs with a force constant k to ensure an equidistant spacing along the pathway. Here, the improved tangent force estimate as described by Henkelmann and Jonsson was applied[17]. The NEB force was used only for selected substrate atoms while the environment was allowed to relax freely. Additionally, in each optimization step the spring force is fully optimized to enforce equidistant spacing of NEB images. This ensured that the resulting MEPs are well suited to serve as path CVs[18] for path WTM-eABF free energy simulations[19–21]. Hence, each image of the optimized minimum energy path was heated and equilibrated for subsequent free energy simulations, confining each MD simulation to its corresponding path node using a harmonic potential with force constant 120 kcal/mol.

For the Glu305 base-assisted mechanisms both reaction channels we observe an almost collinear alignment of the O_{wat}, P_γ and O_{3B} atoms around the TS, before the nucleophilic attack occurs. For channel B the water molecule forms a $\Psi = 135°$ angle in the reactant state, whereas the channel A water molecule forms a $\Psi = 150°$ angle, making it better positioned for the attack.

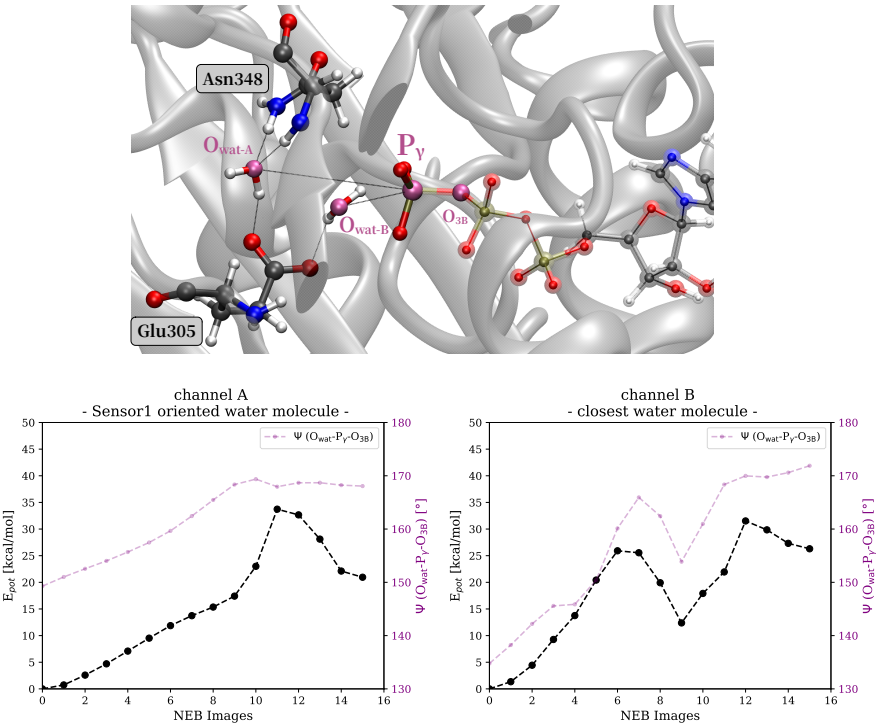


Figure S9: Base-assisted mechanism (Glu). **Top:** the binding site of p97 with ATP, two water molecules corresponding to channel A and B, Sensor 1 Asn348 and the proton acceptor GluE305. Atoms marked with purple have an almost collinear alignment before the nucleophilic attack. **Bottom:** $\Psi(O_{wat} - P_\gamma - O_{3B})$ angle values for images of the channel A and channel B NEB paths.

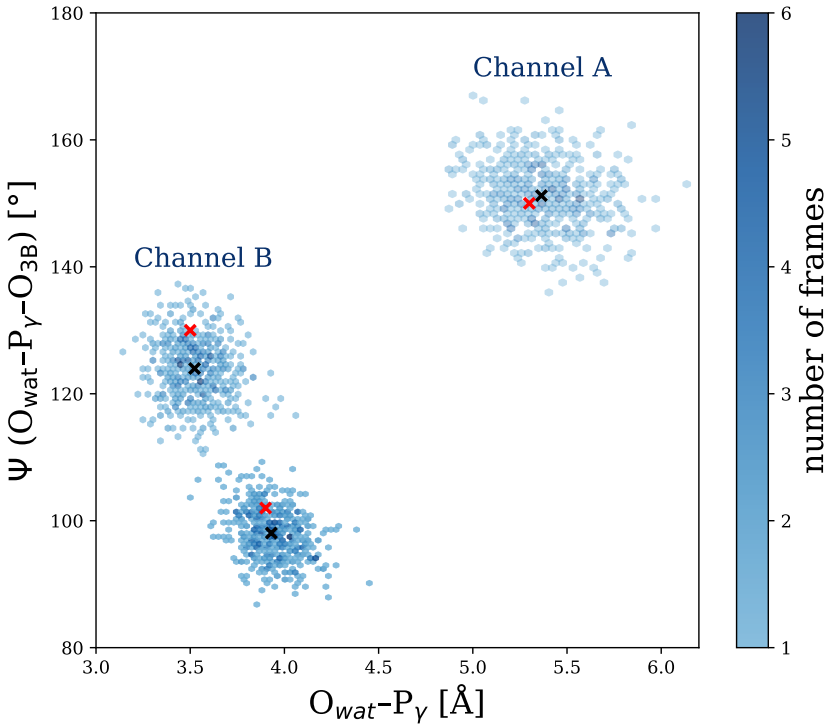


Figure S10: 2D hexbin plot depicting orientations of the closest buried water molecules to the P_γ atom in the active site during the 1-microsecond 1 μs MM-MD trajectory [9]. Black crosses indicate the median values of the $\Psi(O_{\text{wat}} - P_\gamma - O_{3\text{B}})$ angle and $O_{\text{wat}} - P_\gamma$ distance, while red crosses mark the educt structure used in this study. The nucleophilic attack of the water molecule found at $\Psi(O_{\text{wat}} - P_\gamma - O_{3\text{B}}) = 100°$ in the hexbin plot leads to the substrate-assisted mechanism.

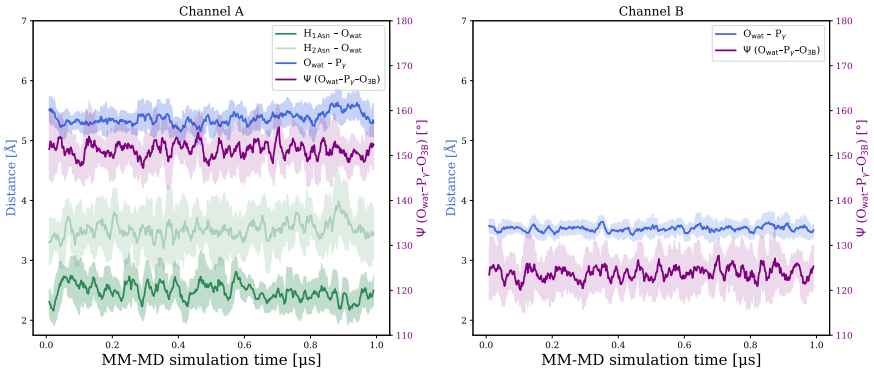


Figure S11: Rolling averages (window size = 10 frames) over the MM-MD trajectory[9] capturing the educt state. Shown are: $O_{wat}-P_\gamma$ distances, H-bonds formed with Asn348 and the $\Psi(O_{wat}-P_\gamma-O_{3B})$ and the attacking angle for the channel A and channel B water molecules. Shaded areas indicate the range of ±1 standard deviation around the rolling average.

4 Benchmark calculations - the influence of different DFT functionals and basis sets

We computed the single point energies of the images from the refined NEB path with various DFT functionals and basis sets using ORCA [22]. We used the following DFT functionals: PBEh-3c, ωB97M-V, B3LYP-D3(BJ) with double-ζ and triple-ζ basis sets. The images of the NEB optimized path obtained with PBEh-3c were used to estimate the electronic energies at different theory levels.

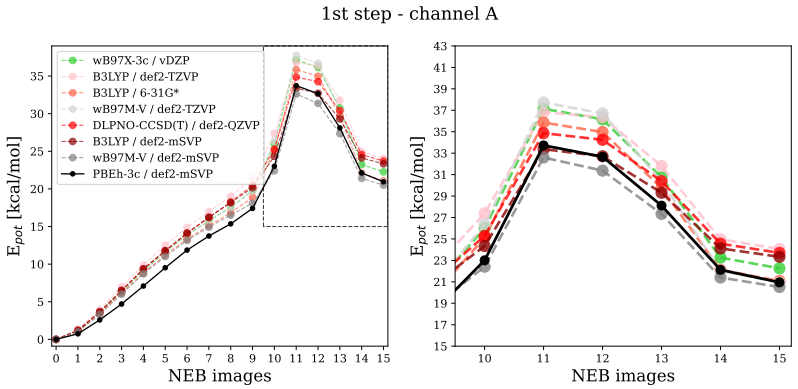


Figure S12: Benchmarking the energetics of the first step, channel A.

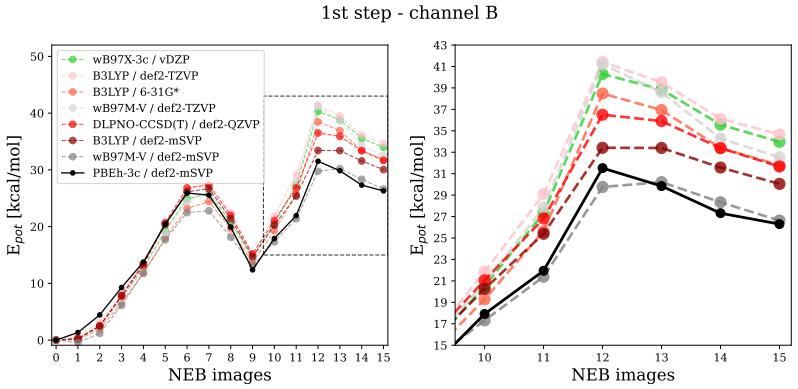


Figure S13: Benchmarking the energetics of the first step, channel B.

For both channels PBEh-3c[4] underestimates the barrier compared to ωB97X-3c[23] by 5 kcal (channel A) and by 10 kcal/mol (channel B). The ωB97X-3c range-separated composite method builds on the ωB97X-V functional, using a molecule-optimized polarized valence double-zeta (vDZP) basis set and a tailored D4 dispersion correction. We find a strong basis set influence. Using triple-ζ basis sets we get very close to the DLPNO-CCSD(T)[24] and ωB97X-3c barrier. We chose the PBEh-3c approach with the double-ζ basis set to enable extensive sampling. Additionally, for all functionals, the activation barrier of channel B is higher than that of channel A, and the intermediate in channel B (6 kcal/mol), representing a shallow minimum, is less stable than the intermediate in channel A (12 kcal/mol).

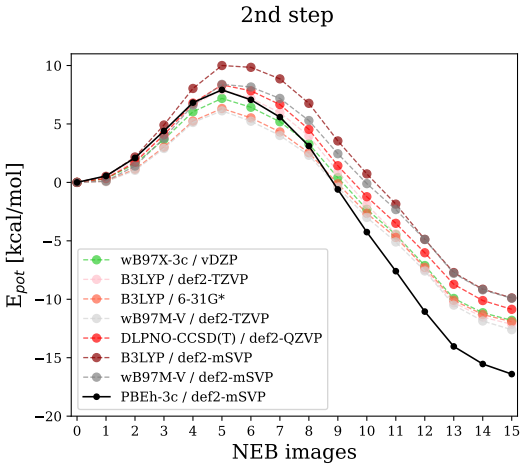


Figure S14: Benchmarking the energetics of the second step.

5 Benchmark calculations - the influence of the cutoff of electrostatic interactions with the MM subsystem

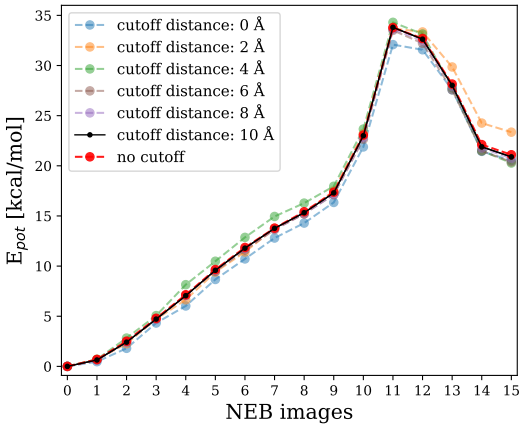


Figure S15: The influence of the cutoff used for the QM–MM electrostatic interactions.

We maintain high accuracy by applying a 10 Å cutoff (black curve) and achieving approximately a 4-fold speedup with respect to those calculations, where the electrostatic interactions between the QM and the MM subsystem were not cut off (red curve). The overall small consequence of fully neglecting the electrostatic interaction with MM atoms (cutoff 0 Å) indicates that the QM region captures the electrostatics of the active site sufficiently well.

6 Benchmark calculations - QM region size

We recomputed the refined NEB path with various QM regions. Water molecules, amino acids or series of amino acids were selected in the QM region is any atoms of these molecules fall within the d distance defined around the reaction center formed by four key atoms from the first reaction step: the two O atoms of Glu305, the O_{wat} atom of the attacking water molecule and P_γ. The cut between the QM and MM regions and the placement of link atoms was made as shown in Figure S1.

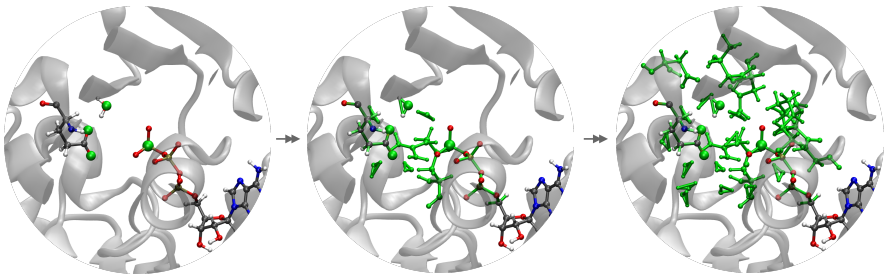


Figure S16: Increasing the QM region around the reaction center defined by the two O atoms of Glu305, O_{wat} and P_γ

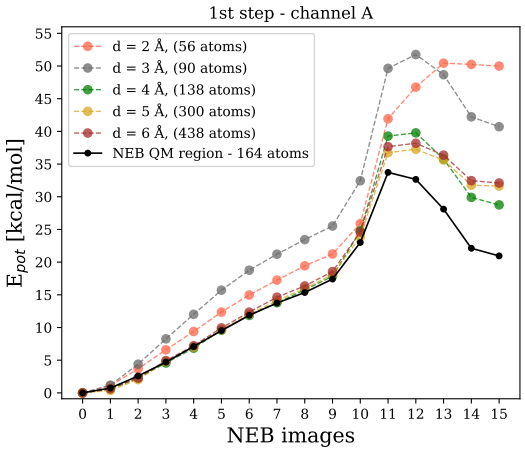


Figure S17: First step: Benchmark calculations for the number of QM atoms used in the QM/MM simulations.

The QM region was systematically increased, including more amino acids and neighboring solvent molecules. All QM regions include the phosphate backbone of the ATP molecule, the nucleophilic water molecule, Glu305, the Mg^{2+} ion, the nucleophilic water and Arg (R359). The smallest QM region (d = 2 Å) does not contain Asn348, without this key residue the minimum corresponding to the intermediate cannot be located. Nevertheless, for cutoff 4 Å(138 atoms) or larger the energy curve is largely conserved, suggesting that the NEB QM region of 164 atoms is a safe choice.

Energy barriers slightly increase for all tested QM regions compared to the QM region used in NEB optimizations. This can be understood by considering that for the test QM regions only single-point energies were computed without re-optimizing the NEB path. The 4.3 kcal/mol difference between the barriers of the converged paths with 138–438 QM atoms and the NEB path with 164 QM atoms can be attributed to the effect of the NEB optimization. The 1000 iterations refer to 1000 steps of the Nudged Elastic Band (NEB) method. During these steps, the reaction path and the images along the path are iteratively refined. The first and last NEB images were frozen during this re-optimization, only focusing on replacing the energy barrier. As shown in Fig. S18 and Fig. S20, this gradual optimization improves the minimum energy path and reduces the energy barrier by approximately 4 kcal/mol, such that the energy gap between the NEB-optimized and recalculated pathways disappears.

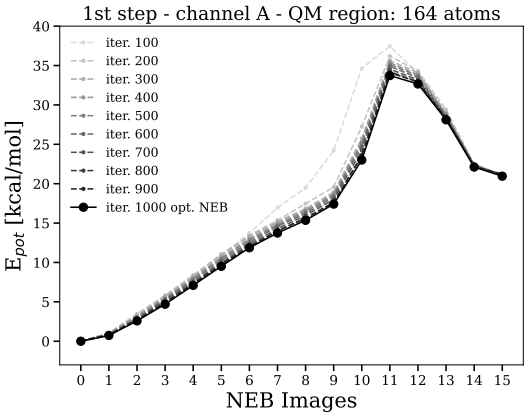


Figure S18: First step: the effect of 1000 NEB iterations on the minimum energy path

Below the same benchmark is shown for the second step: proton transfer from GluE305 to a phosphate oxygen. However, the influence of the QM region on results is less severe than for the first step, and the H^+ transfer process observed in the second step is more local. Therefore, we choose a smaller QM region, consisting of 123 QM atoms.

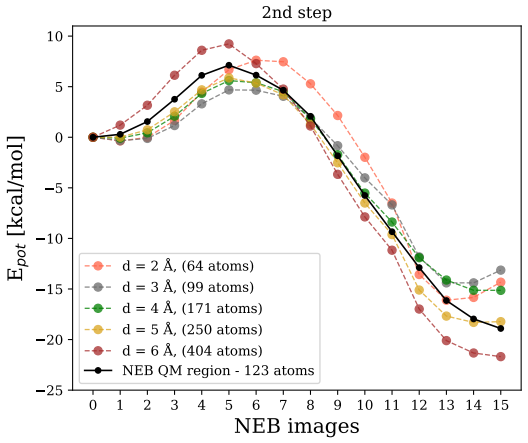


Figure S19: Second step: Benchmark calculations for the number of QM atoms used in the QM/MM simulations.

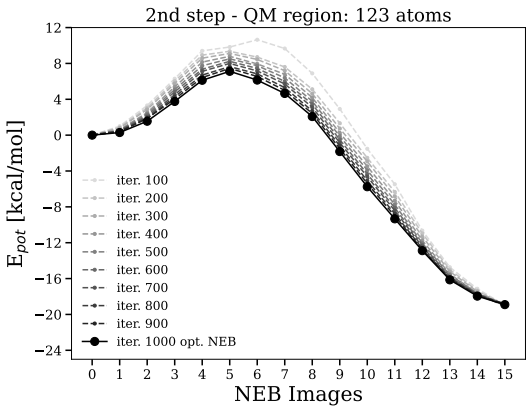


Figure S20: Second step: the effect of 1000 NEB iterations on the minimum energy path.

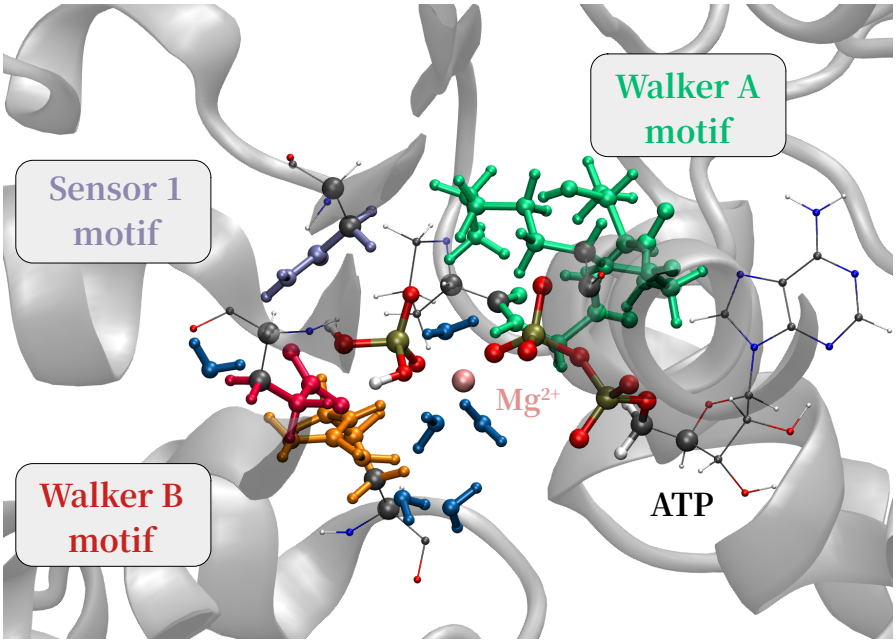


Figure S21: Second step: the QM region consist of 123 atoms - the phosphate backbone of the ATP molecule, the six closest water molecules (blue), Glu305 from the Walker B motif (red), Asn348 from the Sensor 1 motif (purple), Gly248, Thr249, Gly250, and Lys251 from the Walker A motif (green), as well as Arg359 from the adjacent protein subunit (orange). Grey spheres indicate carbon atoms at the QM/MM boundaries, where hydrogen link atoms were introduced; only non-polar C–C bonds were cut to define the QM region.

7 Enhanced sampling using the WTM-eABF method

Reaction free energy profiles were computed from biased QM/MM-MD simulations using the path Well-Tempered Metadynamics extended-system Adaptive Biasing Force (WTM-eABF) algorithm as implemented in the adaptive-sampling package [21]. In the extended-system formulation, the bias forces are not applied directly to the collective variable (CV) but to a fictitious particle. This particle, with a mass of 40 a.u., is coupled to the CV via a harmonic spring, with a thermal width set to 0.01. The CV space is discretized with a bin width of 0.01, and the WTM and ABF forces are accumulated on this grid. The ABF force was scaled up using a linear ramping scheme, where the force applied in each bin was proportional to the number of collected samples with the full force applied only in bins with 200 samples or more. For the metadynamics potential, 4000 K was set as the WTM bias temperature and every 10 fs a new Gaussian hill with a height of 0.1 kJ/mol and with a variance of 0.03 was deposited. The initial height of the Gaussian hills decreases over time due to the well-tempered scaling.

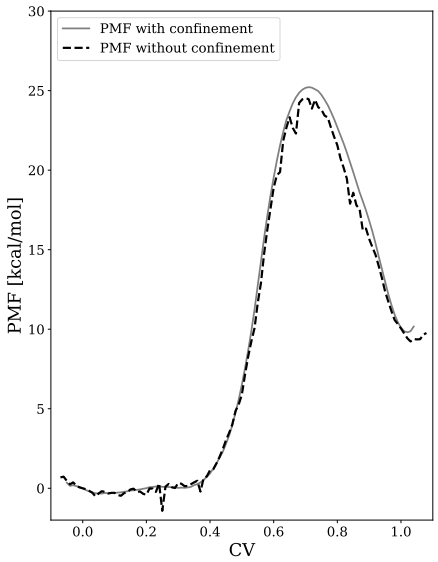


Figure S22: Effect of removing the confinement to the path CV on the PMF profile.

During sampling, harmonic walls with a force constant of 100 kJ/molÅ^2 are applied to confine the CV to the range of interest, thus preventing the system from exploring unwanted configurations. The Multistate Bennett's Acceptance Ratio (MBAR) was used as estimator to compute ensemble averages and PMFs using the unbiased weights of the simulation frames [25, 26]. The MBAR equations are solved self consistently and the

PMF profiles are computed. The PMF can be reconstructed without the bias introduced with the confinement of the path CV, by removing the harmonic confinement potential from the PMF. We apply the geometric path CV definition [18], which requires the selection of an appropriate CV space. While the environment is not confined, the four breaking and forming bond distances (show in Fig. S23) that are included in the CV space of the path CV are forced to stay close to the MEP during the simulation.

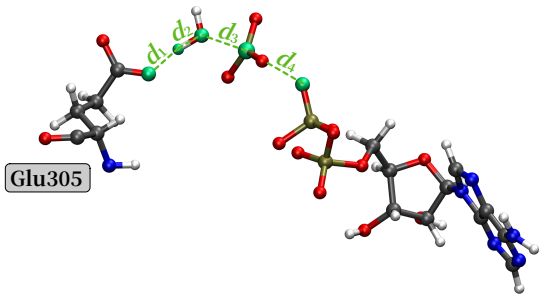


Figure S23: Enhanced sampling of the 1st step in ATP hydrolysis: bond distances employed to build the CV space in the WTM-eABF simulations.

As shown in Fig. S24 and S25, the process starts with the cleavage of the scissile $O_{3B}-P_\gamma$, followed by the nucleophilic attack of a water molecule, concerted with its deprotonation by Glu. We see that the distance to the path remains relatively constant over the duration of the sampling period, suggesting that the MEB and minimum free energy path (MFEP) are well aligned.

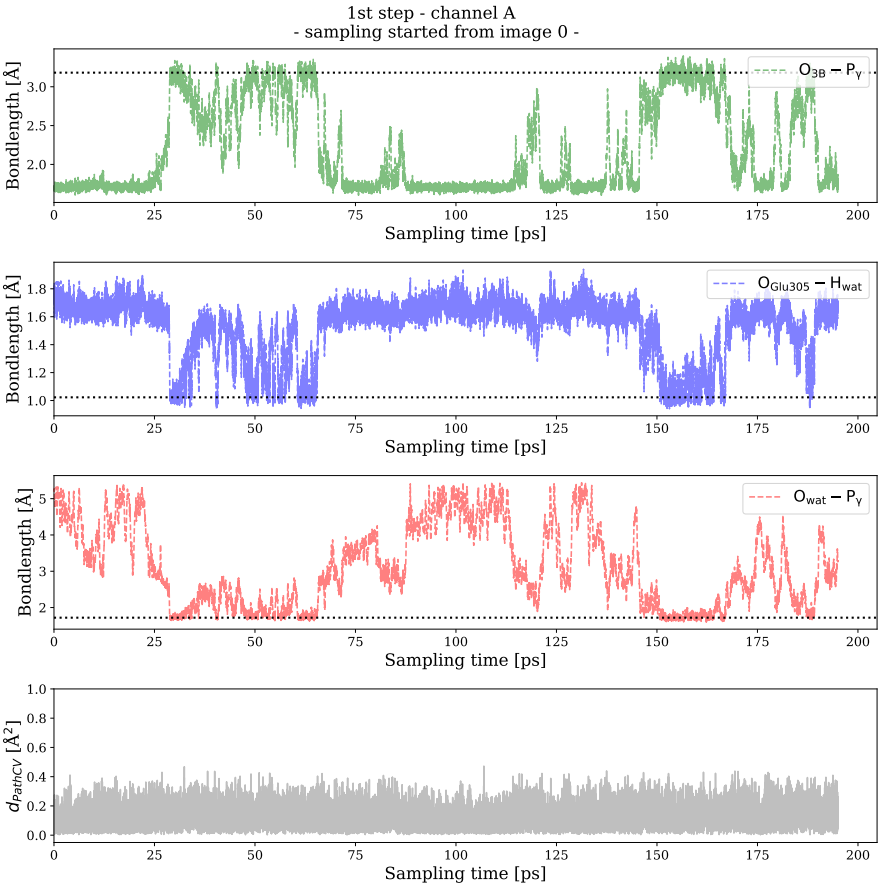


Figure S24: The evolution of key interatomic distances and the distance from the path CV along the sampling time. The WTM-eABF sampling is started from the educt structure (0^{th} NEB image), the dashed lines mark the distances measured in the intermediate structure (15^{th} NEB image), where the H^+ from the water molecule gets transferred to the Glu305, the O_{wat} - P_γ bond is formed and the O_{3B} - P_γ bond is broken.

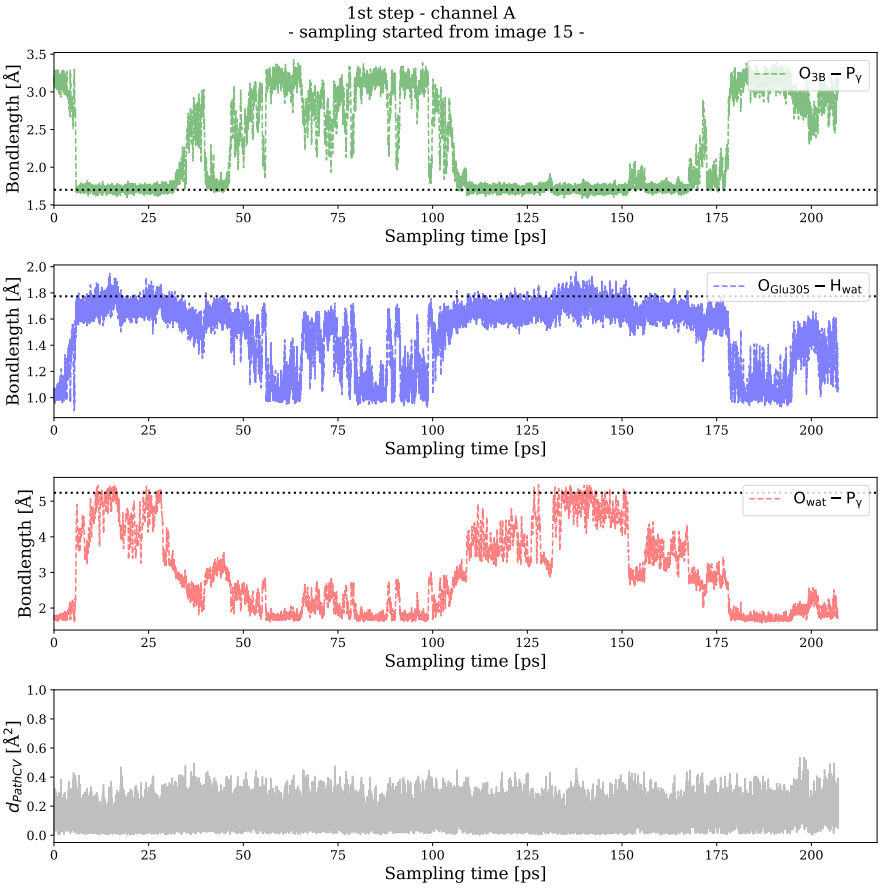


Figure S25: The evolution of key interatomic distances and the distance from the path CV along the sampling time. The WTM-eABF sampling is started from the intermediate structure (15$^{\text{th}}$ NEB image), the dashed lines mark the distances measured in the educt structure (0$^{\text{th}}$ NEB image), where the O_{3B} - P_γ bond forms again and the water goes away from the terminal phosphate.

8 The educt and the product state - H-bond networks at the binding site

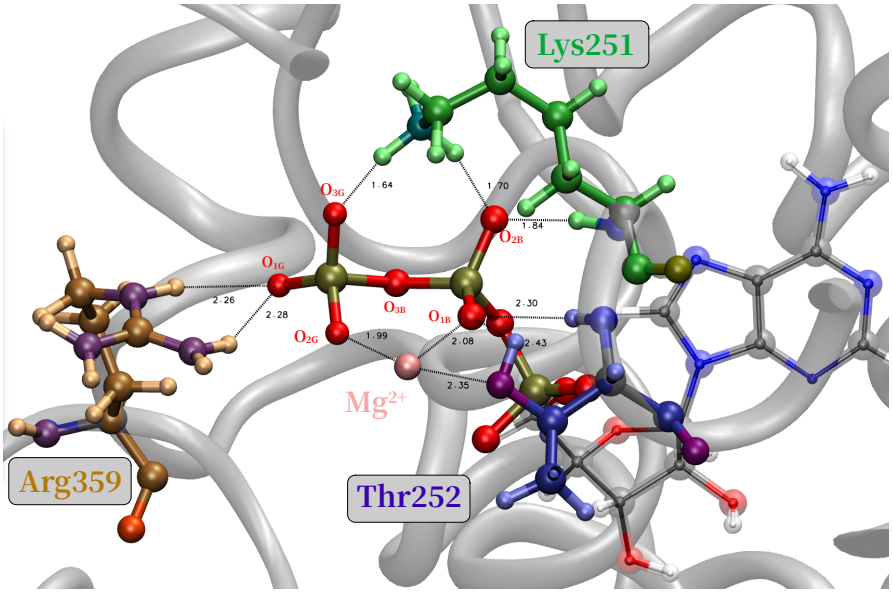


Figure S26: Educt state: H-bond network formed between the substrate, Arg359, Lys251, and Thr252 with the Mg^{2+} coordination sphere.

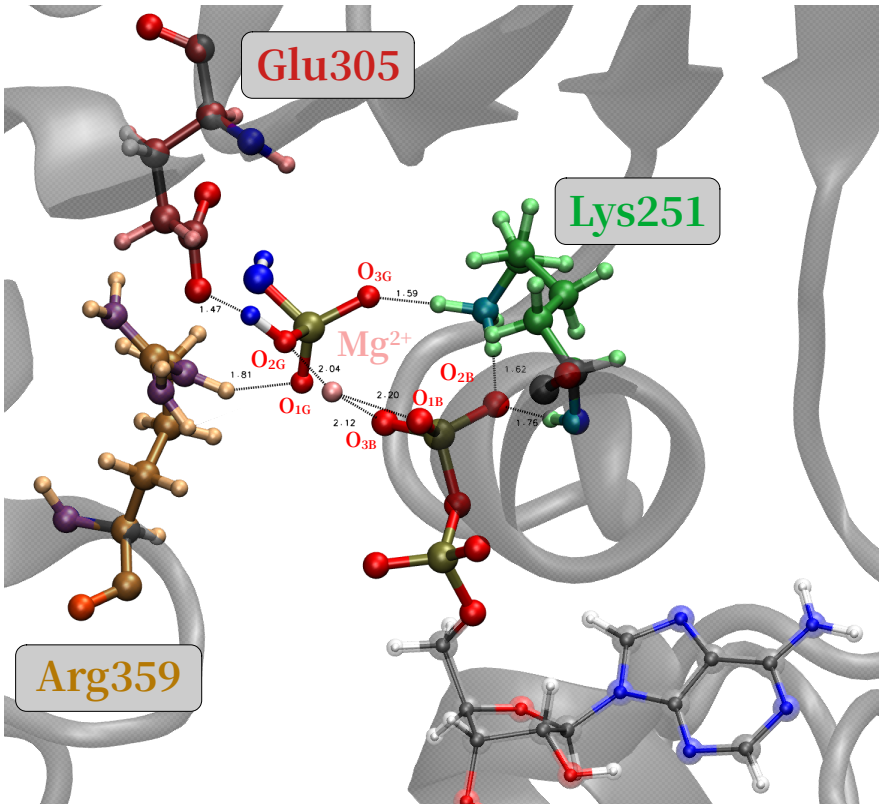


Figure S27: Product state: the Mg^{2+} coordination sphere bridges ADP^{3-} and $H_2PO_4^-$. H-bond network formed between the substrate and key amino acids: Glu305, Arg359, and Lys251.

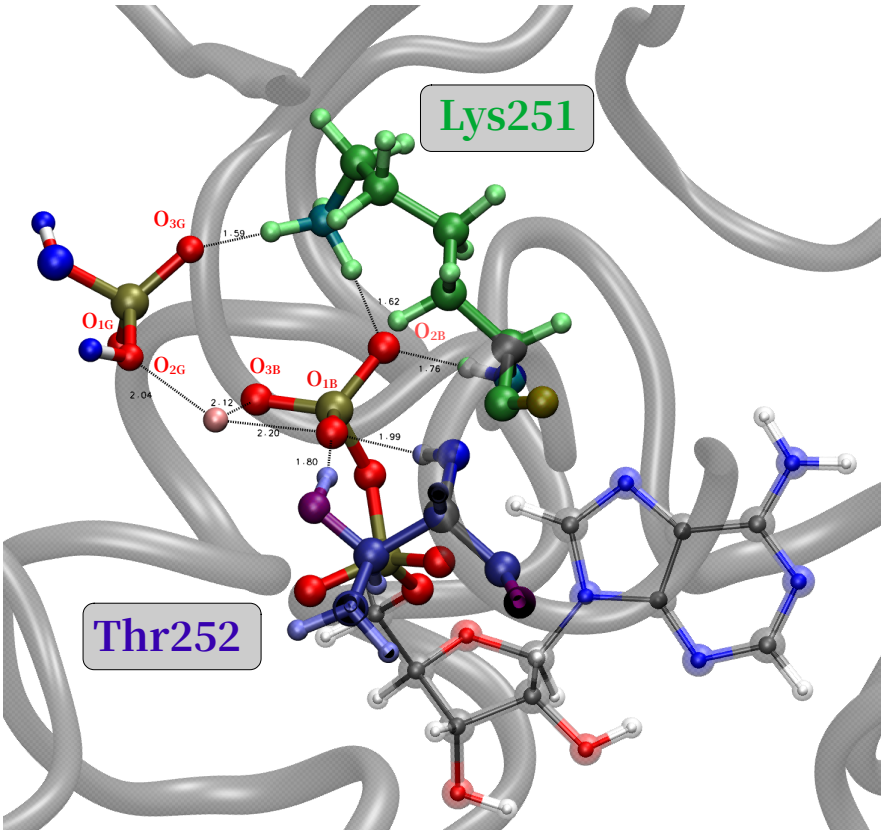


Figure S28: Product state: the Mg^{2+} coordination sphere bridges ADP^{3-} and $H_2PO_4^-$. H-bond network formed between the substrate, Thr252 and Lys251.

In the product state (see Fig. S27), the Mg^{2+} ion is tightly coordinated by oxygen atoms of ADP, P_i (O_{2G}, O_{3B}, O_{1B}) and three water molecules forming an octahedral arrangement.

The NEB pathway that leads to the product state captures two events: the O_{3B} atom enters the Mg^{2+} coordination shell and Thr252 leaves it.

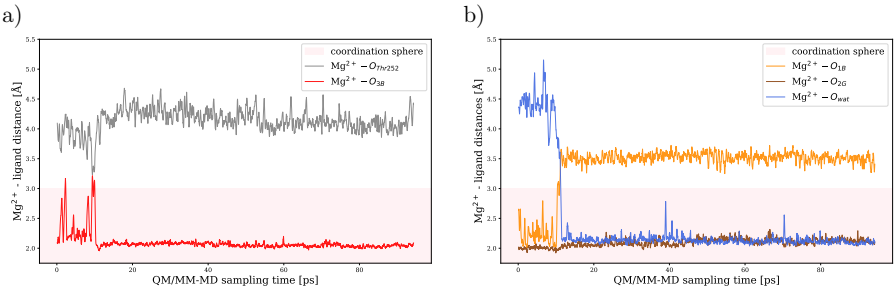


Figure S29: Mg^{2}+ coordination in the product state over 100 ps.

The 100 ps long unbiased QM/MM-MD simulation of the product state shows that the O_{3B} atom remains strongly coordinated to the Mg^{2+}, while Thr252 stays outside the coordination sphere. The 3-fold Mg^{2+} coordination to the ADP + P_i is stable for 10 ps, after which the O_{1B} atom leaves the coordination shell.

9 DFT NMR calculations

The chemical shift of the P_α atom barely changes as we transition from the educt to the product structure; therefore, this nucleus was used as a reference to convert absolute magnetic shieldings (σ) into chemical shifts (δ). All computed isotropic shieldings along the reaction path were shifted such that the P_α from the educt (0^{th} NEB image of the first step) matches the experimentally measured P_α shift of the ATP at the active site of p97[27].

$$\text{ref} = \sigma_{\text{calc. } (P_\alpha \text{ in ATP})} + \delta_{\text{exp. } (P_\alpha \text{ in ATP})}$$

10 Enhanced sampling using the WTM-eABF method - trajectories and histograms for the first reaction step

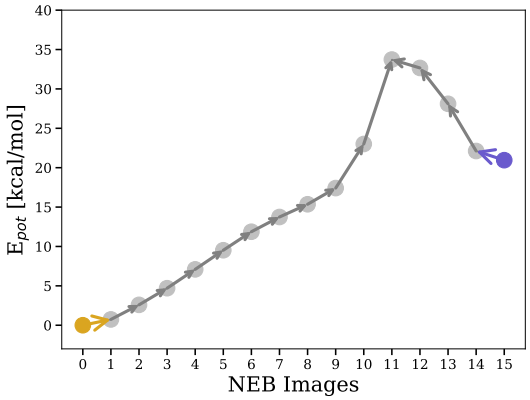


Figure S30: Minimum Energy Pathway from NEB optimizations for channel A. Arrows indicate QM/MM-MD trajectories initiated from NEB images, with bias applied to the path collective variable for sampling. The 0^{th} (gold) and the 15^{th} NEB image (purple) correspond to the educt and the ADP + HPO_4^{2-} intermediate structure.

Sampling data from trajectories started from the 5^{th} and 6^{th} NEB image (Fig. S28 and S29) were excluded from the PMF profile calculation, because these trajectories were trapped in the product state after the first transition.

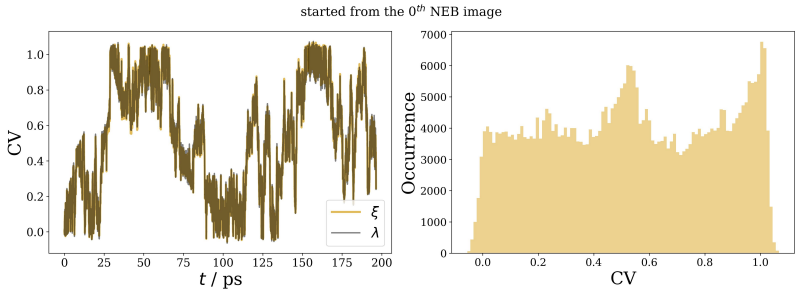


Figure S31: Trajectory and histogram of the path CV for QM/MM-MD sampling started from the NEB image 0.

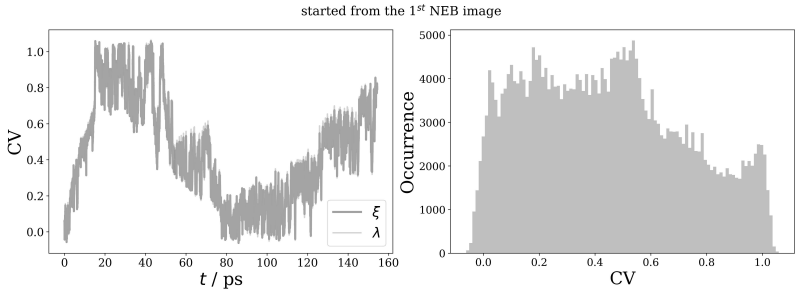


Figure S32: Trajectory and histogram of the path CV for QM/MM-MD sampling started from the NEB image 1.

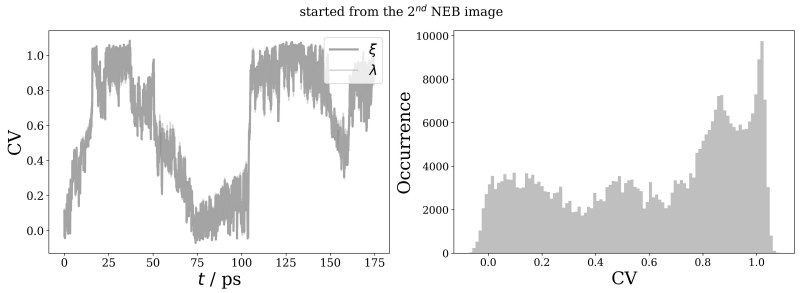


Figure S33: Trajectory and histogram of the path CV for QM/MM-MD sampling started from the NEB image 2.

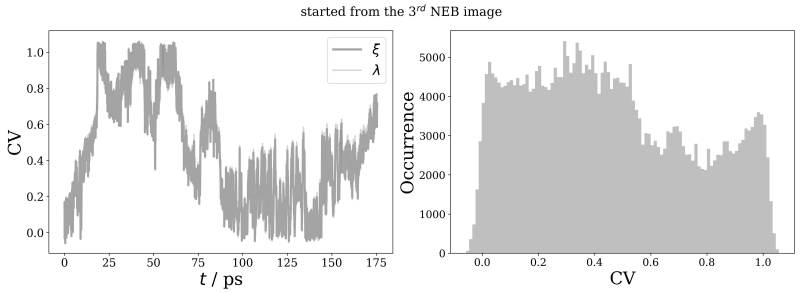


Figure S34: Trajectory and histogram of the path CV for QM/MM-MD sampling started from the NEB image 3.

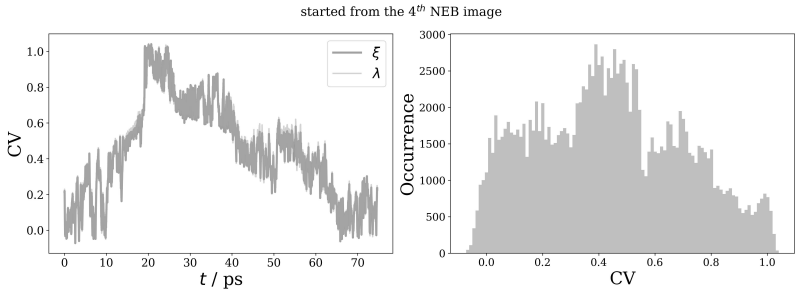


Figure S35: Trajectory and histogram of the path CV for QM/MM-MD sampling started from the NEB image 4.

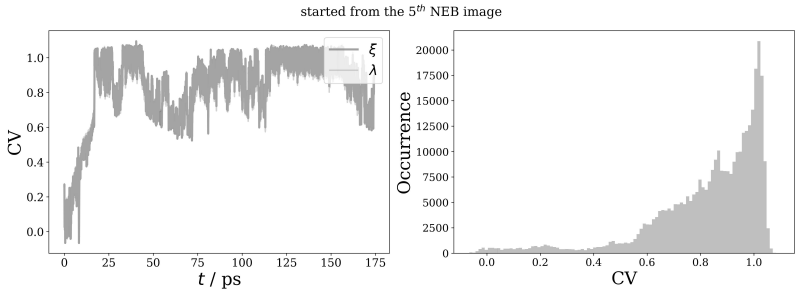


Figure S36: Trajectory and histogram of the path CV for QM/MM-MD sampling started from the NEB image 5.

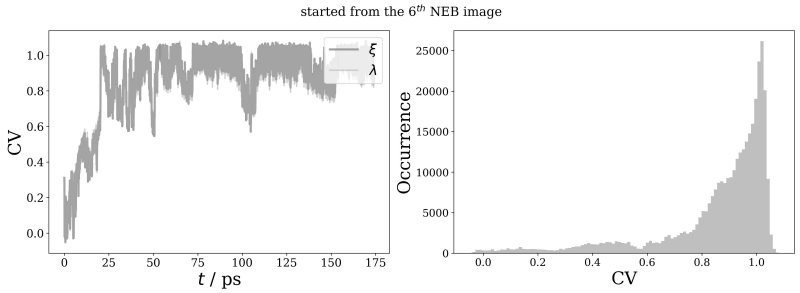


Figure S37: Trajectory and histogram of the path CV for QM/MM-MD sampling started from the NEB image 6.

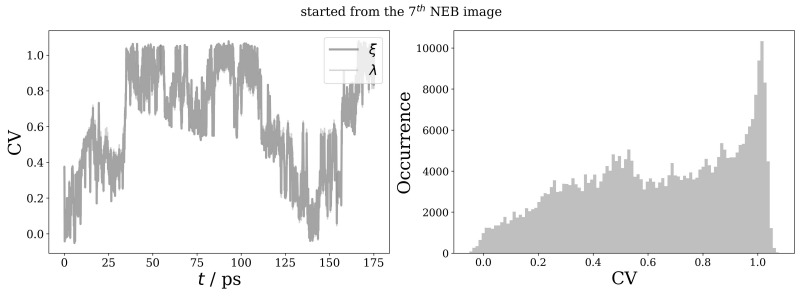


Figure S38: Trajectory and histogram of the path CV for QM/MM-MD sampling started from the NEB image 7.

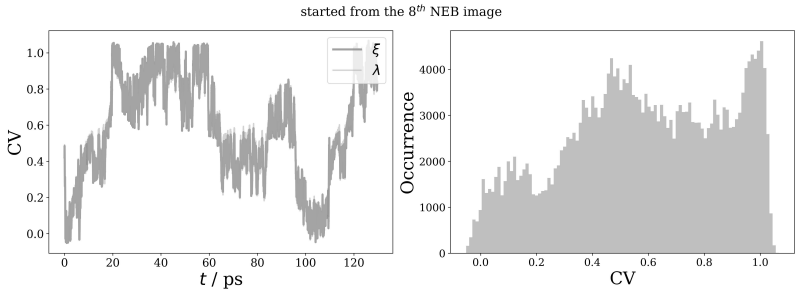


Figure S39: Trajectory and histogram of the path CV for QM/MM-MD sampling started from the NEB image 8.

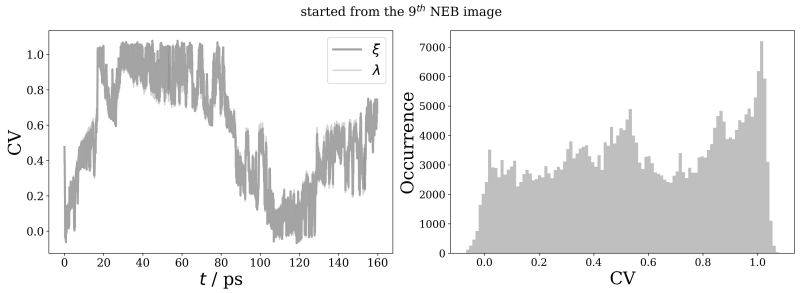


Figure S40: Trajectory and histogram of the path CV for QM/MM-MD sampling started from the NEB image 9.

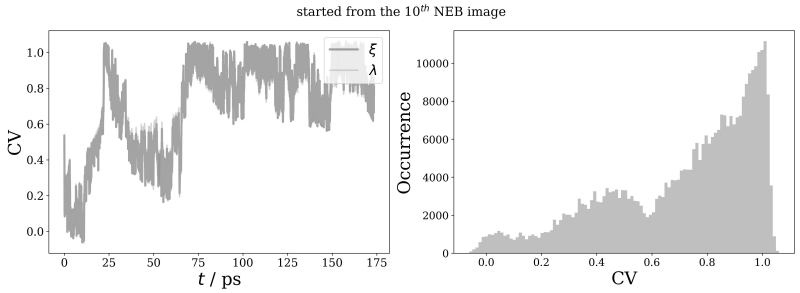


Figure S41: Trajectory and histogram of the path CV for QM/MM-MD sampling started from the NEB image 10.

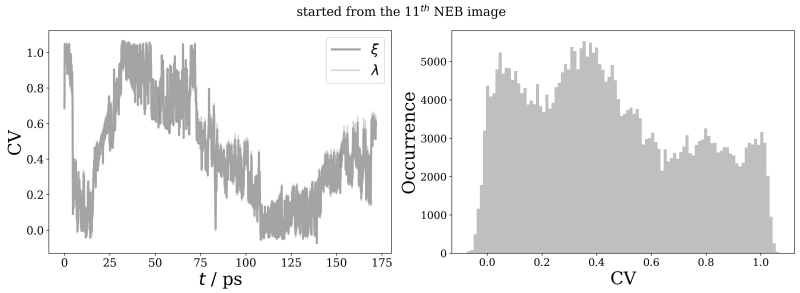


Figure S42: Trajectory and histogram of the path CV for QM/MM-MD sampling started from the NEB image 11.

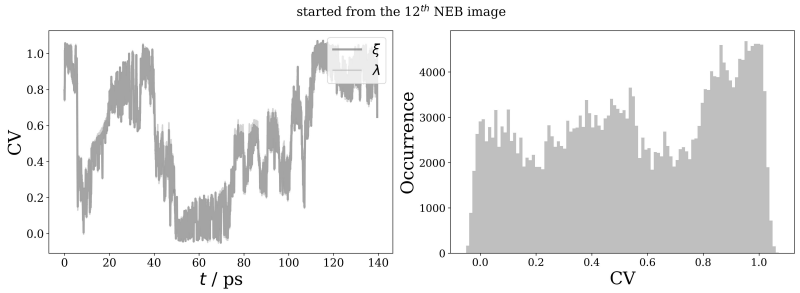


Figure S43: Trajectory and histogram of the path CV for QM/MM-MD sampling started from the NEB image 12.

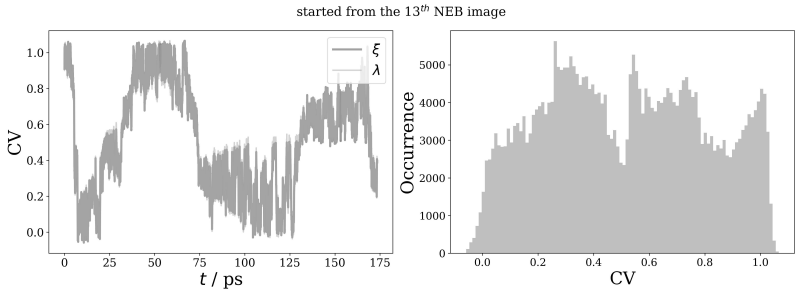


Figure S44: Trajectory and histogram of the path CV for QM/MM-MD sampling started from the NEB image 13.

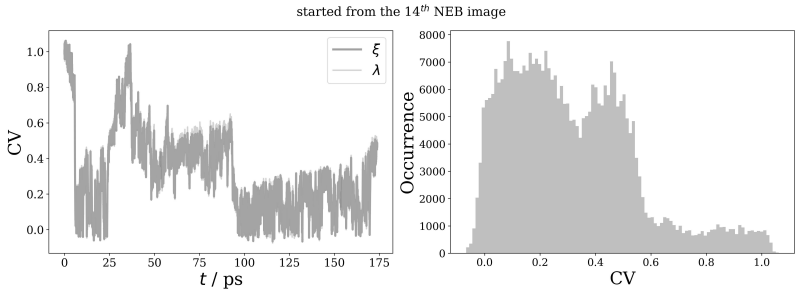


Figure S45: Trajectory and histogram of the path CV for QM/MM-MD sampling started from the NEB image 14.

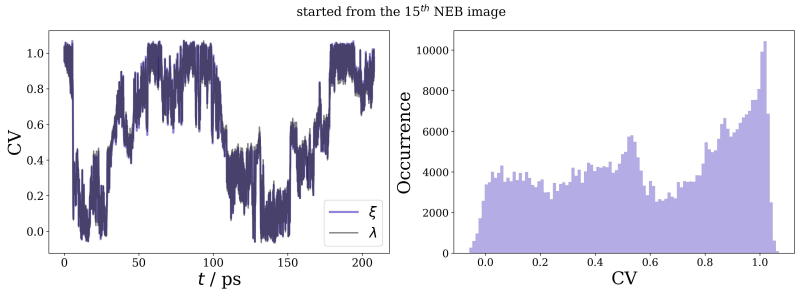


Figure S46: Trajectory and histogram of the path CV for QM/MM-MD sampling started from the NEB image 15.

11 Evaluation of the PMF profile uncertainties

Uncertainties associated with the computed free energy profiles are analyzed in the following ways. Firstly, the variation of the PMF profile as obtained from individual trajectories is computed (shaded region in Figure S47).

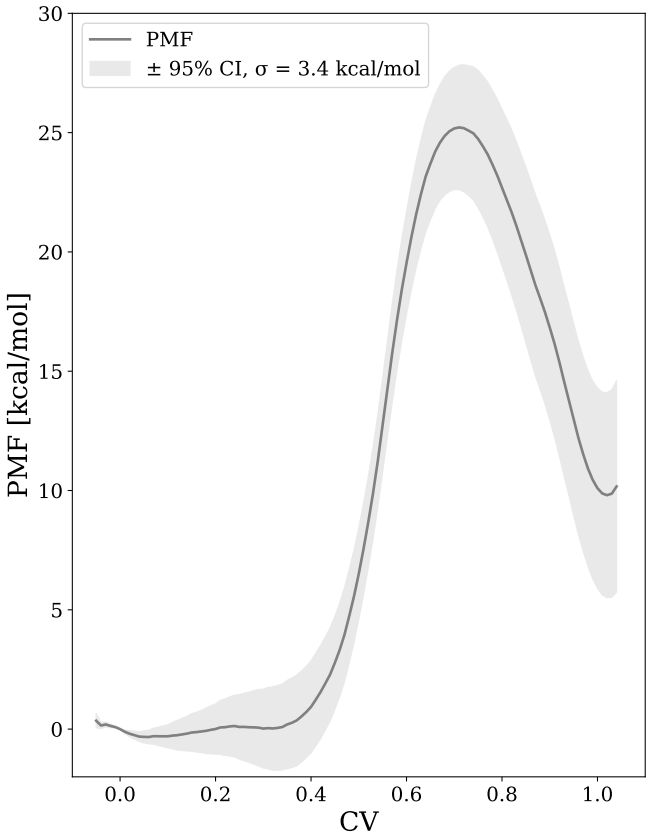


Figure S47: PMF profile uncertainty predicted from individual trajectories. The shaded region illustrates the 95% confidence interval.

Secondly, we use a subsample bootstrap approach to create $N_{sample} = 100$ datasets, each containing as many data points as a single trajectory (200 ps), and recalculate the PMF using the same MBAR protocol.

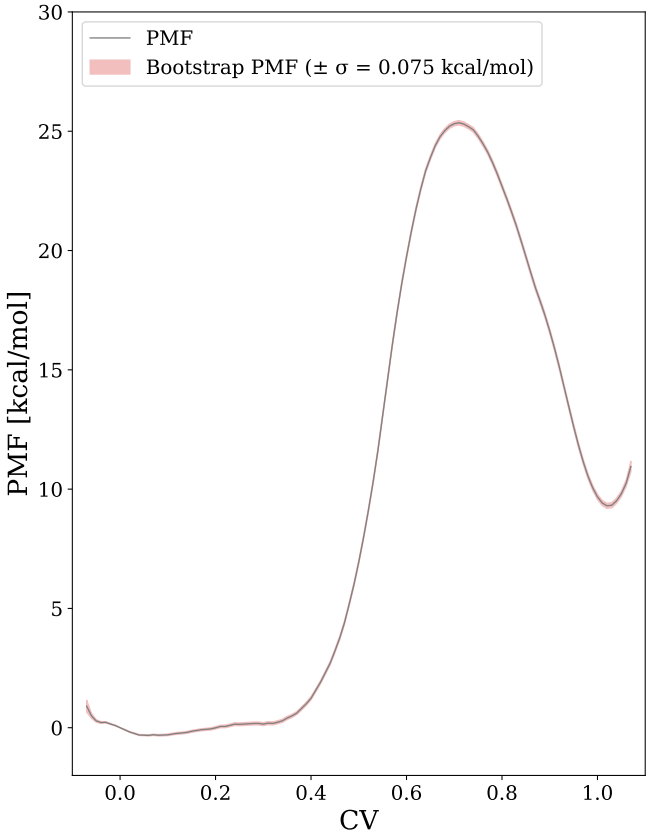


Figure S48: PMF computed from 100 bootstrap subsamples.

In this second approach, the uncertainty can be estimated as the standard deviation across the resulting ensemble of PMFs (Figure S48). The obtained error bars are vanishingly small, reflecting the statistical robustness of the PMF estimation from the global data.

References

(1) Kussmann, J.; Ochsenfeld, C. Pre-selective screening for matrix elements in linear-scaling exact exchange calculations. *J. Chem. Phys.* **2013**, *138*, 134114.

(2) Kussmann, J.; Ochsenfeld, C. Preselective screening for linear-scaling exact exchange-gradient calculations for graphics processing units and general strong-scaling massively parallel calculations. *J. Chem. Theory Comput.* **2015**, *11*, 918–922.

(3) Kussmann, J.; Ochsenfeld, C. Hybrid CPU/GPU integral engine for strong-scaling ab initio methods. *J. Chem. Theory Comput.* **2017**, *13*, 3153–3159.

(4) Grimme, S.; Brandenburg, J. G.; Bannwarth, C.; Hansen, A. Consistent structures and interactions by density functional theory with small atomic orbital basis sets. *J. Chem. Phys.* **2015**, *143*.

(5) Laqua, H.; Thompson, T. H.; Kussmann, J.; Ochsenfeld, C. Highly efficient, linear-scaling seminumerical exact-exchange method for graphic processing units. *J. Chem. Theory Comput.* **2020**, *16*, 1456–1468.

(6) Laqua, H.; Kussmann, J.; Ochsenfeld, C. Accelerating semi-numerical Fock-exchange calculations using mixed single-and double-precision arithmetic. *J. Chem. Phys.* **2021**, *154*, 214116.

(7) Laqua, H.; Dietschreit, J. C.; Kussmann, J.; Ochsenfeld, C. Accelerating Hybrid Density Functional Theory Molecular Dynamics Simulations by Seminumerical Integration, Resolution-of-the-Identity Approximation, and Graphics Processing Units. *J. Chem. Theory Comput.* **2022**, *18*, 6010–6020.

(8) Kussmann, J.; Laqua, H.; Ochsenfeld, C. Highly efficient resolution-of-identity density functional theory calculations on central and graphics processing units. *J. Chem. Theory Comput.* **2021**, *17*, 1512–1521.

(9) Shein, M.; Hitzenberger, M.; Cheng, T. C.; Rout, S. R.; Leitl, K. D.; Sato, Y.; Zacharias, M.; Sakata, E.; Schütz, A. K. Characterizing ATP processing by the AAA+ protein p97 at the atomic level. *Nat. Chem.* **2024**, *16*, 363–372.

(10) Szántó, J. K.; Dietschreit, J. C.; Shein, M.; Schütz, A. K.; Ochsenfeld, C. Systematic QM/MM Study for Predicting 31P NMR Chemical Shifts of Adenosine Nucleotides in Solution and Stages of ATP Hydrolysis in a Protein Environment. *J. Chem. Theory Comput.* **2024**, *20*, 2433–2444.

(11) Kästner, J.; Carr, J. M.; Keal, T. W.; Thiel, W.; Wander, A.; Sherwood, P. DL-FIND: An open-source geometry optimizer for atomistic simulations. *J. Phys. Chem. A* **2009**, *113*, 11856–11865.

(12) Lu, Y.; Farrow, M. R.; Fayon, P.; Logsdail, A. J.; Sokol, A. A.; Catlow, C. R. A.; Sherwood, P.; Keal, T. W. Open-Source, python-based redevelopment of the ChemShell multiscale QM/MM environment. *J. Chem. Theory Comput.* **2018**, *15*, 1317–1328.

(13) Hulm, A.; Lemke, Y.; Johannes, D.; Glinkina, L.; Stan-Bernhardt, A. adaptive_sampling, https://github.com/ochsenfeld-lab/adaptive_sampling.

(14) Tang, W. K.; Xia, D. Altered intersubunit communication is the molecular basis for functional defects of pathogenic p97 mutants. *J. Biol. Chem.* **2013**, *288*, 36624–36635.

(15) Ranaghan, K. E.; Mulholland, A. J. Investigations of enzyme-catalysed reactions with combined quantum mechanics/molecular mechanics (QM/MM) methods. *Int. Rev. Phys. Chem.* **2010**, *29*, 65–133.

(16) Lonsdale, R.; Harvey, J. N.; Mulholland, A. J. A practical guide to modelling enzyme-catalysed reactions. *Chem. Soc. Rev.* **2012**, *41*, 3025–3038.

(17) Henkelman, G.; Jónsson, H. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *J. Chem. Phys.* **2000**, *113*, 9978–9985.

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

(19) Fu, H.; Zhang, H.; Chen, H.; Shao, X.; Chipot, C.; Cai, W. Zooming across the free-energy landscape: shaving barriers, and flooding valleys. *J. Phys. Chem. Lett.* **2018**, *9*, 4738–4745.

(20) Fu, H.; Shao, X.; Cai, W.; Chipot, C. Taming rugged free energy landscapes using an average force. *Acc. Chem. Res.* **2019**, *52*, 3254–3264.

(21) Hulm, A.; Ochsenfeld, C. Improved Sampling of Adaptive Path Collective Variables by Stabilized Extended-System Dynamics. *J. Chem. Theory Comput.* **2023**, *19*, 9202–9210.

(22) Neese, F. Software update: The ORCA program system—Version 5.0. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2022**, *12*, e1606.

(23) Müller, M.; Hansen, A.; Grimme, S. ωB97X-3c: A composite range-separated hybrid DFT method with a molecule-optimized polarized valence double-ζ basis set. *J. Chem. Phys.* **2023**, *158*.

(24) Riplinger, C.; Sandhoefer, B.; Hansen, A.; Neese, F. Natural triple excitations in local coupled cluster calculations with pair natural orbitals. *J. Chem. Phys.* **2013**, *139*.

(25) Shirts, M. R.; Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* **2008**, *129*, 124105.

(26) Hulm, A.; Dietschreit, J. C.; Ochsenfeld, C. Statistically optimal analysis of the extended-system adaptive biasing force (eABF) method. *J. Chem. Phys.* **2022**, *157*.

(27) Rydzek, S.; Shein, M.; Bielytskyi, P.; Schütz, A. K. Observation of a transient reaction intermediate illuminates the mechanochemical cycle of the AAA-ATPase p97. *J. Am. Chem. Soc.* **2020**, *142*, 14472–14480.

3.4 Publication IV: QM/MM Free Energy Calculations of Long-Range Biological Protonation Dynamics by Adaptive and Focused Sampling

Abstract: Water-mediated proton transfer reactions are central for catalytic processes in a wide range of biochemical systems, ranging from biological energy conversion to chemical transformations in the metabolism. Yet, the accurate computational treatment of such complex biochemical reactions is highly challenging and requires the application of multiscale methods, in particular hybrid quantum/classical (QM/MM) approaches combined with free energy simulations. Here, we combine the unique exploration power of new advanced sampling methods with density functional theory (DFT)-based QM/MM free energy methods for multiscale simulations of long-range protonation dynamics in biological systems. In this regard, we show that combining multiple walkers/well-tempered metadynamics with an extended system adaptive biasing force method (MWE) provides a powerful approach for exploration of water-mediated proton transfer reactions in complex biochemical systems. We compare and combine the MWE method also with QM/MM umbrella sampling and explore the sampling of the free energy landscape with both geometric (linear combination of proton transfer distances) and physical (center of excess charge) reaction coordinates and show how these affect the convergence of the potential of mean force (PMF) and the activation free energy. We find that the QM/MM-MWE method can efficiently explore both direct and water-mediated proton transfer pathways together with forward and reverse hole transfer mechanisms in the highly complex proton channel of respiratory Complex I, while the QM/MM-US approach shows a systematic convergence of selected long-range proton transfer pathways. In this regard, we show that the PMF along multiple proton transfer pathways is recovered by combining the strengths of both approaches in a QM/MM-MWE/focused US (FUS) scheme and reveals new mechanistic insight into the proton transfer principles of Complex I. Our findings provide a promising basis for the quantitative multiscale simulations of long-range proton transfer reactions in biological systems.

Reprinted with permission from

M. C. Pöverlein, A. Hulm, J. C. B. Dietschreit, J. Kussmann, C. Ochsenfeld, V. R. I. Kaila. "QM/MM Free Energy Calculations of Long-Range Biological Protonation Dynamics by Adaptive and Focused Sampling." *J. Chem. Theory Comput.* **2024**, 20, 5751–5762.
URL: `https://doi.org/10.1021/acs.jctc.4c00199`.

JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

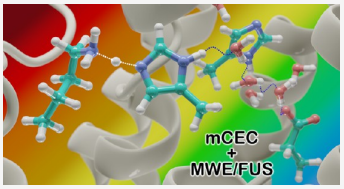
QM/MM Free Energy Calculations of Long-Range Biological Protonation Dynamics by Adaptive and Focused Sampling

Maximilian C. Pöverlein, Andreas Hulm, Johannes C. B. Dietschreit, Jörg Kussmann, Christian Ochsenfeld, and Ville R. I. Kaila*

Cite This: J. Chem. Theory Comput. 2024, 20, 5751−5762 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Water-mediated proton transfer reactions are central for catalytic processes in a wide range of biochemical systems, ranging from biological energy conversion to chemical transformations in the metabolism. Yet, the accurate computational treatment of such complex biochemical reactions is highly challenging and requires the application of multiscale methods, in particular hybrid quantum/classical (QM/MM) approaches combined with free energy simulations. Here, we combine the unique exploration power of new advanced sampling methods with density functional theory (DFT)-based QM/MM free energy methods for multiscale simulations of long-range protonation dynamics in biological systems. In this regard, we show that combining multiple walkers/well-tempered metadynamics with an extended system adaptive biasing force method (MWE) provides a powerful approach for exploration of water-mediated proton transfer reactions in complex biochemical systems. We compare and combine the MWE method also with QM/MM umbrella sampling and explore the sampling of the free energy landscape with both geometric (linear combination of proton transfer distances) and physical (center of excess charge) reaction coordinates and show how these affect the convergence of the potential of mean force (PMF) and the activation free energy. We find that the QM/MM-MWE method can efficiently explore both direct and water-mediated proton transfer pathways together with forward and reverse hole transfer mechanisms in the highly complex proton channel of respiratory Complex I, while the QM/MM-US approach shows a systematic convergence of selected long-range proton transfer pathways. In this regard, we show that the PMF along multiple proton transfer pathways is recovered by combining the strengths of both approaches in a QM/MM-MWE/focused US (FUS) scheme and reveals new mechanistic insight into the proton transfer principles of Complex I. Our findings provide a promising basis for the quantitative multiscale simulations of long-range proton transfer reactions in biological systems.



INTRODUCTION

Proton transfer reactions are essential for many biological processes, ranging from catalytic transformations in the metabolism to cellular respiration and photosynthesis, which are responsible for biological energy conversion.[1,2] Biological proton transfer reactions are often catalyzed by titratable amino acids (His, Lys, Asp, Glu) buried within the protein core that together with water molecules form "proton wires" that facilitate proton transfer via bond-rearrangement in a Grotthuss-type transfer reaction.[3] Respiratory and photosynthetic enzymes employ such water-mediated proton transfer reactions to create a proton motive force across a biological membrane, powering the synthesis of adenosine triphosphate (ATP) and active transport in cells.[4] Yet, despite major advances in understanding these complex biological systems, the mechanistic principles of several bioenergetic enzymes remain unclear and highly debated. In this regard, multiscale simulations provide a key understanding of how the protonation reactions are controlled by the protein structure and dynamics.[1] The mechanistic principles of different proton transfer reactions in several bioenergetic systems have indeed been addressed by various multiscale methods in recent years.[5−8] Yet, long-range biological proton transfer reactions that can extend across large distances still pose major challenges for modern multiscale methods due to a high computational cost, challenges in converging free energy profiles of different reaction pathways, together with the need for the accurate treatment of the electronic structure of the system.

The accurate computational treatment of protonation dynamics requires modeling of both bond-breaking and bond-formation reactions, as well as the conformational dynamics of the system. To lower the computational cost, semi-empirical approaches (e.g., DFTB, SCC-DFTB, PM7) as well as reactive force field methods (e.g., EVB, MS-EVB) have been developed. Although these methods can provide valuable insight into various biological reactions,[8−15] they require pre-

Received: February 16, 2024
Revised: April 16, 2024
Accepted: April 24, 2024
Published: May 8, 2024

https://doi.org/10.1021/acs.jctc.4c00199
J. Chem. Theory Comput. 2024, 20, 5751−5762

parametrization that can be difficult to achieve.[16] In this regard, density functional theory (DFT)-based methods often provide a good compromise between the computational cost and accuracy for many complex systems,[17,18] together with hybrid quantum mechanics/classical mechanics (QM/MM) methods that allow for modeling complex interactions of the biological surroundings at an approximate atomistic force field level.[1,19] Since the protonation reactions can involve significant charge rearrangements that lead to conformational changes in the protein surrounding, they must be sampled by explicit molecular dynamics or enhanced sampling methods on time scales that are often challenging to capture by first principles methods due to the high computational cost.

Recent developments in electronic structure methods, e.g., prescreening of integrals and various linear scaling approaches,[20] can provide a significant reduction of the computational costs, thus allowing modeling of large extended QM systems. Together with new hardware and implementations, such as the accelerations provided by graphics processing units (GPU),[21] they provide new opportunities for the exploration of longer time scales and rare events.

Here, we introduce, implement, and benchmark QM/MM enhanced sampling methods in combination with different reaction coordinates for simulations of long-range proton transfer reactions. More specifically, we study the recent shared-bias well-tempered metadynamics extended adaptive biasing force (MWE) method,[22] and both compare and combine this with umbrella sampling (US) in the context of multiscale QM/MM simulations. We study the performance of geometric [linear combination (LC) of bond-breaking and bond-formation process] and physically motivated [modified center-of-excess charge, mCEC[23]] representations of the transferred proton as reaction coordinates/collective variables (CVs). The former CV is defined *a priori* by manual selection of the involved bond distances, while the global nature of the latter aims for a unified description of all possible proton transfers, which may open up the exploration of new mechanisms on-the-fly.

We show how the unconfined diffusion along CVs in the MWE approach can accelerate the conformational sampling and exploration of various reaction mechanisms relative to the US approach, which by construction confines the simulations to a single reaction channel. However, we also discuss challenges in obtaining converged free energy profiles using MWE due to the sampling of a larger conformational space, which may require combination of multiple simulations and manual tuning of the sampling parameters.

To address the convergence of such challenging potentials of mean force (PMFs), we combine the QM/MM-MWE and QM/MM-US simulation methods to systematically recover the multidimensional PMF along different reaction mechanisms. In the following, after a short theoretical review of the US and MWE methods, we discuss the performance of both sampling strategies on a model system. Finally, we apply the presented framework to the protonation dynamics of respiratory Complex I, a highly challenging biological system, where the different sampling strategies are explored.

THEORY, METHODS, AND MODELS

The potential of mean force (PMF) along a reaction coordinate or collective variable (CV) is defined as

$$A(z) = -k_B T \ln \rho(z) \tag{1}$$

with the Boltzmann constant k_B, the temperature T, and the probability density function $\rho(z)$ of finding the system in a certain state z along the CV $\xi(\boldsymbol{x})$, defined by

$$\rho(z) = \int \delta[\xi(\boldsymbol{x}) - z]\rho(\boldsymbol{x})\mathrm{d}\mathbf{x} \tag{2}$$

where δ denotes the Dirac delta function. The efficient estimation of $\rho(z)$ often requires application of importance sampling strategies.[24−27] The US simulations[27] aim to achieve a uniform exploration of the PMF by performing several equilibrium simulations with predefined restraints $B(\xi(\boldsymbol{x}))$, often modeled as harmonic biasing potentials. Similarly, in the MWE method, the CVs are coupled to harmonic potentials $B(\xi(\boldsymbol{x}),\lambda)$, while the diffusion of the simulation windows along the CV is achieved by the coupling to a fictitious particle λ.[28] The full potential energy of the extended system $(\boldsymbol{x},\lambda)$ can be defined as

$$U(\boldsymbol{x}, \lambda, t) = U_0(\boldsymbol{x}) + B(\xi(\boldsymbol{x}), \lambda) + B_{\mathrm{MWE}}(\lambda, t) \tag{3}$$

where $U_0(\boldsymbol{x})$ is the potential energy of the system. To ensure uniform sampling of the CV along λ, a time-dependent bias potential $B_{\mathrm{MWE}}(\lambda,t)$ can be added. In this regard, the MWE uses two complementary strategies, simultaneously filling the free energy wells (well-tempered metadynamics) and removing the barriers (extended adaptive biasing force). Postprocessing allows recovering the restrained simulation windows $B_i(\xi(\boldsymbol{x}))$ analogous to US windows by separation of the extended system into states with constant $\lambda = \lambda_i$.[29]

As the biasing potentials are known, the unbiased probabilities are obtained via

$$A_i(z) = -k_B T \ln \rho_i^b(z) - B_i(z) + F_i \tag{4}$$

where the integration constants, F_i, for each window, i describe the vertical position of the individual unbiased PMF, while ρ_i^b are the probability distributions obtained from the biased runs. In practice, the unbiased PMF is obtained by the weighted-histogram analysis method (WHAM)[30] or by the multistate Bennett's acceptance ratio (MBAR),[31] a histogram free/zero bin width version of WHAM. The unbiased probabilities in MBAR are obtained by

$$F_i = -\beta^{-1} \ln \sum_{n=1}^{N} \frac{e^{-\beta B_i(\xi(\boldsymbol{x}_n))}}{\sum_{k=1}^{K} N_k e^{-\beta(B_k(\xi(\boldsymbol{x}_n)) - F_k)}} \tag{5}$$

with K simulation windows comprising N_k samples and data points from the pool of all simulations $\boldsymbol{x}_n$. The unbiased weights of the individual frames can thus be recovered as

$$p(\boldsymbol{x}_n) = \frac{\mathcal{N}}{\sum_{k=1}^{K} N_k e^{-\beta(B_k(\xi(\boldsymbol{x}_n)) - F_k)}} \tag{6}$$

where the normalization constant $\mathcal{N}$ is introduced to ensure that $\sum_n p(\boldsymbol{x}_n) = 1$. Recent developments, such as the transition-based reweighting (TRAM,[32] dTRAM[33]) and the dynamic histogram analysis extended to detailed balance (DHAM,[34] DHAMed[35]), which account for the transitions between the restrained simulation windows, can also be used to obtain the free energy estimate F_i as a function of the CV.

Accurate activation free energies, which define the reaction rates, can be determined from the PMF by using a definition yielding consistent energies for CVs with parallel gradients,[36]

https://doi.org/10.1021/acs.jctc.4c00199
J. Chem. Theory Comput. 2024, 20, 5751−5762

Journal of Chemical Theory and Computation pubs.acs.org/JCTC Article

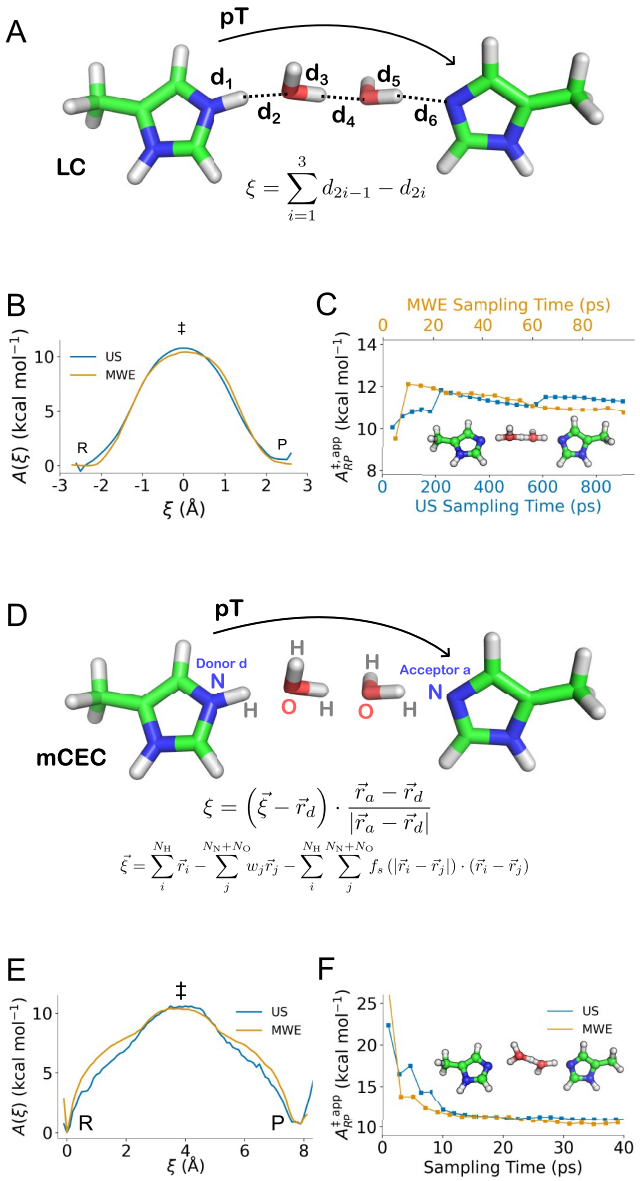


Figure 1. Sampling of water-mediated proton transfer reaction in a His-water model system with QM/MM umbrella sampling (US) and the multiple walker/well-tempered metadynamics/extended system adaptive biasing force (MWE) method. The PMF was explored using a linear combination (LC, panel A) of geometric distance and the modified center of excess charge (mCEC, panel D) as reaction coordinates. (A, D) Definition of the reaction coordinates. (B, E) The PMFs obtained using the LC and mCEC definitions with US and MWE. R, reactant, P, product, and ‡ transition state region. (C, F) Convergence of the max-min difference with increasing sampling time. Representative conformations sampled in the transition state region are shown as an inset.

$$\Delta A^{\ddagger}_{RP} = -k_B T \ln \frac{\rho(z^{\ddagger})\langle\lambda_{\xi}\rangle_{z^{\ddagger}}}{P(R)} \tag{7}$$

where $\rho(z^{\ddagger})$ is the probability density at the dividing surface, $P(R)$ is the probability that the system resides in the reactant state (obtained by integration over the PMF), and $\langle\lambda_{\xi}\rangle_z$ is the conditional average of the de Broglie wavelength[36] related to the mass of the pseudoparticle associated with the CV, m_{ξ} (see

https://doi.org/10.1021/acs.jctc.4c00199
J. Chem. Theory Comput. 2024, 20, 5751−5762

Extended Methods). A comparison of this expression with the often-employed harmonic approximation is given in the SI. Equation 7 removes possible distortions of the Cartesian space by nonlinear CVs and accounts for the mass of the atoms involved in the transition, while apparent free energy barriers obtained from difference between maxima and minima on the PMF ($\Delta A_{RP}^{\ddagger,app} = A^{\ddagger} - A_R$) do not include such corrections. The former treatment correctly reproduces, e.g., isotope effects.[36]

Computational Methods. All QM regions were described at the B3LYP-D3/def2-SVP level of theory,[37−40] which has shown a good balance between computational cost and accuracy for many biological systems (see Figure S15, and ref 41 for detailed benchmarking).[6,42−47] The currently employed density functional level is likely to underestimate barriers by a few kcal mol^{-1} relative to *ab initio* theory (Figure S15, see also refs 48,49). The surroundings of Model 1 were described with COSMO at $\varepsilon = 4$,[50] while those of Model 2 were described explicitly (see below for details on Model 1/2). MD simulations were performed at $T = 310$ K with a time step of 0.5 fs. PMFs were calculated using MBAR,[31,51] as described in ref 29. For the QM/MM sampling, activation free energies ($\Delta A_{RP}^{\ddagger}$) were obtained based on the z-averaged inverse mass of the reaction coordinate.[36] All simulations were implemented in Python and performed with the GPU-accelerated QM code FermiONs++[20,21,52−54] coupled to OpenMM[55] via a QM/MM interface. Significant speed up of QM(/MM)-MD calculations on GPUs was achieved by accelerating evaluations of exact exchange with the sn-LinK method[53,56] and by using the RI-J approximation of the Coulomb energy.[57] To accelerate the sampling, we further applied the multiple walkers (MW)/shared-bias approach, where parallel simulations synchronize the time-dependent bias potential $U_{MWE}(\lambda,t)$ in regular time intervals. Initial system setup and minimization of the reaction pathways were performed using TURBOMOLE v. 7.4.[58,59] Visual molecular dynamics (VMD)[60] and PyMOL[61] were used for visualization.

Model 1: Model System for Water-Mediated Proton Transfer. A system containing two histidine residues (modeled as methyl imidazole), hydrogen-bonded by two water molecules, was employed as a model system to study the free energy profile of proton transfer reactions. To this end, we studied the performance of two different reaction coordinates (Figure 1A,D): (I) a geometric reaction coordinate, defined as a linear combination (LC) of bond-breaking and bond-formation distances (Figure 1A).

$$\xi = \sum_{i=1}^{3} d_{2i-1} - d_{2i} \tag{8}$$

and (II) a modified center of excess charge (mCEC, see also Figure 1D)

$$\vec{\xi} = \sum_{i}^{N_H} \vec{r}_i - \sum_{j}^{N_O+N_N} w_j \cdot \vec{r}_j - \sum_{i}^{N_H} \sum_{j}^{N_O+N_N} f_s(|\vec{r}_i - \vec{r}_j|) \cdot (\vec{r}_i - \vec{r}_j) \tag{9}$$

with the projection onto the donor−acceptor vector

$$\xi = (\vec{\xi} - \vec{r}_d) \cdot \frac{\vec{r}_a - \vec{r}_d}{|\vec{r}_a - \vec{r}_d|} \tag{10}$$

The delta nitrogen ($N\delta$) of the first histidine was defined as the proton donor (d), while the $N\delta$ of the second histidine residue was defined as the proton acceptor (a) (Figure 1D). N_H denotes the number of involved exchangeable protons, and N_O and N_N are the number of oxygen and nitrogen proton acceptors. The weights, w_j, were set to $w_N = 0$ and $w_O = 2$, which describe the minimum number of protons bound to atom j in the reactant or product state. The modification term $f_s(d_{ab})$ introduces a logistic function to provide a smooth switching between the cross-terms and is defined as

$$f_s(d_{ab}) = \left(1 + \exp\left\{\frac{d_{ab} - r_s}{d_s}\right\}\right)^{-1} \tag{11}$$

where d_{ab} is the distance between a heavy atom a and proton b, both participating in the reaction coordinate. The parameters r_s and d_s, define the location and width of the switching regime and were set to 1.3 and 0.05 Å, respectively, unless stated otherwise (Figure S6).

Initial coordinates were obtained by geometry optimization at the B3LYP-D3/def2-SVP level,[37−40] while keeping the distance between the C_β atoms of the histidines fixed to 15 Å. The US simulations were performed using 18 windows sampled for 50 ps with a harmonic bias [$B_i = 1/2\ k\ (z - z_i)^2$] centered at LC coordinates [−2.07, +2.18 Å] every 0.25 Å with a force constant of $k = 100$ kcal mol^{-1} Å^{-2}. The MWE simulations were carried out using two walkers, each sampled for 48 ps. The extended variable was coupled to the system with a coupling width $\sigma = 0.1$ Å, and adaptive forces were accumulated on a grid with a bin width of $\Delta = 0.05$ Å. Well-tempered metadynamics was applied to the extended system with a Gaussian hill width of 0.1 Å and a hill height of 0.12 kcal mol^{-1} (0.5 kJ mol^{-1}) added every 10th fs. The bias factor $\gamma = \Delta T/(T + \Delta T)$, which ensures a smooth hill decay, was set to 0.866.

To explore the mCEC coordinate, US simulations were performed with 33 equally spaced windows, each simulated for 20 ps with a harmonic bias placed on the mCEC reaction coordinate, between [0, 8.75 Å] every 0.25 Å, and a force constant of $k = 100$ kcal mol^{-1} Å^{-2}. MWE simulations along the mCEC were performed with two walkers, each sampled for 50 ps. To this end, the extended variable was coupled to the system with $\sigma = 0.1$ Å and adaptive forces were accumulated on a grid with $\Delta = 0.1$ Å. In the MWE, well-tempered metadynamics was applied to the extended system with a Gaussian hill width of 0.2 Å, a hill height of 0.239 kcal mol^{-1} (1 kJ mol^{-1}), and a frequency of hill creation 10 fs, using a bias factor of $\gamma = 0.866$.

Additional convergence tests of the MWE sampling were performed using 10 walkers each sampled for 50 ps at the semi-empirical GFN2-xTB level.[62,63]

Model 2: Water-Mediated Proton Transfer Reactions in Complex I. The proton channel in Complex I was modeled based on the cryoEM structure of the mouse enzyme (PDB ID: 6ZTQ[64]) that was embedded in a POPC/POPE/cardiolipin (2:2:1) membrane and solvated with water molecules and NaCl (150 mM). The system was relaxed over 1 μs with classical MD simulations, followed by construction of the QM/MM model system. The QM region, comprising 114 atoms, included Lys237, His319, His293, Glu378, Ser289, Ser290, Ser323, Asn366, Asn374 of subunit ND4, as well as 12 water molecules [Figure 3A, inset (ii)]. The QM region was coupled to the MM system via the link atom approach, introduced between the C_α and C_β bonds. The surroundings were described with the CHARMM36 force field[65] and the QM-MM interaction via an additive electro-

https://doi.org/10.1021/acs.jctc.4c00199
J. Chem. Theory Comput. 2024, 20, 5751−5762

Journal of Chemical Theory and Computation pubs.acs.org/JCTC Article

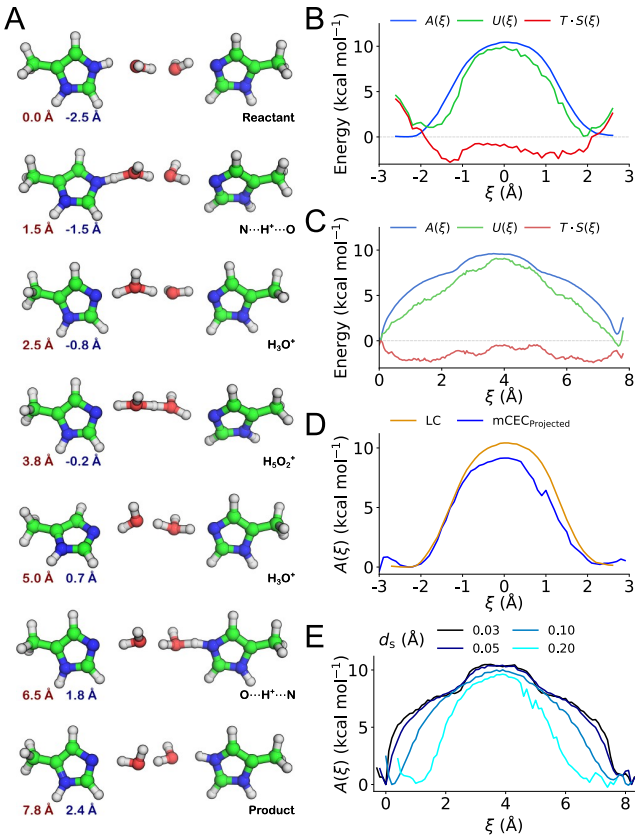


Figure 2. (A) Representative intermediate geometries extracted from the mCEC/MWE sampling with corresponding mCEC (red) and LC (blue) values. (B, C) Free energy A, internal energy U, and entropy $T{\cdot}S$ profile for the His-water array, sampled with MWE as a function of the LC coordinate and mCEC coordinate, respectively. (D) PMF sampled with mCEC projected onto the modified LC in comparison with LC-sampled data, also shown in Figure 1B. (E) PMF profiles obtained from mCEC/MWE sampled data mapped onto the mCEC definitions with different d_s values.

static embedding scheme. The QM/MM system was geometry optimized at B3LYP-D3/def2-SVP level prior to QM/MM free energy sampling.

The water-mediated proton transfer from Lys237 to Glu378 was modeled using the mCEC reaction coordinate with r_s and d_s set to 1.3 and 0.03 Å, respectively, accounting for 27 protons on 17 possible donor/acceptor atoms and including also the 12 nearby water molecules. The US windows were placed between [0.5, 13.75 Å] every 0.25 Å, with a force constant of 100 kcal mol^{-1} Å^{-2}, with additional windows placed in regions of low overlap, resulting in a total simulation time of 300 ps.

The QM/MM-MWE simulations were carried out with a bin width and coupling width of 0.1 Å each with 12 walkers, each sampled for 25 ps. For the well-tempered metadynamics, new Gaussians were added every 10th fs with a bias factor γ = 0.866, a hill height of 0.096 kcal mol^{-1} (0.4 kJ mol^{-1}), and hill width of 0.2 Å. Additional focused US simulations were performed starting from manually selected snapshots of the MWE simulations in chemically interesting and undersampled regions (cf. Figure 4A). To this end, windows were confined to the initial mCEC value of the selected snapshots with a harmonic bias and a force constant of k = 100 kcal mol^{-1}.

Simulations of the proton transfer reaction were also performed with a smaller QM region, comprising 65 atoms (Lys237, His319, His293, and E378, together with six water molecules) to speed up the sampling. The QM region was restrained with positional restraints using a harmonic force constant of 23.9 kcal mol^{-1} Å^{-2} (100 kJ mol^{-1} Å^{-2}) placed on the heavy atoms to limit the sampling along water-mediated pathways in both the US and single-walker/well-tempered metadynamics extended system adaptive biasing force (SWE) simulations. In the SWE, a coupling width of 0.1 Å was employed. A conformational search of the system was additionally carried out using the conformer-rotamer ensemble searching tool (CREST, see Extended Methods for further details).[66]

Convergence of the PMF profiles was assessed by monitoring barrier heights $\Delta A_{RP}^{\ddagger,app}$ as well as free energy

https://doi.org/10.1021/acs.jctc.4c00199
J. Chem. Theory Comput. 2024, 20, 5751−5762

profile differences $\Delta A_{RP}^{app} = A_P - A_R$ a function of the sampling time.

RESULTS

Potential of Mean Force Profiles for Water-Mediated Proton Transfer along a Histidine−Water Array. To test and compare the performance of US and MWE sampling, we first studied a model system (Model 1) of a water-mediated proton transfer reaction using a quasi-one-dimensional water wire connecting two histidine residues (Figure 1). Similar and related model systems have previously been used to explore mechanisms of proton transfer reactions.[11,67−69] The sampling of the proton transfer reaction was performed using (i) the LC (eq 8, Figure 1A) and (ii) the mCEC (eqs 9, 10, Figure 1D) reaction coordinates. Combining these reaction coordinates with both sampling methods (US and MWE) resulted in four conditions tested for Model 1 (Simulations 1−4, see Table S1). The reaction coordinate space was well-sampled and yielded converged PMF profiles for all models (see Figures S1 and S2). Convergence was also probed at the semi-empirical GFN2-xTB level of theory, which supports that the PMF and the apparent free energy barrier converge around 300 ps, in good agreement with the DFT-based sampling (see Figure S2G−I).

Comparison of the sampling methods shows that the MWE method outperforms the US method by reaching convergence in a shorter simulation time. While the US requires around 200 ps of sampling to reach full convergence along the LC reaction coordinate (Figure 1C), we obtain a statistically converged PMF in around 40 ps with MWE. Moreover, for the mCEC coordinate, the convergence with MWE is twice as fast as with US. However, we also note that the apparent free energy barrier ($\Delta A_{RP}^{\ddagger app} = A^{\ddagger} - A_R$) (Figure 1C,F) changes by <5% (0.5 kcal mol^{-1}) after extending the sampling beyond 20 ps per window with US. Both sampling methods predict a transition state that comprises a Zundel ion ($H_5O_2^+$) hydrogen-bonded to both histidine residues, with a similar conformational space explored by both approaches (Figures S3−S5). Both methods also predict activation free energies $\Delta A_{RP}^{\ddagger} \sim 11.1$ kcal mol^{-1} along the LC coordinate, and $\Delta A_{RP}^{\ddagger} \sim 8.5$ kcal mol^{-1} along the mCEC coordinate. This shift could arise from sampling differences in the transitions along the LC and mCEC, although we also observe differences in the predicted entropy profile (Figure 2B,C).

We note that the shape of the PMF is somewhat different along the CVs, which could affect, e.g., the prediction of mean-free passage times.[70,71] Along the LC coordinate, we obtain a PMF profile resembling a Gaussian hill, whereas for the mCEC, the PMF profile has a double-well shape with a broad maximum in the transition state region and steep rise near the reactant and product minima. The saddle points along the MWE profile correspond to configurations where the proton is shared between the histidine and the water molecule (Figure 1E). The MWE minima are also flatter relative to those obtained using US along the LC coordinate, an effect that could arise from the enhanced rotation of the imidazole-ring in the MWE sampling.

Projection of the mCEC sampled conformations onto the LC reaction coordinate suggests that the latter does not uniquely cluster the conformations into reactant and product states. In this regard, we note that sampling along the mCEC coordinate leads to an exchange of the "off-pathway" protons (i.e., protons not directly involved in the Grotthuss wire) with the "on-pathway" protons (Figures S3B and S4A). Such an exchange improves the estimation of entropic effects that are not easily captured when the sampling is performed along the LC coordinate and may in turn lead to an overestimation of the activation free energy (see above).

By accounting for the protons considered for each sampled conformation, we find that a projection of the mCEC conformations onto the LC reaction coordinate can be achieved in *post hoc* analysis. By applying this projection procedure, we recover both the separation between reactant and product regions that remaps the LC coordinate onto the physically relevant range between [−3, +3 Å] (Figure S4B). Moreover, the resulting PMF shows a clear separation between reactant and product minima and the characteristic double-well shape (Figure 2D). However, we note that the profiles still differ in their apparent barrier height, an effect that could arise from sampling differences. To further test the origin of these differences, we computed entropy and internal energy profiles, according to ref 72, by extending the sampling to 300 ps (Simulation 5, see Table S1). For the mCEC coordinate, we observe three peaks in the entropy profile at ξ = 2.5, 3.8, and 5.0 Å (Figure 2C) that arise from the increased combinatorial sampling of atoms in the protonated water species. The four minima on the entropy curve correspond to the restricted configuration space when bonds are broken or formed, as only few orientations are energetically feasible during the proton transfer reaction. Analysis of the O−O distances and the O/H positions reveals that the entropy maxima correspond to hydronium (H_3O^+) and Zundel ($H_5O_2^+$) species (Figures 2A and S3A). The Zundel ion together with the lateral motion of both O atoms (Figure S3C,D) suggest that the process follows a semiconcerted, Grotthuss-like mechanism, with a subtle diffusive component, possibly arising from the constrained His-His distance. Interestingly, the internal energy profile peaks around ξ = 4.0 Å, with the monotonous decreases toward the reactant and product regions, suggesting that the Zundel ion is energetically unstable (Figure 2C). In contrast, for the LC coordinate, the entropy profile shows only one maximum (Figure 2B), as by construction, the PMF along the LC is a superposition of all bond-formation and bond-breaking reactions. Therefore, configurations that comprise Zundel or hydronium ions are mapped to the same LC value around 0 Å. The LC may thus favor a concerted mechanism, while the mCEC would in principle also allow for sequential proton transfers, an aspect that could also contribute to the better estimation of entropic effects along the mCEC coordinate.

We also probed the character of the modification term (eq 9, third term) in the mCEC by mapping the mCEC/MWE data onto the mCEC definitions by testing the effect of different values of the switching width d_s (see eq 11). With an increase in this parameter, we observe a change in the shape of the PMF profile, an increase in the mass of the quasi-particle associated with the CV, and a noticeable decrease of the apparent barrier $\Delta A_{RP}^{\ddagger app}$ from 10.5 to 9.6 kcal mol^{-1} (Figure 2E and Table S1). By increasing d_s, the switching function becomes smoother, which further affects ξ, m_ξ, as well as the shape of $A(\xi)$ (Figure S6). In contrast, we find that the activation free energy $\Delta A_{RP}^{\ddagger}$ is independent of the switching width parameter and remains at 8.5 kcal mol^{-1}.

Protonation Dynamics in the Respiratory Complex I. We next probed the performance of both QM/MM-MWE and QM/MM-US methods on the lateral proton transfer reaction in the membrane domain of respiratory Complex I (Model 2).

https://doi.org/10.1021/acs.jctc.4c00199
J. Chem. Theory Comput. 2024, 20, 5751−5762

Journal of Chemical Theory and Computation pubs.acs.org/JCTC Article

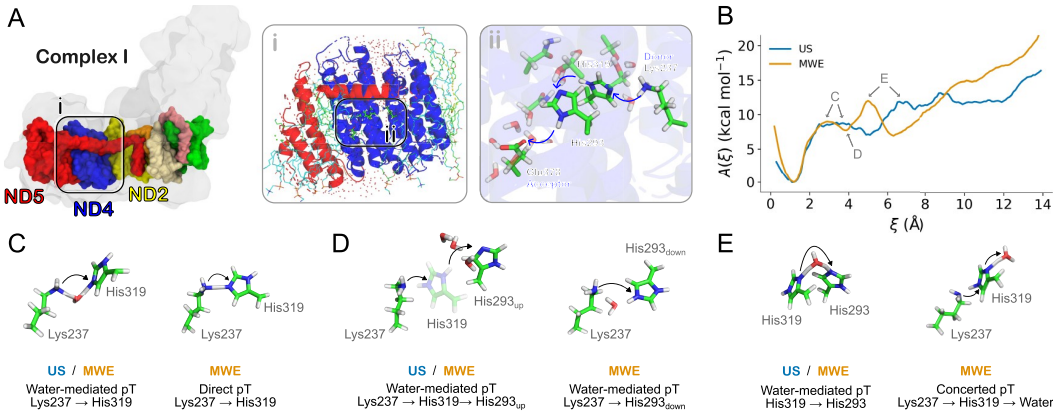


Figure 3. Free energy sampling of proton transfer reactions in the membrane domain of respiratory Complex I (subunit ND4). (A) The mouse Complex I with highlighted core subunits of the membrane domain. Inset i: Overview of the QM/MM system consisting of ND4 and its environment. Inset ii: The QM region with central residues participating in the proton transfer reaction. Blue arrows indicate the proton pathways sampled in the QM/MM free energy calculations. (B) PMFs of the proton transfer reaction obtained with the US and MWE methods. (C–E) Representative conformations for the CV values indicated in panel B, with US (in blue) and MWE (in orange). Due to the explorative sampling by MWE, two representative conformations are shown for the selected states. Additional conformations are shown in Figure S8.

This system is central for understanding biochemical processes underlying cellular respiration. Its mechanistic principles remain elusive and highly debated despite significant efforts over the past decade.[1,73−75] To this end, we explored the proton transfer reaction in a conformational state where the proton transfer reaction is endergonic ("closed ion-pair form", see refs 5,44 for further details). Moreover, we chose to investigate the proton transfer reaction along the mCEC coordinate due to the enhanced sampling properties obtained for Model 1. In this regard, we modeled the proton transfer along the 14 Å water-mediated hydrogen-bonded wire connecting the proton donor Lys237 via His319 and His293 with the proton acceptor Glu378 (Figure 3A).

Both the QM/MM-US and QM/MM-MWE simulations (Simulations 6 and 7, see Table S1) capture the uphill PMF for the proton transfer from Lys237 (ξ < 1 Å) to Glu378 (ξ > 10 Å), favoring the protonated form of Lys237 by around 10 to 12 kcal mol^{-1} (Figures 3B and S7), but with some differences in the shape of the PMF and relative barriers predicted by the two methods. We find that both methods capture the free energy minima for all intermediate states featuring the protonation of Lys237 (ξ = 1 Å), His319 (ξ = 5 Å), His293 (ξ = 8 Å), and Glu378 (ξ = 12 Å). Moreover, both PMF profiles suggest that the initial water-mediated proton transfer from Lys237 to His319 has a $\Delta A_{RP}^{\ddagger,app}$ of ca. 9 kcal mol^{-1}, while the barrier between His319 and His293 is ca. 5 kcal mol^{-1} and shifted toward lower CV value for the MWE PMF (ξ = 5 Å vs ξ = 7 Å for US). The last proton transfer step from His293 to Glu378 is isoenergetic in the US PMF and has an apparent barrier of 2 kcal mol^{-1}, while the MWE PMF suggests that the step is endergonic with an apparent barrier of 8 kcal mol^{-1}.

To understand the differences in the PMF profiles, we compared the conformations, as well as protonation and hydration states. In general, we observe that MWE samples a wider range of states (see below), while the protonation probabilities along the CVs are overall similar for both methods (Figure S8), despite somewhat different protonation profiles around the two intervening histidine residues (Figure S8A,B). In this regard, the saddle point regions comprise Zundel- ($H_5O_2^+$) and Eigen- (H_3O^+)-like species that can be extracted from the US and MWE after clustering the simulation trajectories (Figures 3C–E and S9). Essential dynamics analysis also shows differences in the conformational sampling (Figure S10).

Interestingly, for the initial proton transfer step between Lys237 and His319, the MWE method samples water-mediated conformations (Figure 3C, left), as well as conformations where Lys237 directly donates a proton to His319 (Figure 3C, right). As a consequence, the distance between Lys237 and His319 is reduced, which could explain the shift of the free energy minimum linked to protonation of His319 toward lower values along the reaction coordinate (Figure 3B). Both methods sample a large conformational change featuring a rotation of the His293 ring toward Lys237 ("downward conformation", see Figure 3D for definition) that involves rearrangements of side chains and water molecules. However, with US, this conformation is only sampled for ξ = 12 Å, whereas MWE additionally explores such conformation in the ξ = 3–6 Å regime (Figures 3D and S9), opening an alternative proton transfer pathway, in which His293 could serve as the first intermediate proton acceptor, and thus bypassing His319. These findings suggest that the proton transfer reaction involves conformational changes in His293, which could bridge gaps in the proton wire in Complex I isoforms where His319 is not present.[5,44]

The conformational changes in His293 are further supported by a conformer search conducted using conformer-rotamer ensemble sampling (CREST) (see Extended Methods, Figure S11).[66] Other alternative conformations sampled by MWE involve a hole transfer, comprising an initial exchange of a proton between the two His residues, and leading to the His$^-$/His$^+$ configuration, which is then followed by proton transfer from Lys237 to His$^-$ (on His319) and from HisH$^+$ (on His293) to Glu378 (Figure S9). Although energetically feasible in the sampling, it remains unclear if these states are physically realistic or result, e.g., from DFT charge transfer artifacts.[17]

https://doi.org/10.1021/acs.jctc.4c00199
J. Chem. Theory Comput. 2024, 20, 5751−5762

Journal of Chemical Theory and Computation pubs.acs.org/JCTC Article

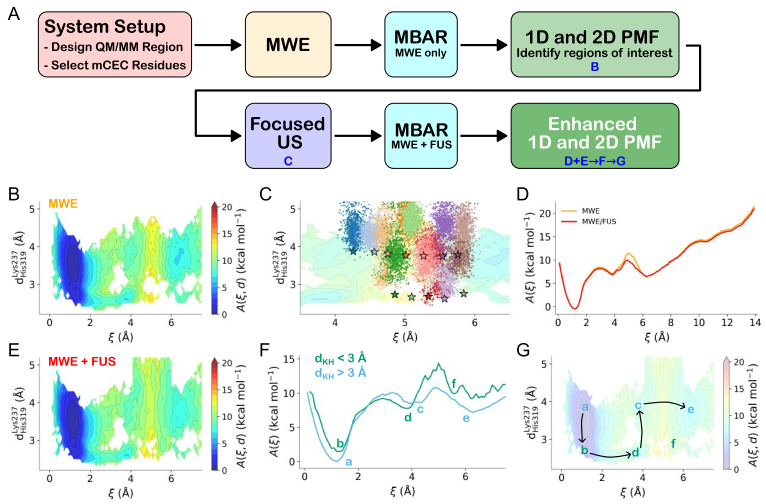


Figure 4. Hybrid MWE/focused US free energy sampling scheme of proton transfer reactions in the membrane domain of Complex I (subunit ND4). (A) Suggested workflow of the MWE/FUS approach. (B–G) Analysis steps as indicated in the flowchart, based on simulation results of Complex I (Model 2). (B) Two-dimensional PMF from MWE trajectories. The two dimensions are the mCEC CV and the distance between Lys237(Nζ) and His319(Nδ) distance (here d_{KH}). (C) 2D-PMF overlaid with initial reaction coordinates of focused US windows and their trajectories. (D) 1D-PMF along the mCEC coordinate before and after addition of the focused sampling. (E) 2D-PMF of combined data from MWE and focused US. (F) PMFs determined for d_{KH} < 3 Å (green) and d_{KH} > 3 Å (cyan). Minima are indicated by a–f. (G) 2D-PMF with minima a–f indicated. The arrows denote possible reaction mechanisms along the a to e pathway.

In general, we note that in addition to the forward proton transfer and reverse hole transfer pathways, the MWE method samples both direct proton transfer between protein residues and water-mediated proton pathways (see above and Figures S9, S12, and S13). Microsecond classical MD indeed supports that the proton pathways in Complex I undergo changes in the overall hydration levels that further affect the free energy of proton transfer and in turn provide possible gating principles (cf. Figure S8 of ref 5). In this regard, the conformational space sampled at the MWE level provides a clear benefit in exploring such alternative reaction mechanisms. However, we note that the direct proton pathways could also arise from the changing biasing force applied to the fictitious particle that may lead to a partial displacement of the intervening water molecule.

To facilitate a better comparison between QM/MM-US and QM/MM-MWE, both sampling methods were constrained to the same reaction mechanism by positionally restraining the heavy atoms participating in the reaction (Simulations 9 and 10, Table S1). This constraint is by construction unphysical but necessary for the direct comparison. As expected, we obtain highly similar PMF profiles for both methods with the constrained sampling (Figure S14). However, the constraints as well as the smaller QM region used to enhance the sampling result in a more endergonic reaction as compared to the unrestrained sampling.

Combining the Advantages of MWE and US—the Hybrid MWE/Focused US Approach. The MWE approach shows a powerful exploration of the reaction phase space, while US has the advantage to systematically show convergence of a given reaction path. We therefore suggest to combine MWE with a focused US to enhance the sampling of phase space regions of special interest or poorly sampled regions. This hybrid MWE/focused US scheme (MWE/FUS, Figure 4A) could improve the accuracy of the obtained PMF for complex systems. To this end, the workflow could involve the following steps, applicable for a biological proton transfer reaction similar to those studied here: (I) system setup, where the QM/MM model is prepared and the residues participating in the mCEC reaction coordinate are selected; (II) an extensive QM/MM-MWE sampling with multiple walkers; (III) clustering of sampled conformations and computation of Boltzmann weights of the microstates using the MBAR; and (IV) compute an initial one-dimensional potential of mean force (1D-PMF) based on the Boltzmann weights and visual inspection to identify further degrees of freedom that are relevant for the process, e.g., characterization of distances between residues or dihedral angles that change during the sampling. These additional degrees of freedom can be used to create two-dimensional potential of mean force (2D-PMF), in which poorly sampled regions of the phase space are identified. (V) Based on the MWE conformations of poorly sampled regions, perform focused US simulations; (VI) derive a refined PMF using MBAR that combines the data from MWE and focused US; (VII) compute new weights that are used to derive refined 1D- and 2D-PMFs. The additional sampling from US improves the reliability of the PMFs, as suggested by sampling the different proton transfer pathways in Complex I (Figure 4E,F).

To illustrate the approach, we next applied the MWE/FUS scheme to study the proton transfer step from His319 to His293, where we observed distinct conformations of Lys237 with His319. This region could be of particular interest as the water-mediated Lys237–His319 connectivity seems to lower the barrier for the His319–His293 reaction relative to the direct proton transfer between Lys237 and His319 (Figure 4B,E). To this end, we initialized five/seven US windows (along low/high d_{KH} values, lysine–histidine distance) for the

https://doi.org/10.1021/acs.jctc.4c00199
J. Chem. Theory Comput. 2024, 20, 5751–5762

two respective connectivity states (Figure 4C, Simulation 8, see Table S1). The Lys237−His319 distance was not constrained to provide an unbiased sampling of the region (Figure 4D).

We find that the focused US indeed improves the 2D-PMF along this region (Figure 4E) and allows the derivation of conformation-dependent 1D-PMFs for the initial reaction steps (Figure 4F). The resulting 2D-PMF reveals that long Lys237−His319 distances are slightly preferred when Lys237 is protonated. Moreover, the apparent barrier is lowered by 1.5 kcal mol^{-1} when the two residues come in close contact, while deprotonation of Lys237 results in longer His−Lys conformations. We therefore propose that the initial proton transfer between Lys237 and His319 could occur by either water-mediated or by direct contact, followed by breaking of the contact and proton transfer from His319 to His293 (Figure 4G). The histidine residues show additional benefits as proton conductors as they can not only exchange protons (e.g., with Nδ as acceptor and Nε as donor) but also show a high conformational flexibility.

DISCUSSION

Our findings suggest that combining accelerated sampling provided by the MWE method with the generalization of the proton transfer reaction through the mCEC reaction coordinate yields a powerful tool for QM/MM free energy calculations of complex (bio)chemical reactions. For our model system of water-mediated proton transfer, we found that the PMF converges faster at the MWE level as compared to US, whereas in the highly intricate proton wires of Complex I, the MWE showed a slower convergence due to the significantly larger conformational space explored. In contrast, QM/MM-US showed a systematic convergence for a given reaction pathway and allowed for an easier parallelization. We also found that by combination of both approaches, the QM/MM-MWE/FUS scheme improves both sampling and accuracy relative to MWE and US (see below).

In this regard, we found that the description of the long-range proton transfer reactions with the mCEC reaction coordinate provides an improved exploration of different regions of conformational space as water molecules along the water array could both reorient and undergo drying/wetting transitions. While conformational sampling with US was locked into predefined pathways, the MWE approach allowed us to map both direct- and water-mediated reaction pathways. This enhancement could provide a benefit in modeling different reaction mechanisms in complex systems, although it also leads to a significantly increased computational cost.

We also found that the system can be biased along the mCEC reaction coordinate such that the sampling displaces intervening water molecules from the "Grotthuss chain". As a consequence, barriers orthogonal to the reaction coordinate are more easily overcome by MWE as compared to US. By introducing positional restraints on the groups participating in the proton transfer reactions, we could show that the US PMF is consistent with the SWE PMF (Figure S14). Indeed, the QM/MM-US method requires also preknowledge of suitable biasing potentials to achieve uniform sampling of the free energy landscape. In this regard, improved initial guesses of the optimal biasing potentials, e.g., from a MWE simulation could significantly enhance the rather slow exploration of the reaction in the QM/MM-US simulations.

To address the sampling and convergence issues, we propose a hybrid MWE/focused US approach for exploration of complex reaction mechanisms. To this end, QM/MM-MWE are performed to explore various putative reaction mechanisms. After identifying different pathways, the PMF is converged using QM/MM-US, while the PMF of the combined data set is obtained using MBAR or related techniques. The high degree of conformational changes and hydration variability in biological systems as observed in atomistic MD simulations highlights the necessity of exploring a wide region of conformational space during free energy sampling and provides a significant challenge particularly for DFT-based QM/MM free energy calculations.

In contrast to previous approaches in which, e.g., metadynamics-explored pathways are subsequently sampled with US,[76,77] the current scheme developed here enables the incorporation of information gained from both MWE and FUS. This combination has the benefit of providing good initial estimates of the PMF with MWE, which can then be greatly improved in the FUS step. Usage of computational resources is thus enhanced, as both extensive and intensive sampling are performed. Nevertheless, FUS also requires a detailed understanding of the studied system in conjunction with a clearly defined target region of the reaction coordinate. Since the MBAR is employed as reweighting procedure, an arbitrary number of US windows can be simulated, allowing for the selection of multiple target regions.

In addition to providing barriers and driving forces, data obtained from free energy calculations can be used to describe the structures of relevant intermediates along the reaction pathways. We suggest that reaction intermediates obtained from the MWE sampling can be extracted using clustering methods. However, the direct time-ordering of physically realistic intermediates for derivation of reaction mechanisms is perhaps more straightforward in the QM/MM-US simulations, as the biasing force reaches a local equilibrium in each simulation window, thus allowing for the characterization of reaction steps along the collective variable. The mCEC/MWE approach suggests that some of the protonation reactions in Complex I may occur via multiple competing reaction pathways, possibly modulated by the hydration and protonation states of the surrounding groups. Although this requires further detailed exploration, it suggests interesting gating principles that could be used to modulate proton conduction, e.g., during reversal of the proton pumping machinery, which takes place during hypoxic conditions in mitochondria.

CONCLUSIONS

We have introduced, implemented, and compared here the multiple walker/well-tempered metadynamics/extended adaptive biasing force (MWE) and umbrella sampling (US) methods for GPU-accelerated QM/MM free energy calculations of water-mediated proton transfer reactions in a complex (bio)chemical system. To this end, we studied the performance of both MWE and US in combination with a geometric and a physical description of the proton transfer reaction coordinate. Our combined findings show that the MWE approach combined with the modified center of excess charge (mCEC) reaction coordinate can efficiently sample the protonation dynamics and orthogonal hydration dynamics, while the US method effectively sampled conformations containing the hydration pattern of the starting conformation. In contrast, we found that the mCEC/MWE approach achieves a multi-pathway exploration of the reaction coordinate that, for the studied model systems, converges

https://doi.org/10.1021/acs.jctc.4c00199
J. Chem. Theory Comput. 2024, 20, 5751−5762

within a time scale similar to that of the US method. For the proton transfer reactions in Complex I, we observe that convergence of the MWE sampling is increased relative to the US method, while combination of both approaches (QM/MM-MWE/FUS) provide an improved exploration of the 2D-PMF profiles. Our study provides key insights into the application of first-principles-based QM/MM free energy methods for mechanistic studies of protonation dynamics in highly intricate biochemical systems and highlights new mechanistic insight into protonation dynamics in Complex I.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.4c00199.

Convergence of the PMF, methods testing, and reaction coordinate values of central states for the studied systems (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Ville R. I. Kaila − *Department of Biochemistry and Biophysics, Stockholm University, 10691 Stockholm, Sweden;* orcid.org/0000-0003-4464-6324; Email: ville.kaila@dbb.su.se

Authors

Maximilian C. Pöverlein − *Department of Biochemistry and Biophysics, Stockholm University, 10691 Stockholm, Sweden*

Andreas Hulm − *Chair of Theoretical Chemistry, Department of Chemistry, University of Munich (LMU), 81377 Munich, Germany;* orcid.org/0000-0003-1268-7578

Johannes C. B. Dietschreit − *Chair of Theoretical Chemistry, Department of Chemistry, University of Munich (LMU), 81377 Munich, Germany; Department of Material Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0002-5840-0002

Jörg Kussmann − *Chair of Theoretical Chemistry, Department of Chemistry, University of Munich (LMU), 81377 Munich, Germany;* orcid.org/0000-0002-4724-8551

Christian Ochsenfeld − *Chair of Theoretical Chemistry, Department of Chemistry, University of Munich (LMU), 81377 Munich, Germany; Max Planck Institute for Solid State Research, D-70569 Stuttgart, Germany;* orcid.org/0000-0002-4189-6558

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.4c00199

Author Contributions

All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was funded by the Knut and Alice Wallenberg Foundation (grant: 2019.0251 and WASPDDLS22:025, V.R.I.K.), the Swedish Research Council (V.R.I.K.), and the German Research Foundation (DFG, TRR235, "Emergence of Life" to V.R.I.K. and C.O.), and by the European Research Council under the European Union's Horizon 2020 research and innovation program/Grant Agreement 715311 (V.R.I.K.). J.C.B.D. acknowledges the support of the Leopoldina Fellowship Program, German National Academy of Sciences Leopoldina, grant number LPDS 2021-08. V.R.I.K. acknowledges support from the German Research Foundation (DFG) via the Collaborative Research Center (SFB1078). This work was also supported by the National Academic Infrastructure for Supercomputing in Sweden (NAISS, NAISS: 2023/1-31, 2023/6-128) and the Swedish National Infrastructure for Computing (SNIC 2022/1-29, SNIC 2022/13-14) at the Center for High-Performance Computing (PDC) Center, partially funded by the Swedish Research Council through grant agreements no. 2022-06725 and no. 2018-05973, and the Leibniz Rechenzentrum (LRZ, project:pr83ro), Germany.

■ ABBREVIATIONS

PMF - potential of mean force; CV - collective variable; RC - reaction coordinate; LC - linear combination of bond-breaking and bond-formation; mCEC - modified center of excess charge; US - umbrella sampling; WTM-eABF - well-tempered metadynamics/extended adaptive biasing force; MWE - multiple-walkers/well-tempered metadynamics/extended adaptive biasing force; SWE - single-walker/well-tempered metadynamics/extended adaptive biasing force; QM/MM - hybrid quantum mechanics/molecular mechanics; MD - molecular dynamics; DFTB - density functional based tight binding; SCC-DFTB - self-consistent charge DFTB; PM7 - parametric method 7; EVB - empirical valence bond; MS-EVB - multistate EVB

■ REFERENCES

(1) Kaila, V. R. I. Resolving Chemical Dynamics in Biological Energy Conversion: Long-Range Proton-Coupled Electron Transfer in Respiratory Complex I. *Acc. Chem. Res.* **2021**, *54* (24), 4462−4473.

(2) Kaila, V. R. I.; Wikström, M. Architecture of bacterial respiratory chains. *Nat. Rev. Microbiol.* **2021**, *19* (5), 319−330.

(3) Agmon, N. The Grotthuss mechanism. *Chem. Phys. Lett.* **1995**, *244* (5), 456−462.

(4) Mitchell, P. Coupling of Phosphorylation to Electron and Hydrogen Transfer by a Chemi-Osmotic type of Mechanism. *Nature* **1961**, *191* (4784), 144−148.

(5) Röpke, M.; Saura, P.; Riepl, D.; Pöverlein, M. C.; Kaila, V. R. I. Functional Water Wires Catalyze Long-Range Proton Pumping in the Mammalian Respiratory Complex I. *J. Am. Chem. Soc.* **2020**, *142* (52), 21758−21766.

(6) Röpke, M.; Riepl, D.; Saura, P.; Di Luca, A.; Mühlbauer, M. E.; Jussupow, A.; Gamiz-Hernandez, A. P.; Kaila, V. R. I. Deactivation blocks proton pathways in the mitochondrial complex I. *Proc. Natl. Acad. Sci. U.S.A.* **2021**, *118* (29), No. e2019498118.

(7) Mader, S. L.; Lopez, A.; Lawatscheck, J.; Luo, Q.; Rutz, D. A.; Gamiz-Hernandez, A. P.; Sattler, M.; Buchner, J.; Kaila, V. R. I. Conformational dynamics modulate the catalytic activity of the molecular chaperone Hsp90. *Nat. Commun.* **2020**, *11* (1), No. 1410.

(8) Liang, R.; Swanson, J. M. J.; Peng, Y.; Wikström, M.; Voth, G. A. Multiscale simulations reveal key features of the proton-pumping mechanism in cytochrome *c* oxidase. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113* (27), 7420−7425.

(9) Li, C.; Yue, Z.; Espinoza-Fonseca, L. M.; Voth, G. A. Multiscale Simulation Reveals Passive Proton Transport Through SERCA on the Microsecond Timescale. *Biophys. J.* **2020**, *119* (5), 1033−1040.

(10) Pisliakov, A. V.; Sharma, P. K.; Chu, Z. T.; Haranczyk, M.; Warshel, A. Electrostatic basis for the unidirectionality of the primary proton transfer in cytochrome *c* oxidase. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105* (22), 7726−7731.

https://doi.org/10.1021/acs.jctc.4c00199
J. Chem. Theory Comput. 2024, 20, 5751−5762

(11) Maag, D.; Mast, T.; Elstner, M.; Cui, Q.; Kubař, T. O to bR transition in bacteriorhodopsin occurs through a proton hole mechanism. *Proc. Natl. Acad. Sci. U.S.A.* **2021**, *118* (39), No. e2024803118.
(12) Elstner, M. The SCC-DFTB method and its application to biological systems. *Theor. Chem. Acc.* **2006**, *116* (1), 316−325.
(13) Mlýnský, V.; Banáš, P.; Šponer, J.; van der Kamp, M. W.; Mulholland, A. J.; Otyepka, M. Comparison of ab Initio, DFT, and Semiempirical QM/MM Approaches for Description of Catalytic Mechanism of Hairpin Ribozyme. *J. Chem. Theory Comput.* **2014**, *10* (4), 1608−1622.
(14) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Phys. Rev. B* **1998**, *58* (11), 7260−7268.
(15) Stewart, J. J. P. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. *J. Mol. Model.* **2013**, *19* (1), 1−32.
(16) Christensen, A. S.; Kubař, T.; Cui, Q.; Elstner, M. Semiempirical Quantum Mechanical Methods for Noncovalent Interactions for Chemical and Biochemical Applications. *Chem. Rev.* **2016**, *116* (9), 5301−5337.
(17) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Challenges for Density Functional Theory. *Chem. Rev.* **2012**, *112* (1), 289−320.
(18) Siegbahn, P. E. M.; Blomberg, M. R. A. Transition-Metal Systems in Biochemistry Studied by High-Accuracy Quantum Chemical Methods. *Chem. Rev.* **2000**, *100* (2), 421−438.
(19) Kubař, T.; Elstner, M.; Cui, Q. Hybrid Quantum Mechanical/Molecular Mechanical Methods For Studying Energy Transduction in Biomolecular Machines. *Annu. Rev. Biophys.* **2023**, *52* (1), 525−551.
(20) Kussmann, J.; Ochsenfeld, C. Preselective Screening for Linear-Scaling Exact Exchange-Gradient Calculations for Graphics Processing Units and General Strong-Scaling Massively Parallel Calculations. *J. Chem. Theory Comput.* **2015**, *11* (3), 918−922.
(21) Kussmann, J.; Ochsenfeld, C. Hybrid CPU/GPU Integral Engine for Strong-Scaling Ab Initio Methods. *J. Chem. Theory Comput.* **2017**, *13* (7), 3153−3159.
(22) Fu, H.; Shao, X.; Cai, W.; Chipot, C. Taming Rugged Free Energy Landscapes Using an Average Force. *Acc. Chem. Res.* **2019**, *52* (11), 3254−3264.
(23) König, P. H.; Ghosh, N.; Hoffmann, M.; Elstner, M.; Tajkhorshid, E.; Frauenheim, T.; Cui, Q. Toward Theoretical Analysis of Long-Range Proton Transfer Kinetics in Biomolecular Pumps. *J. Phys. Chem. A* **2006**, *110* (2), 548−563.
(24) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99* (20), 12562−12566.
(25) Darve, E.; Pohorille, A. Calculating free energies using average force. *J. Chem. Phys.* **2001**, *115* (20), 9169−9183.
(26) Sørensen, M. R.; Voter, A. F. Temperature-accelerated dynamics for simulation of infrequent events. *J. Chem. Phys.* **2000**, *112* (21), 9599−9606.
(27) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **1977**, *23* (2), 187−199.
(28) Lesage, A.; Lelièvre, T.; Stoltz, G.; Hénin, J. Smoothed Biasing Forces Yield Unbiased Free Energies with the Extended-System Adaptive Biasing Force Method. *J. Phys. Chem. B* **2017**, *121* (15), 3676−3685.
(29) Hulm, A.; Dietschreit, J. C. B.; Ochsenfeld, C. Statistically optimal analysis of the extended-system adaptive biasing force (eABF) method. *J. Chem. Phys.* **2022**, *157* (2), No. 024110.
(30) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* **1992**, *13*, 1011−1021.
(31) Shirts, M. R.; Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* **2008**, *129* (12), No. 124105.
(32) Wu, H.; Paul, F.; Wehmeyer, C.; Noé, F. Multiensemble Markov models of molecular thermodynamics and kinetics. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113* (23), E3221−E3230.
(33) Wu, H.; Mey, A. S. J. S.; Rosta, E.; Noé, F. Statistically optimal analysis of state-discretized trajectory data from multiple thermodynamic states. *J. Chem. Phys.* **2014**, *141* (21), No. 214106.
(34) Rosta, E.; Hummer, G. Free Energies from Dynamic Weighted Histogram Analysis Using Unbiased Markov State Model. *J. Chem. Theory Comput.* **2015**, *11* (1), 276−285.
(35) Stelzl, L. S.; Kells, A.; Rosta, E.; Hummer, G. Dynamic Histogram Analysis To Determine Free Energies and Rates from Biased Simulations. *J. Chem. Theory Comput.* **2017**, *13* (12), 6328−6342.
(36) Dietschreit, J. C. B.; Diestler, D. J.; Hulm, A.; Ochsenfeld, C.; Gómez-Bombarelli, R. From free-energy profiles to activation free energies. *J. Chem. Phys.* **2022**, *157* (8), No. 084113.
(37) Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **1988**, *38* (6), 3098−3100.
(38) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **1988**, *37* (2), 785−789.
(39) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7* (18), 3297−3305.
(40) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132* (15), No. 154104.
(41) Mangiatordi, G. F.; Brémond, E.; Adamo, C. DFT and Proton Transfer Reactions: A Benchmark Study on Structure and Kinetics. *J. Chem. Theory Comput.* **2012**, *8* (9), 3082−3088.
(42) David, R.; Jamet, H.; Nivière, V.; Moreau, Y.; Milet, A. Iron Hydroperoxide Intermediate in Superoxide Reductase: Protonation or Dissociation First? MM Dynamics and QM/MM Metadynamics Study. *J. Chem. Theory Comput.* **2017**, *13* (6), 2987−3004.
(43) Duster, A. W.; Lin, H. Tracking Proton Transfer through Titratable Amino Acid Side Chains in Adaptive QM/MM Simulations. *J. Chem. Theory Comput.* **2019**, *15* (11), 5794−5809.
(44) Mühlbauer, M. E.; Saura, P.; Nuber, F.; Di Luca, A.; Friedrich, T.; Kaila, V. R. I. Water-Gated Proton Transfer Dynamics in Respiratory Complex I. *J. Am. Chem. Soc.* **2020**, *142* (32), 13718−13728.
(45) Yagi, K.; Ito, S.; Sugita, Y. Exploring the Minimum-Energy Pathways and Free-Energy Profiles of Enzymatic Reactions with QM/MM Calculations. *J. Phys. Chem. B* **2021**, *125* (18), 4701−4713.
(46) Dürr, S. L.; Bohuszewicz, O.; Berta, D.; Suardiaz, R.; Jambrina, P. G.; Peter, C.; Shao, Y.; Rosta, E. The Role of Conserved Residues in the DEDDh Motif: the Proton-Transfer Mechanism of HIV-1 RNase H. *ACS Catal.* **2021**, *11* (13), 7915−7927.
(47) Kim, H.; Saura, P.; Pöverlein, M. C.; Gamiz-Hernandez, A. P.; Kaila, V. R. I. Quinone Catalysis Modulates Proton Transfer Reactions in the Membrane Domain of Respiratory Complex I. *J. Am. Chem. Soc.* **2023**, *145* (31), 17075−17086.
(48) Sheng, X.; Himo, F. The Quantum Chemical Cluster Approach in Biocatalysis. *Acc. Chem. Res.* **2023**, *56* (8), 938−947.
(49) Siegbahn, P. E. M. A quantum chemical approach for the mechanisms of redox-active metalloenzymes. *RSC Adv.* **2021**, *11* (6), 3495−3508 10.1039/D0RA10412D.
(50) Klamt, A.; Schüürmann, G. COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc., Perkin Trans. 2* **1993**, No. 5, 799−805.
(51) Shirts, M. R.; Ferguson, A. L. Statistically Optimal Continuous Free Energy Surfaces from Biased Simulations and Multistate Reweighting. *J. Chem. Theory Comput.* **2020**, *16* (7), 4107−4125.

https://doi.org/10.1021/acs.jctc.4c00199
J. Chem. Theory Comput. 2024, 20, 5751−5762

(52) Kussmann, J.; Ochsenfeld, C. Pre-selective screening for matrix elements in linear-scaling exact exchange calculations. *J. Chem. Phys.* **2013**, *138* (13), No. 134114.
(53) Laqua, H.; Kussmann, J.; Ochsenfeld, C. Accelerating seminumerical Fock-exchange calculations using mixed single- and double-precision arithmethic. *J. Chem. Phys.* **2021**, *154* (21), No. 214116.
(54) Kussmann, J.; Laqua, H.; Ochsenfeld, C. Highly Efficient Resolution-of-Identity Density Functional Theory Calculations on Central and Graphics Processing Units. *J. Chem. Theory Comput.* **2021**, *17* (3), 1512−1521.
(55) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; et al. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* **2017**, *13* (7), No. e1005659.
(56) Laqua, H.; Thompson, T. H.; Kussmann, J.; Ochsenfeld, C. Highly Efficient, Linear-Scaling Seminumerical Exact-Exchange Method for Graphic Processing Units. *J. Chem. Theory Comput.* **2020**, *16* (3), 1456−1468.
(57) Laqua, H.; Dietschreit, J. C. B.; Kussmann, J.; Ochsenfeld, C. Accelerating Hybrid Density Functional Theory Molecular Dynamics Simulations by Seminumerical Integration, Resolution-of-the-Identity Approximation, and Graphics Processing Units. *J. Chem. Theory Comput.* **2022**, *18* (10), 6010−6020.
(58) TURBOMOLE V7.4 2019, A Development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH; TURBOMOLE GmbH, 1989−2007.
(59) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. Electronic structure calculations on workstation computers: The program system turbomole. *Chem. Phys. Lett.* **1989**, *162* (3), 165−169.
(60) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graphics* **1996**, *14* (1), 33−38.
(61) Delano, W. L. *PyMOL Molecular Graphics System*; Schrödinger LLC, 2002. https://sourceforge.net/projects/pymol/.
(62) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15* (3), 1652−1671.
(63) Ehlert, S.; Stahn, M.; Spicher, S.; Grimme, S. Robust and Efficient Implicit Solvation Model for Fast Semiempirical Methods. *J. Chem. Theory Comput.* **2021**, *17* (7), 4250−4261.
(64) Bridges, H. R.; Fedor, J. G.; Blaza, J. N.; Di Luca, A.; Jussupow, A.; Jarman, O. D.; Wright, J. J.; Agip, A.-N. A.; Gamiz-Hernandez, A. P.; Roessler, M. M.; et al. Structure of inhibitor-bound mammalian complex I. *Nat. Commun.* **2020**, *11* (1), No. 5261.
(65) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D., Jr. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain χ1 and χ2 Dihedral Angles. *J. Chem. Theory Comput.* **2012**, *8* (9), 3257−3273.
(66) Pracht, P.; Bohle, F.; Grimme, S. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Phys. Chem. Chem. Phys.* **2020**, *22* (14), 7169−7192.
(67) Kaila, V. R. I.; Hummer, G. Energetics and dynamics of proton transfer reactions along short water wires. *Phys. Chem. Chem. Phys.* **2011**, *13* (29), 13207−13215 10.1039/C1CP21112A.
(68) Kaila, V. R. I.; Hummer, G. Energetics of Direct and Water-Mediated Proton-Coupled Electron Transfer. *J. Am. Chem. Soc.* **2011**, *133* (47), 19040−19043.
(69) Saura, P.; Frey, D. M.; Gamiz-Hernandez, A. P.; Kaila, V. R. I. Electric field modulated redox-driven protonation and hydration energetics in energy converting enzymes. *Chem. Commun.* **2019**, *55* (43), 6078−6081.
(70) Berezhkovskii, A. M.; Szabo, A. Committors, first-passage times, fluxes, Markov states, milestones, and all that. *J. Chem. Phys.* **2019**, *150* (5), No. 054106.
(71) Chupeau, M.; Gladrow, J.; Chepelianskii, A.; Keyser, U. F.; Trizac, E. Optimizing Brownian escape rates by potential shaping. *Proc. Natl. Acad. Sci. U.S.A.* **2020**, *117* (3), 1383−1388.
(72) Dietschreit, J. C. B.; Diestler, D. J.; Gómez-Bombarelli, R. Entropy and Energy Profiles of Chemical Reactions. *J. Chem. Theory Comput.* **2023**, *19* (16), 5369−5379.
(73) Kaila, V. R. I. Long-range proton-coupled electron transfer in biological energy conversion: towards mechanistic understanding of respiratory complex I. *J. R. Soc., Interface* **2018**, *15* (141), No. 20170916.
(74) Chung, I.; Grba, D. N.; Wright, J. J.; Hirst, J. Making the leap from structure to mechanism: are the open states of mammalian complex I identified by cryoEM resting states or catalytic intermediates? *Curr. Opin. Struct. Biol.* **2022**, *77*, No. 102447.
(75) Kampjut, D.; Sazanov, L. A. Structure of respiratory complex I—An emerging blueprint for the mechanism. *Curr. Opin. Struct. Biol.* **2022**, *74*, No. 102350.
(76) Autieri, E.; Sega, M.; Pederiva, F.; Guella, G. Puckering free energy of pyranoses: A NMR and metadynamics-umbrella sampling investigation. *J. Chem. Phys.* **2010**, *133* (9), No. 095104.
(77) Babin, V.; Roland, C.; Darden, T. A.; Sagui, C. The free energy landscape of small peptides as obtained from metadynamics with umbrella sampling corrections. *J. Chem. Phys.* **2006**, *125* (20), No. 204909.

https://doi.org/10.1021/acs.jctc.4c00199
J. Chem. Theory Comput. 2024, 20, 5751−5762

Supplementary Information

QM/MM Free Energy Calculations of Long-Range Biological Protonation Dynamics by Adaptive and Focused Sampling

Maximilian C. Pöverlein[1], Andreas Hulm[2], Johannes C. B. Dietschreit[2,3], Jörg Kussmann[2], Christian Ochsenfeld[2,4], Ville R. I. Kaila[1,*]

[1]Department of Biochemistry and Biophysics, Stockholm University, 10691, Stockholm, Sweden.
[2]Chair of Theoretical Chemistry, Department of Chemistry, University of Munich (LMU), 81377 Munich, Germany.
[3]Department of Material Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.
[4]Max Planck Institute for Solid State Research, D-70569 Stuttgart, Germany.
[*] To whom correspondence should be addressed: Ville R. I. Kaila, ville.kaila@dbb.su.se

Content

Extended Methods

Essential Dynamics Analysis

Essential dynamics analysis (EDA) for the His-water model (Model 1; Figure 1) was performed on the Cartesian coordinates of all atoms in the system. For the proton transfer reaction in Complex I (Model 2), the EDA was performed on the Cartesian coordinates of the QM atoms of the QM/MM system (Figure 3). The EDA was performed using ProDy.[1] Similar to principal component analysis (PCA), the EDA orthogonalizes the covariance matrix to obtain principle modes of motion. The eigenvectors corresponding to the largest eigenvalues are interpreted as the dominant modes of motion.[2]

Semi-Empirical Conformer Search

Based on the initial conformation obtained from free energy sampling of the proton transfer reactions in Complex I, a conformational search was performed using the conformer-rotamer ensemble sampling tool (CREST).[3] The conformational search was performed on the QM region and its immediate surroundings, by extending the QM region from 114 atoms to 216 atoms (Figure S7A). The sidechains of amino acids were modeled by replacing the C_α atoms by hydrogen atoms, which were fixed to the C_α positions in the reference structure. After geometry optimization at the GFN2-xTB level of theory,[4] the conformational search was also performed at the GFN2-xTB level.[4]

CV mass and de Broglie thermal wavelength

The inverse mass of the pseudo-particle associated with the reaction coordinate can be defined as (*cf.* also Ref. [5], Eqn. 27),

$$m_\xi^{-1} = \left(\vec{\nabla}_x \xi\right)^{\mathrm{T}} \mathbf{M^{-1}} \left(\vec{\nabla}_x \xi\right) = \sum_i^{3N} m_i^{-1} \left(\frac{\partial \xi}{\partial x_i}\right)^2 \qquad (1)$$

where $\vec{\nabla}_x \xi$ denotes the gradient of the reaction coordinate with respect to the nuclear coordinates and $\mathbf{M}$ is the $3N \times 3N$ diagonal matrix of atomic masses. The de Broglie thermal wavelength of the reaction coordinate-associated pseudo-particle (Eqn. 7, *cf.* also Ref. [5]) is defined as,

$$\lambda_\xi \equiv \frac{h}{\sqrt{2\pi m_\xi k_B T}} \qquad (2)$$

where T is the temperature of the system.

Derivation of Eqn. 7. for the activation free energy for a 1D harmonic oscillator

Let us assume that the potential at the reactant minimum (R_{min}) and transition state (TS) can be approximated harmonically by,

$$H_{R_{min}} = U_{R_{min}} + \frac{1}{2}(kq^2 + \frac{p^2}{m}) \qquad H_{TS} = U_{TS} \tag{3}$$

where q is the normal mode and p the conjugate momentum and U_α the potential energy of configuration, with the imaginary mode for the transition state excluded from the Hamiltonian. The activation free energy is the difference of the transition state and reactant free energies,

$$\Delta F^{\ddagger} = F_{TS} - F_R = -\beta^{-1} \ln \frac{Q_{TS}}{Q_R} = -\beta^{-1} \ln \frac{Z_{TS}\, \Lambda_R}{\Lambda_{TS}\, Z_R} \tag{4}$$

Using the harmonic Hamiltonians, the analytical configuration integrals Z_α and product of thermal wavelengths Λ_α for reactant and transition state are,

$$Z_R = e^{-\beta U_R} \int_{-\infty}^{\infty} \mathrm{d}q\, e^{-\frac{\beta k q^2}{2}} = e^{-\beta U_R} \sqrt{\frac{2\pi}{\beta k}} \qquad Z_{TS} = e^{-\beta U_{TS}} \tag{5}$$

$$\Lambda_R{}^{-1} = \frac{1}{h} \int_{-\infty}^{\infty} \mathrm{d}p\, e^{-\frac{\beta p^2}{2m}} = \sqrt{\frac{2\pi m}{\beta h^2}} \qquad \Lambda_{TS}{}^{-1} = 1 \tag{6}$$

Z_R and Z_{TS} differ by the dimension of one normal mode, and Λ_R and Λ_{TS} by exactly one thermal wavelength. Hence, the argument of the logarithm in the equation for the activation free energy retains the thermal wavelength,

$$\Delta F^{\ddagger} = -\beta^{-1} \ln \left(\frac{e^{-\beta U_{TS}} \sqrt{\beta h^2 / 2\pi m}}{e^{-\beta U_R} \sqrt{2\pi / \beta k}} \right) \tag{7}$$

Comparing the harmonic approximation with Eq. 7 yields the corresponding terms,

$$\frac{Z_{TS}}{Z_R} = \frac{e^{-\beta U_{TS}}}{e^{-\beta U_R} \sqrt{2\pi/\beta k}} \mathrel{\widehat{=}} \frac{\rho(z^{\ddagger})}{P(\mathrm{R})} \quad \text{and} \quad \frac{\Lambda_R}{\Lambda_{TS}} = \sqrt{\beta h^2 / 2\pi m} \mathrel{\widehat{=}} \langle \lambda_\xi \rangle_{z^{\ddagger}} \tag{8}$$

Supplementary Figures

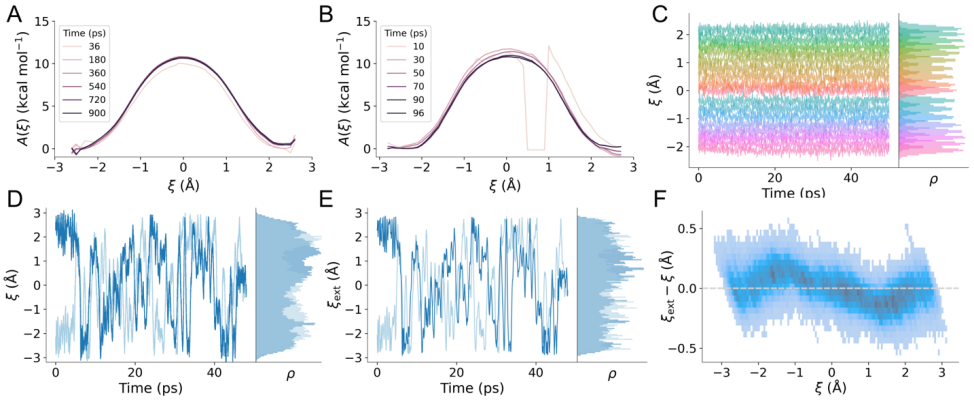


Figure S1. Sampling of the proton transfer reaction in the His-water model along the linear combination (LC) reaction coordinate. (A) Convergence of the PMF obtained from US. (B) Convergence of the PMF from MWE. (C) Sampling of the reaction coordinate using QM/MM-US. (D) Sampling of the reaction coordinate using QM/MM-MWE. (E) Sampling of the extended variable (eABF) reaction coordinate in MWE. (F) 2D histogram of the instantaneous difference between the extended system variable and the reaction coordinate as a function of the reaction coordinate.

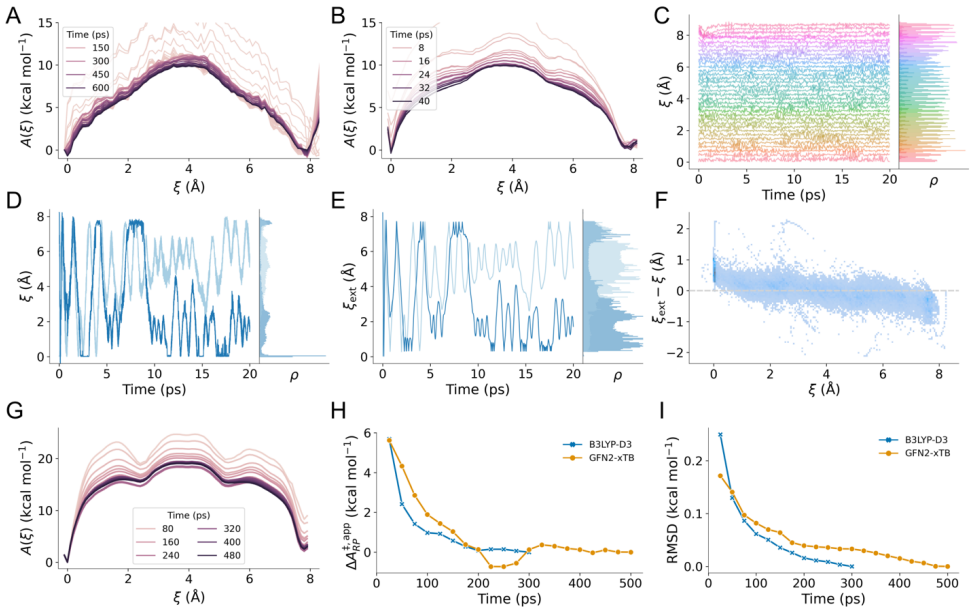


Figure S2. Sampling of the proton transfer reaction in the His-water model system using the modified center of excess charge reaction coordinate. (A) Convergence of the PMF obtained from US. (B) Convergence of the PMF from MWE. (C) Sampling of the extended variable reaction coordinate by US. (D) Sampling of the reaction coordinate using QM/MM-MWE. (E) Sampling of the extended variable (eABF) reaction coordinate in MWE. (F) 2D histogram of the instantaneous difference between the extended system variable and the reaction coordinate as a function of the reaction coordinate. (G) Convergence of the PMF from MWE at the GFN2-xTB level of theory. (H) Convergence of apparent barrier height ($\Delta A_{RP}^{\ddagger,\mathrm{app}} = A^{\ddagger} - A_R$) and (I) evolution of the PMF RMSD both with respect to final PMF profile.

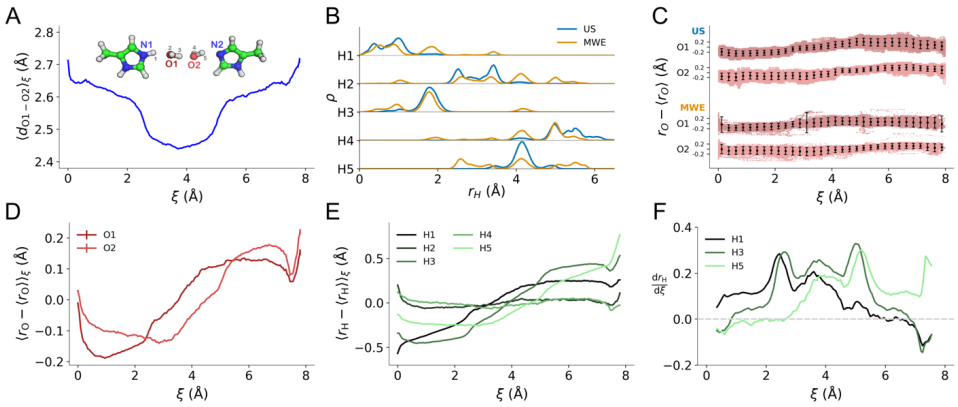


Figure S3. Conformational sampling of proton transfer in the His-water model. (A) Weighted average of O-O distance. (B) Histogram of the proton positions projected onto the donor-acceptor vector. (C) Projection of the water-O1 and water-O2 positions onto the reaction coordinate vector plotted versus reaction coordinate value. Error bars indicate 5th and 95th percentile. (D) Relative position of the O atoms projected onto the donor-acceptor vector. (E) Projected relative positions of the H atoms. (F) Derivative of the projected positions of the H atoms H1, H3, and H5. MWE data shown in A-F were obtained from 300 ps run of MWE.

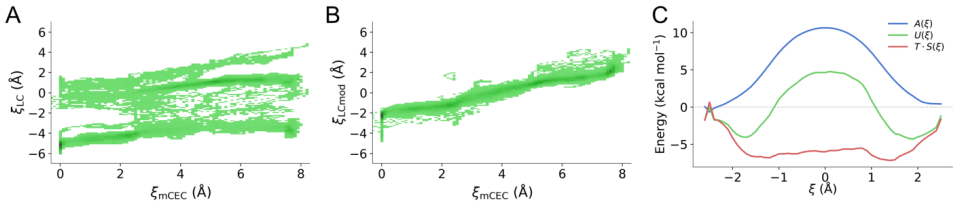


Figure S4. Projection of the mCEC-CV on the LC-CV. (A) Mapping of the mCEC sampling onto LC reaction coordinate. Reactant and product minima are indicated by dashed lines. Proton exchange during mCEC sampling leads to absence of separation between reactant and product state for the LC RC. (B) Projection of a modified LC reaction coordinate re-maps the reactant and product state to the physical RC range. (C) Free energy A, internal energy U, and entropy $T \cdot S$ profile of the His-water model as a function of the mCEC coordinate sampled with MWE.

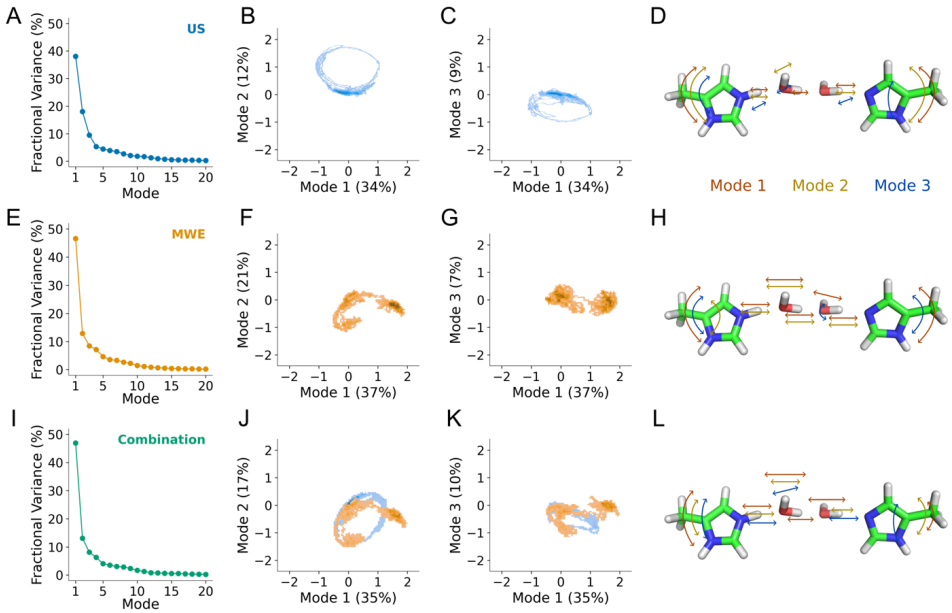


Figure S5. Essential dynamics analysis of the His-water model with the modified center of excess charge reaction coordinate using (A-D) US, (E-H) MWE. (I-L) Combination of the data from US and MWE. (A, E, I) Fractional variance of the first 20 principal components. (B, F, J) Projection of the trajectory data onto PC 1 and PC 2. (C, G, K) Projection of the trajectory data onto PC 1 and PC 3. (D, H, L) Visualization of the movements described by PC 1, PC 2, and PC 3.

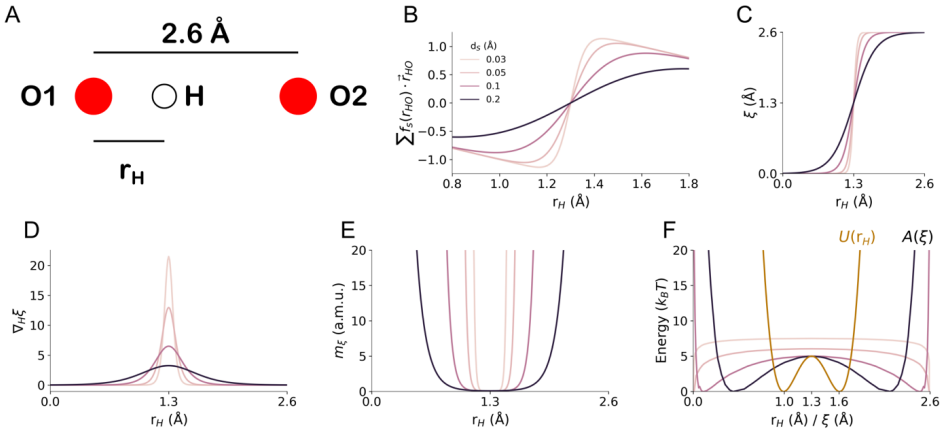


Figure S6. Effect of switching parameter on mCEC-CV and PMF. (A) Illustration of the of a O-H-O test system. (B) Sum of the mCEC correction terms as a function of the hydrogen position r_H for different d_s. (C) mCEC-CV value as a function of r_H. (D) Gradient of the mCEC with respect to r_H. (E) Mass of the mCEC. (F) PMFs $A(\xi;d_s)$ when the model system is sampled assuming a double-well potential (red) $U(x)=565.8436\,(x^4 - 5.2x^3 + 9.9518x^2 - 8.2987x + 2.5469)$, which was obtained by fitting a fourth degree polynomial to the points ((0.7,40), (1,0), (1.3,5), (1.6,0), (1.9,40)).

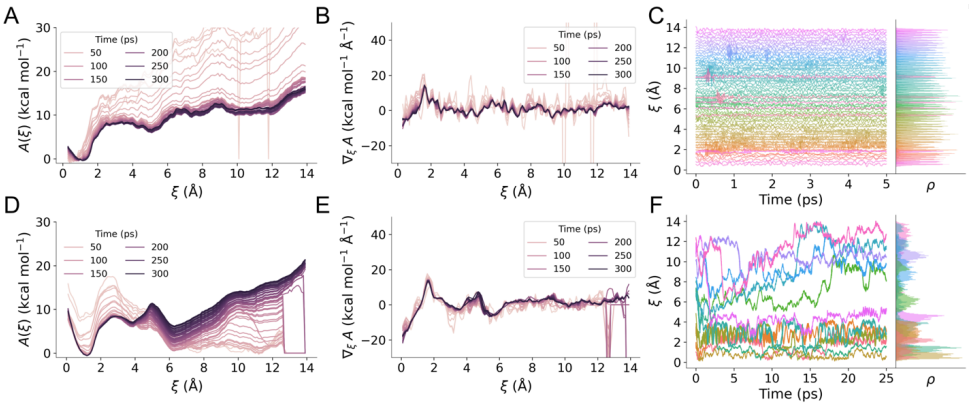


Figure S7. Sampling of the proton transfer reactions in the membrane domain of Complex I with the US and MWE methods. (A,D) Convergence of the PMF obtained from US and MWE, respectively. (B,E) Convergence of the mean force profile obtained from US and MWE, respectively. (C,F) Sampling of the reaction coordinate using US and MWE.

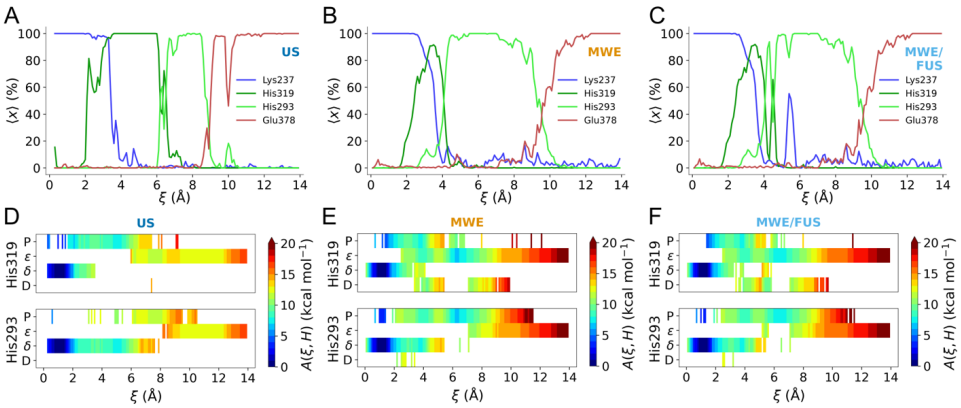


Figure S8. Sampling of protonation states in the membrane domain of Complex I with the US and MWE methods. (A,B,C) Protonation probability of the titratable residues as a function of the collective variable for US, MWE, and MWE/FUS, respectively. (D,E,F) Protonation state PMF profile of His319 (top) and His293 (bottom) for US, MWE, and MWE/FUS, respectively.

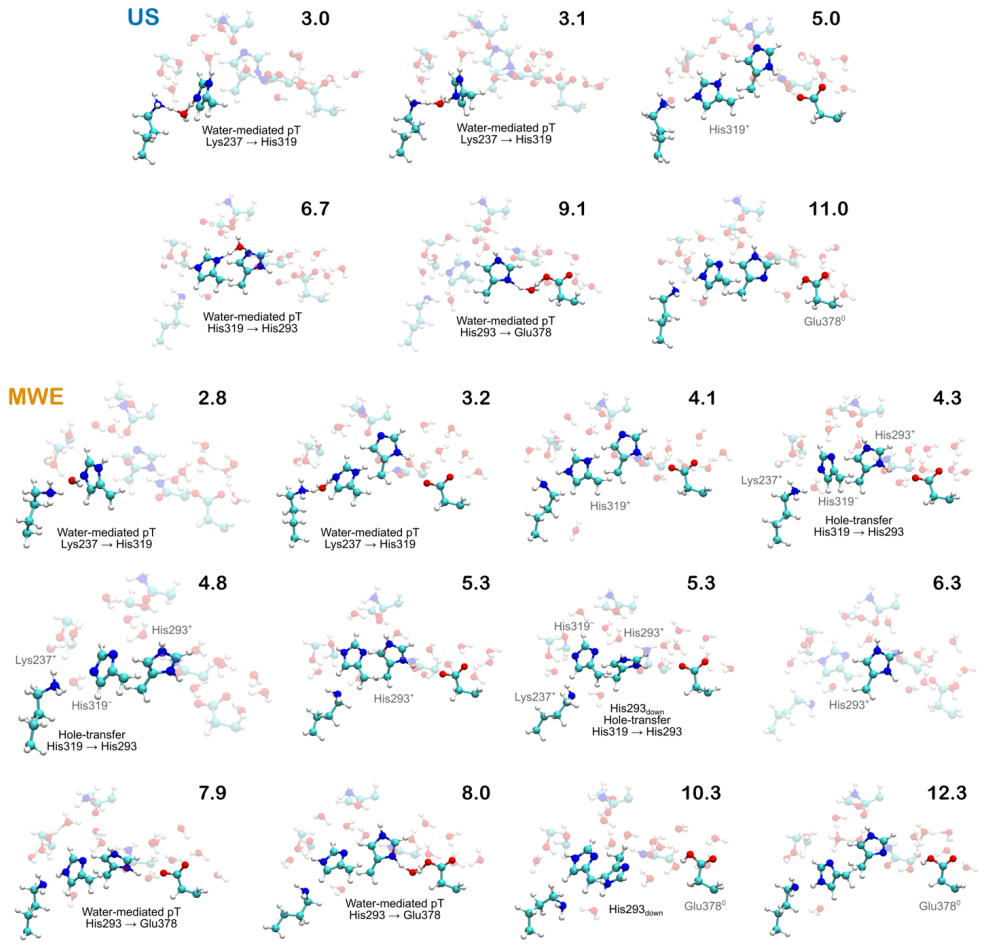


Figure S9. Intermediate structures sampled during proton transfer in Complex I from US and MWE. RC values (in Å) are indicated for each conformation.

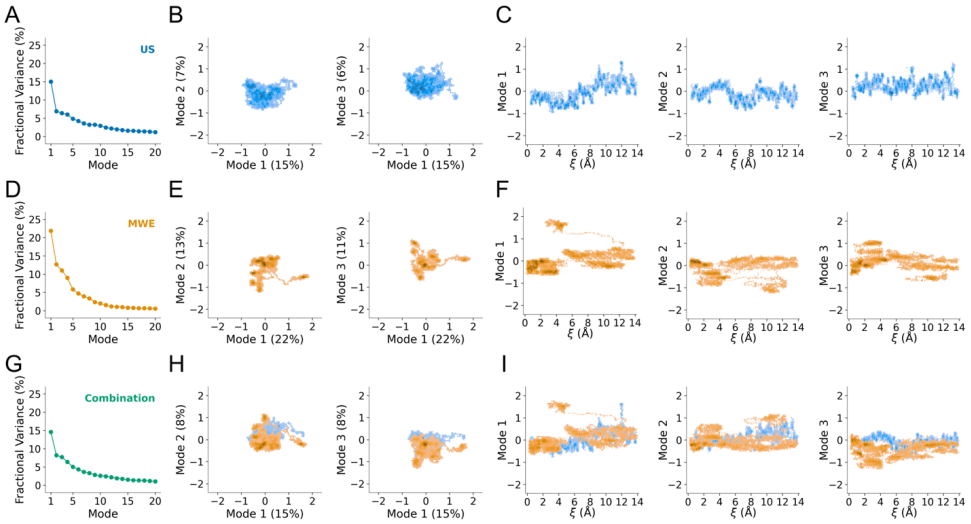


Figure S10. Essential dynamics analysis of proton transfer reactions sampled in the membrane domain of Complex I with the modified center of excess charge reaction coordinate using (A-C) US, (D-F) MWE, (G-I) combination of the data from US and MWE. (A, D, G) Fractional variances of essential dynamics analysis of the QM region. (B, E, H) Projection of the trajectory plotted onto principal components 1 and 2 (left) and 1 and 3 (right). (C, F, I) Projection of the trajectory onto principal components 1/2/3 as a function of collective variable.

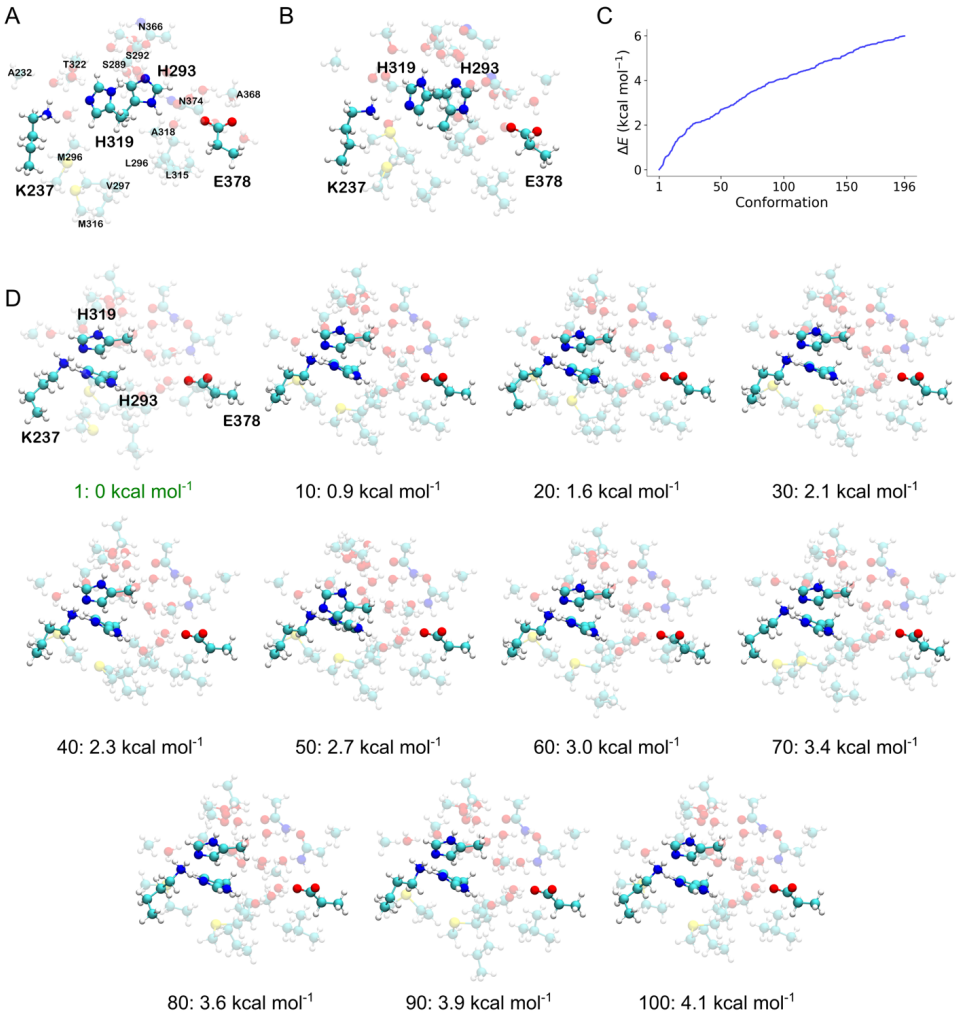


Figure S11. Conformer-rotamer search obtained in CREST sampling for the Complex I model. (A) Input conformation and residue numbering. (B) Structure obtained after optimization at the GFN2-xTB level. (C) Relative energies of the 196 conformers found below the threshold of 6 kcal mol^{-1}. (D) Every 10th conformer from the initial 100 conformers, labeled with their respective energy differences in comparison to the lowest energy conformer. The lowest energy conformer is highlighted in green.

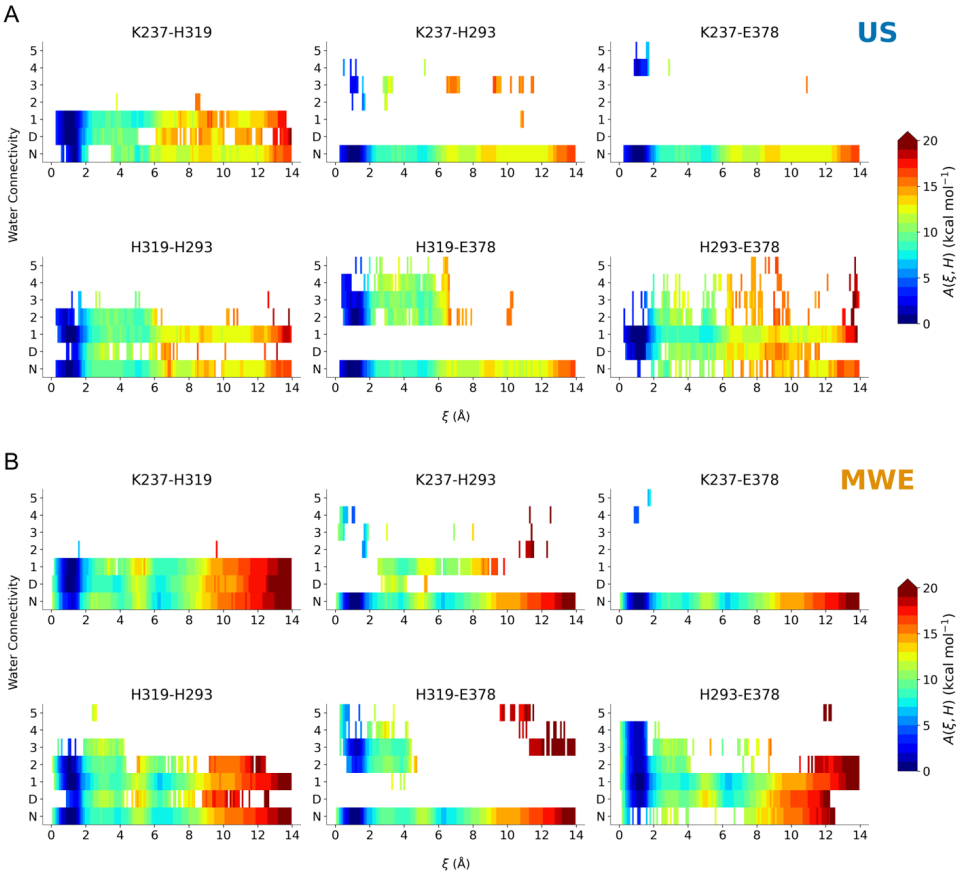


Figure S12. Sampled hydration states during proton transfer in the Complex I model as a function of mCEC-CV value. *N* stands for no water-mediated connection between two residues, *D* stands for a direct connection between two residues. (A) Water connectivity sampled with US. (B) Water connectivity sampled with MWE.

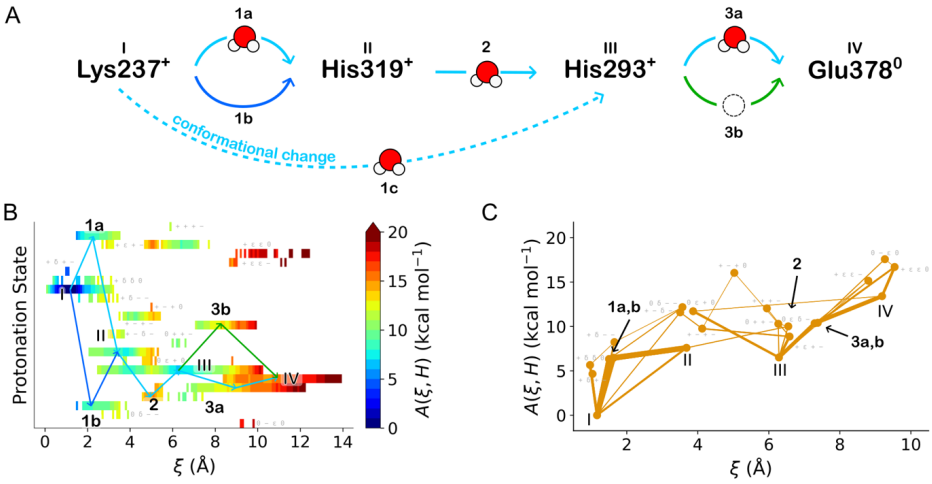


Figure S13. Sampling of protonation states during proton transfer in the Complex I model. (A) Reaction mechanism scheme sampled by MWE. Water-mediated reaction steps are denoted by a cartoon water molecule, hole transfer is denoted by a dashed circle, direct protonation. Direct protonation is denoted by an unadorned arrow. (B) 2D-PMF along the mCEC-CV and the protonation states of the protein residues participating in the pT. Intermediates and arrows correspond to mechanism scheme in A. Additional protonation states are annotated in grey using the following scheme: protonation state of Lys237, His319, His293, and Glu378. (C) Protonation states plotted according to their weighted mCEC-CV average and their free energy. Transitions between two states during sampling are depicted as connections between the two respective states. Thicker lines denote more frequent transitions. Intermediate states are annotated using arrows.

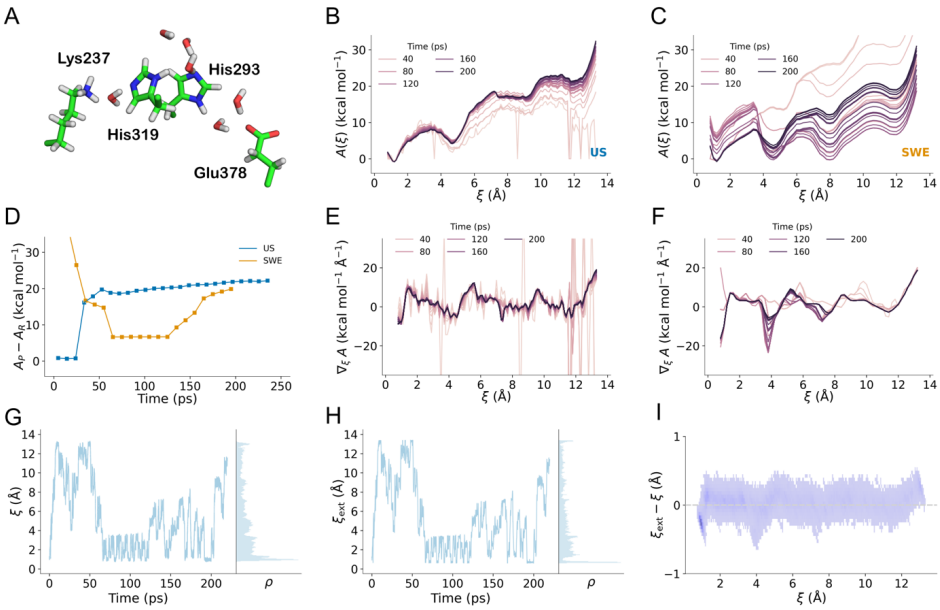


Figure S14. Sampling of proton transfer reactions in the Complex I modeled, using a smaller QM region and positional restraints. (A) Structure of the QM region. (B, C) Convergence of the PMF profile obtained with US and SW-WTM-eABF, respectively. (D) Convergence of PMF difference between reactant and product state. (E, F) Convergence of mean force profile obtained with US and SWE, respectively. (G) Sampling of the reaction coordinate using SWE. (H) Sampling of the extended variable reaction coordinate using SWE. (I) 2D histogram of the instantaneous difference between the extended system variable and the reaction coordinate as a function of the reaction coordinate.

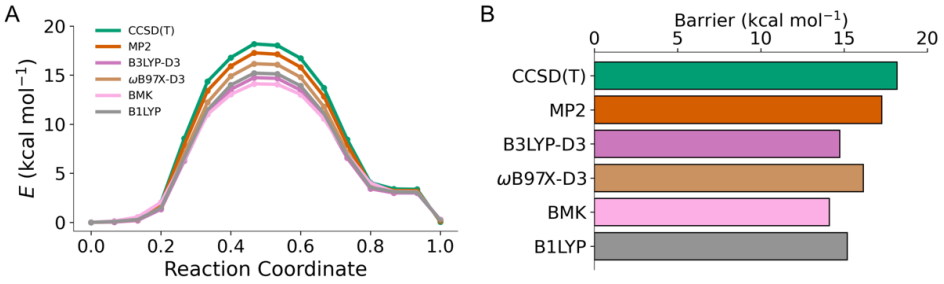


Figure S15. Benchmarking the proton transfer energetics for the His-water model (model 1). (A) Minimum energy profiles and (B) barriers at different theory levels. B3LYP-D3/def2-SVP geometries were used for estimation of electronic energies with the def2-TZVPPD basis set at the B3LYP-D3,[6-8] ωB97X-D3,[9] BMK,[10] B1LYP,[11] and compared against PNO-based CCSD(T) and MP2 level of theory.

Table S1. List of simulations.

Simulation	Model	Reaction Coordinate	Sampling Algorithm	Sampling Time (ps)			Total Sampling Time (ps)
1	1	LC	US	18	×	50	900
2			MWE	2	×	48	96
3		mCEC	US	33	×	20	660
4			MWE	2	×	50	100
5			MWE	10	×	30	300
6	2	mCEC	US	60	×	5	300
7			MWE	12	×	25	300
8			FUS	12	×	10	120
9	3	mCEC	US	48	×	5	240
10			SWE	1	×	220	220

Table S2. Apparent barriers and activation free energies for Model 1 for different switching parameter d_s.

d_s (Å)	$m_{z^\ddagger}$ (a.m.u.)	$\Delta A^\ddagger$ (kcal mol^{-1})	$A^\ddagger - A_R = \Delta A_{RP}^{\ddagger app}$ (kcal mol^{-1})
0.03	0.017	8.4	10.5
0.05	0.026	8.5	10.4
0.10	0.049	8.7	10.0
0.20	0.086	8.6	9.6

References

(1) Bakan, A.; Meireles, L. M.; Bahar, I. ProDy: Protein Dynamics Inferred from Theory and Experiments. *Bioinformatics* **2011**, *27* (11), 1575-1577. DOI: 10.1093/bioinformatics/btr168.
(2) Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. Essential dynamics of proteins. *Proteins: Structure, Function, and Bioinformatics* **1993**, *17* (4), 412-425. DOI: 10.1002/prot.340170408.
(3) Pracht, P.; Bohle, F.; Grimme, S. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Physical Chemistry Chemical Physics* **2020**, *22* (14), 7169-7192. DOI: 10.1039/C9CP06869D.
(4) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *Journal of Chemical Theory and Computation* **2019**, *15* (3), 1652-1671. DOI: 10.1021/acs.jctc.8b01176.
(5) Dietschreit, J. C. B.; Diestler, D. J.; Hulm, A.; Ochsenfeld, C.; Gómez-Bombarelli, R. From free-energy profiles to activation free energies. *The Journal of Chemical Physics* **2022**, *157* (8). DOI: 10.1063/5.0102075.
(6) Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical Review A* **1988**, *38* (6), 3098-3100. DOI: 10.1103/PhysRevA.38.3098.
(7) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **1988**, *37* (2), 785-789. DOI: 10.1103/physrevb.37.785.
(8) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132* (15), 154104. DOI: 10.1063/1.3382344.
(9) Chai, J.-D.; Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Physical Chemistry Chemical Physics* **2008**, *10* (44), 6615-6620. DOI: 10.1039/B810189B.
(10) Boese, A. D.; Martin, J. M. L. Development of density functionals for thermochemical kinetics. *The Journal of Chemical Physics* **2004**, *121* (8), 3405-3416. DOI: 10.1063/1.1774975.
(11) Adamo, C.; Barone, V. Toward reliable adiabatic connection models free from adjustable parameters. *Chemical Physics Letters* **1997**, *274* (1), 242-250. DOI: 10.1016/S0009-2614(97)00651-9.

3.5 Publication V: Combining Fast Exploration With Accurate Reweighting In the OPES-eABF Hybrid Sampling Method

Abstract: On-the-fly probability enhanced sampling (OPES) has recently been introduced [Invernizzi, M.; Parrinello, M., *J. Chem. Theory Comput.* **2022**, *18*, 3988–3996], with important improvements over the highly popular metadynamics methods. In our work, we introduce a new combination of OPES with the extended-system adaptive biasing force (eABF) method. We show that the resulting OPES-eABF hybrid is highly robust to the choice of input parameters, while ensuring faster exploration of configuration space than the original OPES. The only critical parameter of OPES-eABF is the coupling width to the extended-system, for which we introduce an automatic algorithm based on a short initial unbiased simulation, such that the OPES-eABF requires minimal user intervention. Additionally, we show that due to the decoupling of the physical system from the time-dependent potential, unbiased probabilities of visited configurations are recovered highly accurately.

Reprinted with permission from

A. Hulm; R. P. Schiller; C. Ochsenfeld. "Combining Fast Exploration With Accurate Reweighting In the OPES-eABF Hybrid Sampling Method" *J. Chem. Theory Comput.* **2025**, 21, 6434–6445.
URL: `https://doi.org/10.1021/acs.jctc.5c00395`.

JCTC Journal of Chemical Theory and Computation

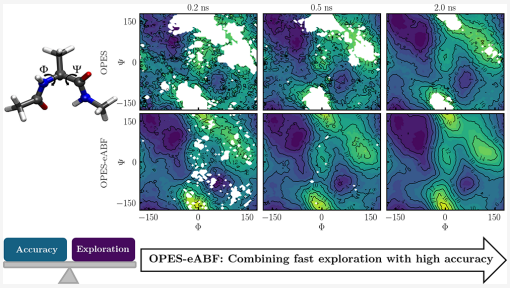
pubs.acs.org/JCTC Article

Combining Fast Exploration with Accurate Reweighting in the OPES-eABF Hybrid Sampling Method

Andreas Hulm, Robert P. Schiller, and Christian Ochsenfeld*

Cite This: J. Chem. Theory Comput. 2025, 21, 6434−6445

ABSTRACT: On-the-fly probability enhanced sampling (OPES) has recently been introduced [Invernizzi, M.; Parrinello, M. *J. Chem. Theory Comput.* **2022**, *18*, 3988−3996], with important improvements over the highly popular metadynamics methods. In our work, we introduce a new combination of OPES with the extended-system adaptive biasing force (eABF) method. We show that the resulting OPES-eABF hybrid is highly robust to the choice of input parameters, while ensuring faster exploration of configuration space than the original OPES. The only critical parameter of OPES-eABF is the coupling width to the extended-system, for which we introduce an automatic algorithm based on a short initial unbiased simulation, such that OPES-eABF requires minimal user intervention. Additionally, we show that due to the decoupling of the physical system from the time-dependent potential, unbiased probabilities of visited configurations are recovered highly accurately.



INTRODUCTION

Molecular dynamics (MD) has become a most powerful tool for the characterization of many-particle systems, and is able to provide significant insights into the dynamics of various chemical systems, ranging from biological macromolecules such as enzymes or nucleic acids[1−3] to solid state systems.[4−6] However, as many processes only occur on macroscopic time scales that are out of reach for conventional MD, importance sampling algorithms must be applied, that are able to selectively accelerate rare transitions.[7,8]

For this purpose, highly successful techniques have been developed over the years, many of which are based on biasing potentials which are applied to low dimensional collective variables (CVs) that represent reaction coordinates. Such approaches trace back to umbrella sampling (US), where simulations are encouraged to visit high free-energy regions with static biasing potentials.[9,10] Today, adaptive potential methods like the (well-tempered) metadynamics (WTM) and its variants are highly popular, which use a memory kernel to build suitable biasing potentials based on information on the trajectory and encourage the exploration of undersampled regions of CV space.[11−13] Recently, on-the-fly probability enhanced sampling (OPES) emerged, significantly improving upon WTM by shifting the focus from iteratively estimating the potential of mean force (PMF) (i.e., free energy surface), to directly reconstructing the underlying probability distribution based on a kernel density estimation.[14]

In contrast to WTM, OPES very quickly converges to a quasi-static biasing potential. This entails two major advantages: First, the reweighting of the configurations to the unbiased probability distribution can be done in the same way as for US. The reweighting of WTM is more complicated because the time dependence of the bias potential must be taken into account.[15] Second, OPES eliminates the danger of pushing the system into high free-energy transitions by too harsh repulsion from the current state, as it can occur in WTM. However, the downside of this is that for the case of imperfect CVs, transitions can still be relatively slow compared to WTM.[16] For this reason, a complementary variant of OPES was introduced, termed OPES explore ($OPES_E$), where the biasing potential is designed to continuously change and push the system out of equilibrium.[17] Naturally, from $OPES_E$ one cannot obtain accurate equilibrium properties.

An alternative to adaptive biasing potential based methods is the adaptive biasing force (ABF), where instead of the PMF one obtains an on-the-fly estimate of the mean force acting along CVs, which is integrated to obtain the PMF.[18,19] By applying the negative of this force estimate as a bias one aims for uniform sampling. To circumvent the associated strict technical requirements on CVs, the ABF is commonly applied

Received: March 9, 2025
Revised: May 15, 2025
Accepted: May 29, 2025
Published: June 18, 2025

https://doi.org/10.1021/acs.jctc.5c00395
J. Chem. Theory Comput. 2025, 21, 6434−6445

Journal of Chemical Theory and Computation pubs.acs.org/JCTC Article

to fictitious particles that are coupled to CVs, which yields the extended-system ABF (eABF).[20−22] Even more, the eABF enables remarkable algorithmic flexibility that can be harnessed to obtain highly efficient hybrid methods like the WTM-eABF,[23,24] or to accurately retrieve probabilities of the sampled states based on the multistate Bennett acceptance ratio (MBAR) estimator,[25,26] making use of the decoupling of the physical system from time-dependent biasing potentials and forces.

In this contribution, we introduce a new OPES-eABF hybrid method, that combines many of the advantages of its predecessors. On three test examples, namely the asymmetric double-well potential, the Müller−Brown potential, and the alanine dipeptide system in vacuum, we show that OPES-eABF is highly robust to the choice of parameters as well as CVs, provides fast sampling without pushing the system to high-energy transitions, and preserves accurate reweighting from the extended-system trajectories. Additionally, we provide an algorithm to automatically obtain suitable coupling width for the extended-system from short unbiased MDs, which is the most critical parameter for eABF, WTM-eABF, and OPES-eABF. Overall, we show that OPES-eABF can serve as unified tool for diverse sampling problems.

THEORY

We consider the classical dynamics of an N particle systems in $\mathbb{R}^3$ with $3N$ Cartesian coordinates $\mathbf{x}^T = (x_1, \cdots, x_{3N})$, momenta $\mathbf{p}^T = (p_1, \cdots, p_{3N})$ and potential energy $U(\mathbf{x})$, whose probability density follows the Boltzmann-distribution

$$\rho(\mathbf{x}) = \frac{e^{-\beta U(\mathbf{x})}}{\int e^{-\beta U(\mathbf{x})} d\mathbf{x}} = \frac{1}{Z} e^{-\beta U(\mathbf{x})} \tag{1}$$

with inverse temperature $\beta = (k_B T)^{-1}$, Boltzmann constant k_B, and configurational integral Z. Ensemble averages, indicated by $\langle \cdots \rangle$, of any observable $O(\mathbf{x})$ are given by

$$\langle O(\mathbf{x}) \rangle = \int O(\mathbf{x}) \rho(\mathbf{x}) d\mathbf{x} \tag{2}$$

In principle, estimates for $\rho(\mathbf{x})$ can be obtained from molecular dynamics (MD) or Monte Carlo (MC) simulations under the assumption of ergodicity, which states that trajectory averages will eventually converge to the ensemble average. However, from eq 1 it is clear that visiting configurations with higher potential energy is exponentially less likely, such that simulations are often trapped in certain regions of configuration space. For this reason conventional MD simulations are quasi-nonergodic, and accurate estimates of $\rho(\mathbf{x})$ are hard to compute. Importance sampling techniques aim to circumvent this problem by modifying the potential energy with some biasing potential

$$\tilde{U}(\mathbf{x}) = U(\mathbf{x}) + U^{\text{bias}}(\mathbf{x}) \tag{3}$$

such that energy barriers are reduced and ergodicity can at least partially be restored. The modified probability density reads

$$\tilde{\rho}(\mathbf{x}) = \frac{e^{-\beta(U(\mathbf{x}) + U^{\text{bias}}(\mathbf{x}))}}{\int e^{-\beta(U(\mathbf{x}) + U^{\text{bias}}(\mathbf{x}))} d\mathbf{x}} = \rho(\mathbf{x}) \frac{Z}{\tilde{Z}} e^{-\beta U^{\text{bias}}(\mathbf{x})} \tag{4}$$

where $\tilde{Z}$ is the modified configurational integral. By insertion into eq 2, biased ensemble averages can be reweighted to the physical distribution via

$$\langle O(\mathbf{x}) \rangle = \int O(\mathbf{x}) \tilde{\rho}(\mathbf{x}) \frac{\tilde{Z}}{Z} e^{\beta U^{\text{bias}}(\mathbf{x})} d\mathbf{x} = \frac{\langle O(\mathbf{x}) e^{\beta U^{\text{bias}}(\mathbf{x})} \rangle_{\tilde{U}}}{\langle e^{\beta U^{\text{bias}}(\mathbf{x})} \rangle_{\tilde{U}}} \tag{5}$$

where $\langle \cdots \rangle_{\tilde{U}}$ denotes averages over the biased ensemble.

Most importance sampling techniques rely on the definition of collective variables (CVs), which are functions $\xi: \mathbb{R}^{3N} \to \mathbb{R}^d$ with $d \ll N$ that map high dimensional systems onto a low dimensional representation $\mathbf{z} = (z_1, z_2, \cdots, z_d) = \xi(\mathbf{x})$ which serves as reaction coordinate. The marginal probability distribution along $\mathbf{z}$ is defined via

$$\rho(\mathbf{z}) = \int \delta[\mathbf{z} - \xi(\mathbf{x})] \rho(\mathbf{x}) d\mathbf{x} = \langle \delta[\mathbf{z} - \xi(\mathbf{x})] \rangle \tag{6}$$

with multivariant Dirac delta distribution $\delta[\cdots]$. From $\rho(\mathbf{z})$ one obtains the potential of mean force (PMF), or free-energy surface

$$A(\mathbf{z}) = -\beta^{-1} \ln \rho(\mathbf{z}) \tag{7}$$

which is the main target of most importance sampling algorithms, as it allows for the calculation of the reaction and activation free energy of the underlying process. Assuming that the CV space provides a good separation of two metastable states A and B, the reaction free energy can simply be obtained from

$$\Delta A_{A \to B} = -\beta^{-1} \ln \frac{\int_B \rho(\mathbf{z}) d\mathbf{z}}{\int_A \rho(\mathbf{z}) d\mathbf{z}} = -\beta^{-1} \ln \frac{Z_B}{Z_A} \tag{8}$$

integrating over the corresponding domains of CV space.[27]

More involved is the calculation of the activation free energy, for which recently an analytical expression was provided by Dietschreit et al.[28]

$$\Delta A^{\ddagger}_{A \to B} = -\beta^{-1} \ln \frac{Z_{TS} \langle \lambda_\xi \rangle_{z^\ddagger}}{Z_A} \tag{9}$$

Here, $z^\ddagger$ denotes the position of the transition state ensemble (TSE) of a one-dimensional CV and $\lambda_\xi = \sqrt{\beta h^2 / 2\pi m_\xi}$ with Planck constant h is the thermal de-Broglie wavelength of the pseudoparticle associated with this CV

$$m_\xi^{-1} = (\nabla_{\mathbf{x}} \xi)^T \mathbf{M}^{-1} (\nabla_{\mathbf{x}} \xi) \tag{10}$$

where $\mathbf{M}$ is the diagonal mass matrix. Besides accounting for the mass of atoms involved in the transition, for example reproducing isotope effects, $\langle \lambda_\xi \rangle$ removes distortions of the Cartesian space by nonlinear CVs. Note that instead in the literature the simple difference of maxima and minima on the PMF is frequently employed, which can be seen as approximation of eq 9, ignoring distortions of the coordinate system and the influence of mass while assuming the probability density to be sharply peaked in the reactant minimum. Reaction rate constants, which are of high interest as they are often experimentally observable, can be obtained from Eyring's equation

$$k_{A \to B} = \frac{1}{\tau_{A \to B}} = \frac{\kappa}{\beta h} e^{-\beta \Delta A^{\ddagger}_{A \to B}} \tag{11}$$

where $\tau_{A \to B}$ is the first passage time from metastable state A to B and κ the transmission coefficient, which is often taken to be one assuming that all trajectories that reach the TS also cross

https://doi.org/10.1021/acs.jctc.5c00395
J. Chem. Theory Comput. 2025, 21, 6434−6445

it. In practice, it is assumed that $Z_{TS} = \rho(z^{\ddagger})$, which is the reason for the high sensitivity of results on the choice of CV, as it needs to correctly capture the TSE. Alternatively, there are a number of methods for the estimation of kinetic rates that are based on directly observing the first passage time from multiple simulations, avoiding the computation of TSEs. The underlying assumption is, that the TSE remains bias free, which means that only Z_A is modified in eq 9. This allows for the rescaling of biased passage times $\tilde{\tau}_{A\to B}$ via

$$\tau_{A\to B} = \int_0^{\tilde{\tau}_{A\to B}} e^{\beta U^{bias}(\mathbf{x}(t))} dt \tag{12}$$

However, such simulations do not allow for the simultaneous computation of the $\Delta A_{A\to B}$ as the PMF is never computed. Hence, in this contribution we want to analyze the convergence of both eq 8 and 9, filling a gap in the importance sampling literature that often exclusively focuses on the $\Delta A_{A\to B}$.

Adaptive Biasing Potential Methods. Adaptive potential methods aim to learn a good biasing function on-the-fly from information from the trajectory. Although many strategies are reported in literature, the today by far most influential is (well-tempered) metadynamics (WTM).[11,12,29] Here, a time-dependent repulsive potential is built from the superposition of Gaussian hills

$$G(\mathbf{z}, \mathbf{z}_t) = h_G e^{-(\mathbf{z}-\mathbf{z}_t)^2/2\sigma_G^2} \tag{13}$$

with Gaussian height h_G, standard deviation $\sigma_G = (\sigma_1, \cdots, \sigma_d)$, and Gaussian center $\mathbf{z}_t$. New hills are created in fixed time intervals τ_G according to

$$U^{WTM}(\mathbf{z}, t) = \sum_{t=0,\tau_G,2\tau_G,\cdots} e^{-\beta U^{WTM}(\mathbf{z},t-1)/(\gamma-1)} G(\mathbf{z}, \mathbf{z}_t) \tag{14}$$

where to ensure smooth convergence, the height of Gaussian hills is scaled down depending on the previously deposited bias, such that bias factor $\gamma > 1$ controls how much the original distribution is smoothed out.[12] Note that for $\gamma \to \infty$ the scaling factor is 1 and the Gaussian height remains constant, such that conventional metadynamics (MtD) is recovered. It can be mathematically proven,[30] that the WTM potential converges to

$$\begin{aligned} U^{WTM}(\mathbf{z}, t) &= -\left(1 - \frac{1}{\gamma}\right) A(\mathbf{z}) + \beta^{-1} \ln \frac{\int d\mathbf{z} e^{-\beta A(\mathbf{z})}}{\int d\mathbf{z} e^{-\beta(A(\mathbf{z}) + U^{WTM}(\mathbf{z},t))}} \\ &= -\left(1 - \frac{1}{\gamma}\right) A(\mathbf{z}) + C(t) \end{aligned} \tag{15}$$

such that an unbiased estimate of the PMF can directly be obtained from U^{WTM}. Here, $C(t)$ is a time-dependent constant, which can be ignored for estimating the PMF, as we are only interested in relative free energies, but is important for proper reweighting.[15] Most importantly, after an initial transient the time dependencies of $U^{WTM}(\mathbf{z},t)$ and $C(t)$ cancel, such that a time-independent statistical estimator is given by

$$\langle O(\mathbf{x}) \rangle = \langle O(\mathbf{x}) e^{\beta(U^{WTM}(\mathbf{z},t) - C(t))} \rangle_{\tilde{U}} \tag{16}$$

Recently, on-the-fly probability enhanced sampling (OPES) emerged as a new alternative to WTM, shifting the focus from estimating the biasing potential to directly estimating the underlying probability density.[14,17] The estimate of the probability density is obtained from the kernel density estimation

$$\tilde{\rho}(\mathbf{z}, t) = \frac{\sum_{t=\tau_G,2\tau_G,\cdots} w_t G(\mathbf{z}, \mathbf{z}_t)}{\sum_{t=\tau_G,2\tau_G,\cdots} w_t} \tag{17}$$

with $w_t = e^{\beta U^{OPES}(\mathbf{z}_t, t-1)}$. Note that unlike for WTM the height of Gaussians is given by $h_G = \prod_i (\sigma_i \sqrt{2\pi})^{-1}$ and changing it would only lead to a change in the normalization. The biasing potential is given by

$$U^{OPES}(\mathbf{z}, t) = \left(1 - \frac{1}{\gamma}\right) \beta^{-1} \log\left(\frac{\tilde{\rho}(\mathbf{z}, t)}{Z_t} + \epsilon\right) \tag{18}$$

where Z_t is a norm factor, normalizing over the explored space $|\Omega_t|$ via

$$Z_t = \frac{1}{|\Omega_t|} \int_{\Omega_t} \tilde{\rho}(\mathbf{z}, t) d\mathbf{z} \tag{19}$$

and ϵ ensures that the logarithm is always defined. By choosing $\epsilon = e^{-\beta \Delta E/(1-1/\gamma)}$, OPES allows for setting an upper bound to the biasing potential, which is given by the parameter ΔE. Note that in contrast to WTM, where the biasing potential builds up slowly, already at the beginning of the simulation U^{OPES} is in the order of ΔE. Therefore, one obtains fast initial transitions after which the bias quickly becomes quasi-static and the details of the PMF are slowly refined. As opposed to WTM, reweighting is simply possible via eq 5 and much more stable, especially for the initial part of the trajectory where U^{WTM} would still change drastically.[14]

Extended-System Dynamics. We use the term *extended-system* for methods in which artificial particles are coupled to CVs. These additional degrees of freedom are subject to the same dynamics as the physical system and serve as a proxy for the application of adaptive potential methods.[20,22−24] The full extended-system potential thus reads

$$\begin{aligned} U^{ext}(\mathbf{x}, \lambda) = U(\mathbf{x}) &+ \sum_{i=1}^{d} \frac{1}{2\beta\sigma_{ext,i}^2} (\xi_i(\mathbf{x}) - \lambda_i)^2 \\ &+ U^{bias}(\lambda_1, \cdots, \lambda_d, t) \end{aligned} \tag{20}$$

where fictitious particles $(\lambda_1, \cdots, \lambda_d)$ are tightly coupled to CVs with coupling widths $\sigma_{ext,i} = \sqrt{1/(\beta k_i)}$, k_i being harmonic force constants. Thus, the physical system just experiences the time-independent harmonic coupling potential, and any time-dependent biasing potential $U^{bias}(\lambda_1, \cdots, \lambda_d, t)$ can be applied to fictitious particles to accelerate sampling. An especially efficient method has emerged by combining two complementary biasing strategies in the WTM-eABF hybrid method, where WTM pushes the system away from the already sampled states and an adaptive biasing force (ABF) removes barriers along the way.[23,24] In spirit of the WTM-eABF, we introduce a new and most efficient variant, replacing the WTM with the more recent OPES to yield the OPES-eABF hybrid method.

For reweighting only information from the trajectories of CVs and λ's is required. This means that results are independent of the convergence of U^{bias}, making results robust against the choice of parameters of the chosen sampling accelerator (WTM, OPES, and/or ABF). Previously, we have shown that accurate reweighting is always possible using standard importance sampling techniques like the MBAR,[25,26]

https://doi.org/10.1021/acs.jctc.5c00395
J. Chem. Theory Comput. 2025, 21, 6434−6445

while fast on-the-fly estimates of the PMF are available based on thermodynamic integration.[22,31] Overall, because of these properties we are convinced that the combination of OPES with extended-system dynamics provides an ideal foundation for the development of a unified importance sampling algorithm, as outlined below.

COMPUTATIONAL DETAILS

Implementation of the OPES and OPES-eABF Method. A python based implementation of the OPES method is provided in our in-house adaptive-sampling package,[26,32] following the original implementation of Invernizzi and Parrinello, which includes a kernel compression algorithm, a shrinking bandwidth to converge details of the PMF, and an algorithm for the efficient numeric computation of the norm factor (eq 19).[14] We provide the option to store biasing potentials and forces on a grid instead of computing the sum of kernels in every step. Note that while this is highly efficient in low dimensional CV spaces it may become disadvantageous in higher dimensional spaces because of the "curse of dimensionality", which is elegantly circumvented by the kernel compression algorithm of OPES. If no bias factor γ is provided, it is set to $\gamma = \beta\Delta E$.[14] An adaptive bandwidth algorithm is implemented based on an exponentially decaying average using Welford's online algorithm with decay time τ.[17,33] Similarly, the initial kernel bandwidth can be obtained from a short unbiased simulation over τ steps. Hence, the only parameters that have to directly be set are the frequency of kernel creation and the barrier factor ΔE. Although not further discussed here, we also provide an implementation of the explore variant of OPES (OPES_E).[17]

We leveraged the extended-system formalism to develop a modular approach for OPES-eABF (respectively, OPES_E), analogous to the WTM-eABF.[23,24] The fictitious particle is propagated using a Langevin integrator[34] identical to that of the physical system. It experiences a combined bias composed of OPES and/or ABF, which are evaluated on the same grid. For ABF, N_full is the only input parameter, which describes a linear ramp function that controls how fast the ABF force is scaled up. To set up the extended-system one has to choose two empirical parameters for each CV: masses m_{λ_i} of fictitious particles (or the oscillator periods $\tau_i = 2\pi\sqrt{m_{\lambda_i}/k_i}$), which have only a marginal influence on results, and coupling widths $\sigma_{\text{ext},i}$.[22] The latter are critical, as too loose coupling will result in insufficient sampling of the physical system, which does not remain coupled, and too tight coupling will hinder convergence. The parameters of U^OPES, respectively U^WTM, are not as critical, as they do not enter reweighting and only control how fast the CV space is explored. Therefore, we aim to obtain an automatic algorithm for the estimation of suitable coupling widths $\sigma_{\text{ext},i}$ to make the methods as easy to use as their non-extended counterparts. To motivate our approach, we start by approximating the probability density in metastable state A with a Gaussian

$$\tilde{p}_\text{A}(\mathbf{z}) \propto \mathrm{e}^{-\frac{1}{2\sigma_\text{A}^2}(\mathbf{z}-\mathbf{z}_0)^2} \tag{21}$$

with equilibrium position $\mathbf{z}_0$ and standard deviation σ_A. Inserting into the definition of the PMF (eq 7), we obtain

$$\tilde{A}(\mathbf{z}) = -\beta^{-1}\ln(\mathrm{e}^{-\frac{1}{2\sigma_\text{A}^2}(\mathbf{z}-\mathbf{z}_0)^2}) = -\frac{1}{2\beta\sigma_\text{A}^2}(\mathbf{z}-\mathbf{z}_0)^2 \tag{22}$$

To ensure tight coupling as long as $(\mathbf{z}-\mathbf{z}_0) \sim (\mathbf{z}-\boldsymbol{\lambda})$, we require the force along the PMF $-\frac{\partial}{\partial\mathbf{z}}\tilde{A}(\mathbf{z})$ to be less than the coupling force,

$$\frac{1}{\beta\sigma_\text{A}^2}(\mathbf{z}-\mathbf{z}_0) < \frac{1}{\beta\sigma_\text{ext}^2}(\mathbf{z}-\lambda) \tag{23}$$

from which we obtain the condition $\sigma_\text{ext} < \sigma_\text{A}$, serving as an upper bound for the choice of coupling width. This follows from the intuition that for $(\mathbf{z}-\mathbf{z}_0) \gg (\mathbf{z}-\boldsymbol{\lambda})$ eq 21 does not hold, and instead, as desired, one escapes to another metastable state. Hence, a suitable σ_ext can be found by estimating σ_A from a short initial MD and scaling by a factor smaller than 1 (0.5 is used throughout this work, unless otherwise noted). Note that for this the probability density is assumed to be Gaussian-shaped, which is accurate for a wide range of processes, e.g., chemical reactions that require bond breaking, but is not always the case. Starting from a diffuse, non-Gaussian-shaped state may lead to estimates of σ_ext that are too loose. Hence, for multiple states (A, B, ···), we propose to use the minimum of all (σ_A, σ_B, ···) to ensure tight coupling in all states. For open exploration runs, where intermediate metastable states are not known beforehand, it is advisable to choose a more tight coupling width, either by reducing the scaling factor of the automatic algorithm, e.g., to 0.1, or by manually setting a tight parameter. As discussed further below, this slightly reduces the speed of convergence by introducing more noise in the coupling force, but ensures that the system is always able to efficiently escape from different metastable states. Further, the algorithm extends to multidimensional CVs, by separately obtaining σ_ext for all degrees of freedom.

Numerical Potentials. As simple test cases, we consider the Langevin dynamics of a single particle on numerical 2D potentials. We start with the asymmetric double-well (ADW) potential

$$U^\text{ADW}(x, y) = ax^2 - bx^3 + cx^4 + dy^2 + e \tag{24}$$

with the empirical parameters given in the Supporting Information (SI). MD simulations are initialized at the global minimum according to Boltzmann statistics and run for 500 ps with 1 fs time step (500,000 steps) and a friction constant of 1 ps^{-1} at an equilibrium temperature of 300 K. Unless otherwise noted, WTM potentials are updated every 100th step with initial hill height 0.239 kcal/mol (1 kJ/mol), kernel standard deviation 0.07, and bias factor 15 (which is equivalent to a bias temperature of 4200 K). For OPES, comparable parameters are used, but hills are only created every 500 steps, as due to the normalization for OPES the pace of hill creation has no influence on how fast the biasing potential grows. The barrier factor ΔE is set to 30.4 kcal/mol, which corresponds to the analytical barrier for the forward reaction starting from the global minimum. The mass of fictitious particles for extended-systems is set to 20 a.u., and the coupling width is obtained from 5000 steps of unbiased MD with a scaling factor of 0.5, unless otherwise noted. The ABF force is scaled up with a linear ramp and fully applied in bins with at least 500 samples. The x-axis is always employed as CV, such that the reference PMF is given by numerical integration of the analytical probability density along the y-axis.

As a second test case, we consider the Müller–Brown potential

https://doi.org/10.1021/acs.jctc.5c00395
J. Chem. Theory Comput. 2025, 21, 6434−6445

$$U^{\mathrm{MB}}(x, y) = B \sum_{i=1}^{4} A_i \exp[\alpha_i (x - x_i)^2 + \beta_i (x - x_i)(y - y_i) + \gamma_i (y - y_i)^2] \quad (25)$$

with B = 1 kJ/mol, and all other numerical parameters given in the SI. Langevin dynamics simulations are performed as before, but for 10 ns. WTM potentials are updated every 500 steps with Gaussian hills of height 1.0 kJ/mol, standard deviation 0.1 and bias factor 15. For OPES, the barrier factor is set to 5 kcal/mol, always applying an adaptive bandwidth using a running average with a decay time of 5000 steps. Extended-systems are parameterized exactly like for the ADW potential. The ABF force is scaled up with a linear ramp and fully applied in bins with at least 500 samples. Again, the x-axis is employed as CV, computing the reference PMF by numerical integration of the analytical probability density along the y-axis. Furthermore, to test sampling along a good CV for the MB potential, path CVs (PCVs)[35,36] are employed, as described in our previous study[37] and detailed in the SI. Both potential energy surfaces are shown in Figure 1.

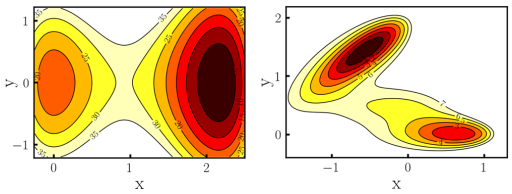


Figure 1. Asymmetric double-well (left) and Müller−Brown potential (right). Inline numbers denote the potential energy in kcal/mol.

Alanine Dipeptide. Simulations of alanine dipeptide in vacuum are performed using the OpenMM molecular dynamics library[38] and AMBER ff14SB parameters.[39] For this purpose, a python interface to the adaptive-sampling package is developed based on the CustomExternalForce module of OpenMM. The Langevin integrator as implemented in OpenMM is employed at 300 K with a time step of 2 fs and damping of 1 ps^{-1}, keeping covalent bond distances of hydrogen atoms constrained. For each sampling method, 11 independent simulations are performed, 10 ns each. For WTM and OPES hills are created every 500 steps. An initial hill height of 1.2 kJ/mol, kernel standard deviation of 0.35 radians and bias factor 15 was used for WTM potentials. For OPES, the parameters of ref 14. are reproduced, such that results are comparable, and our implementation can be verified. Most importantly, the barrier factor is set to 50 kJ/mol (∼12 kcal/mol). The extended-system is initialized exactly like for numerical potentials and for eABF hybrids WTM and OPES parameters are identical to nonextended simulations. Again, for 1D simulations the ABF force is fully applied in bins with at least 500 samples. However, for 2D simulations this parameter is reduced to 100 samples. Additionally, for 2D simulations the mass of fictitious particles for the extended-system is increased to 200 a.u. To ensure tight coupling it is better to estimate σ_{ext} from the $C7_{\mathrm{ax}}$ state. Estimating σ_{ext} from the non-Gaussian shaped and more diffuse $C7_{\mathrm{eq}}$ state results in only loose coupling, as discussed in the Implementation of the OPES and OPES-eABF Method section. In Figure 2 on the left, the alanine dipeptide molecule is shown, where the Φ and Ψ dihedral angles are marked, which are employed as CVs. On the right, a reference PMF is given, as obtained from the combined data of 10 independent 10 ns OPES simulations.

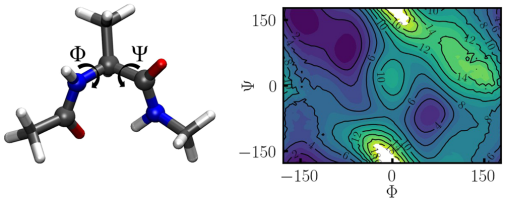


Figure 2. On the left, the $C7_{\mathrm{eq}}$ state of alanine dipeptide is shown. The ϕ and ψ dihedral are marked, which represent the most important CVs, involved atoms shown as balls. The graphic was Graphic created with VMD.[40] On the right, a reference PMF is shown as obtained from long OPES simulations. Inline numbers on contours denote energies in kcal/mol.

RESULTS AND DISCUSSION

In the following we discuss the performance of WTM, WTM-eABF, OPES, and OPES-eABF on three different scenarios: first, the ADW potential, representing a system with high ΔA, where an optimal reaction coordinate is given by the x-axis. Second, the MB potential, which is a prototypical example for a system with a more complicated nonlinear reaction coordinate, and last, the $C7_{\mathrm{eq}}$ to $C7_{\mathrm{ax}}$ transition of alanine dipeptide, which can undergo two competing reaction channels.

The Case of High ΔA: Asymmetric Double-Well Potential. We start by analyzing the performance of different sampling algorithms for a particle on an asymmetric double-well (ADW) potential where the x-axis represents an optimal CV. The potential energy surface is shown in Figure 1 on the left. In Figure 3a,b we compare the performance of WTM and WTM-eABF, respectively, where all parameters of WTM potentials are identical. On the left, the mean final PMFs, in the middle the convergence of the reaction free energy, and on the right the convergence of the activation free energy are shown, averaged over 11 simulations with standard deviations given by light areas. As it is well-known in the literature, the convergence of WTM strongly depends on its parameterization.[13] For example, since with smaller values of γ the hill height of new Gaussian's decreases faster, the WTM potential grows more slowly, as it is also evident in Figure 3a. Figure 3b shows that due to the additional ABF, WTM-eABF is much more robust against the choice of WTM parameters and all simulations converge already after about 20 ps to the analytic result with vanishing standard deviations. The application of extended-system dynamics with an ABF or WTM bias alone leads to dramatically reduced convergence (Figure S1), which underlines the advantage of combining two complementary biasing strategies. Overall, this demonstrates the robustness of the WTM-eABF method, for which the main critical parameter is the coupling width σ_{ext}.

Therefore, we aim to further simplify the process of setting up the extended-system by automatically obtaining a suitable σ_{ext} from the CVs standard deviation in a short initial MD. As discussed in the Computational Details section we use the CVs unbiased standard deviation as an upper bound to the coupling width, and to ensure tight coupling, propose to scale it with a factor smaller than 1 to obtain the final estimate of σ_{ext}. In Figure 4, we compare results for scaling factors 0.1 (blue), 0.5

https://doi.org/10.1021/acs.jctc.5c00395
J. Chem. Theory Comput. 2025, 21, 6434−6445

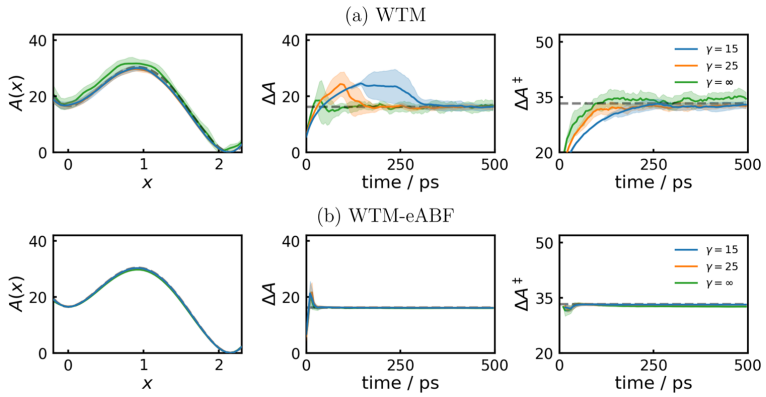


Figure 3. On the left the mean PMFs from 11 independent 500 ps MD simulations are shown, with standard deviations denoted by light areas. Additionally, the convergence behavior of ΔA (middle column) and $\Delta A^{\ddagger}$ (right) are shown. Dashed gray lines denote analytic results.

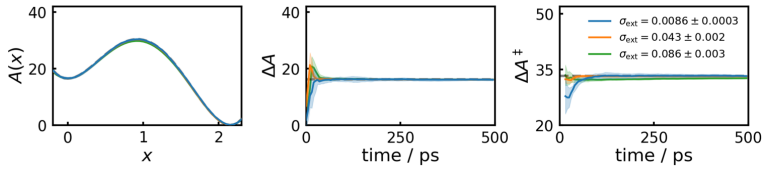


Figure 4. On the left the mean PMFs from 11 independent 500 ps WTM-eABF runs with different coupling widths are shown, with standard deviations denoted by light areas. Coupling widths were obtained from the CVs standard deviation from 5000 initial MD steps, and scaled with 0.1 (blue), 0.5 (orange), and 1.0 (green), respectively, mean and standard deviation of the resulting coupling width shown as inset. Additionally, the convergence behavior of ΔA (middle column) and $\Delta A^{\ddagger}$ (right) are shown. Dashed gray lines denote analytic results.

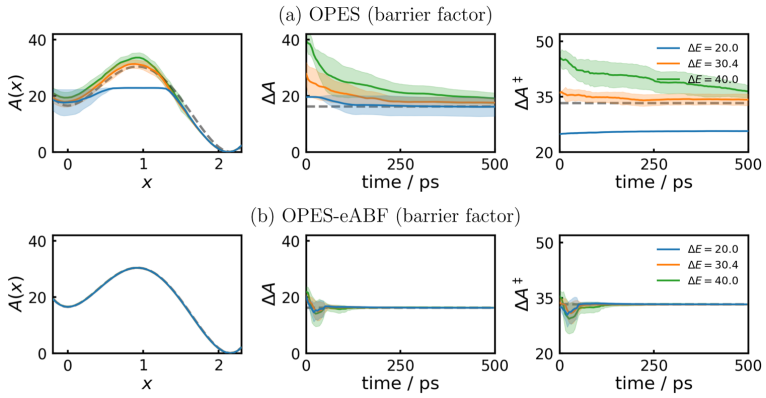


Figure 5. On the left, the mean PMFs from 11 independent 500 ps MD runs are shown, with standard deviations denoted by light areas. Additionally, the convergence behavior of ΔA (middle column) and $\Delta A^{\ddagger}$ (right) are shown. Dashed gray lines denote analytic results. Dashed gray lines denote analytic results.

(orange), and 1.0 (green). The resulting values for σ_{ext} obtained from 5000 initial MD steps are shown as inset on the right. For direct use of the unscaled σ_{ext} (green), the PMF at the transition state and the activation free energy are slightly underestimated, showing that the coupling is barely tight enough. Using scaling factors of 0.5 (orange) or 0.1 (blue), the PMF fully converges to the analytic result. However, with the smallest σ_{ext} the initial convergence begins to deteriorate as the coupling forces become larger and more noisy. Therefore, we always use a scaling factor of 0.5 in this work, which seems to provide a good balance between tight coupling and fast convergence. The broad validity of this choice for other systems is shown by its application to simulations of the MB potential and alanine dipeptide system further below.

We next switch from WTM to the more recent OPES. In Figure 5a,b we compare the performance of OPES (upper

https://doi.org/10.1021/acs.jctc.5c00395
J. Chem. Theory Comput. 2025, 21, 6434−6445

row) to the new OPES-eABF counterpart (lower row). We note that as discussed by Invernizzi and Parrinello,[14] OPES has only a small sensitivity on the choice of bias factor (see also Figure S2) as it does not influence the growth rate of the OPES potential. However, in OPES the barrier parameter ΔE is introduced, which should be chosen as the height of barriers one wishes to overcome as it sets roughly an upper bound to the OPES potential. If ΔE is chosen too small (blue), transitions are rare, as the difference to $\Delta A^{\ddagger}$ remains as effective barrier. On the other hand, choosing ΔE too high, the OPES biasing potential initially overshoots and convergence is slowed down as well. In orange, results are shown for setting ΔE to the analytic activation free energy. As expected, this enables almost instant convergence of $\Delta A^{\ddagger}$. However, because of the high analytic ΔA of the asymmetric double-well potential, convergence of ΔA is still relatively slow. This is because for the higher-energy state (left minimum) ΔE is chosen much too high, which leads to an initial overshooting of the OPES potential for this state. Again, as shown in Figure 5 we do not observe the same sensitivity to ΔE for OPES-eABF simulations, which always converge to the analytic result relatively quickly. For smaller values of ΔE convergence is reached a bit faster as here the OPES potential does not overshoot, quickly reaching a quasi-static regime, such that the ABF can gently remove the remaining barrier.

A similar but more severe effect arises for initializing simulations in the higher-energy state (local minimum). The reason is that after the first transition to the global minimum, due to the mechanism of OPES the effective barrier for the back transition is roughly given by $\Delta A + \Delta E$. Hence, since the bias potential is still limited to ΔE, simulations get trapped in the global minimum. Figure 6 demonstrates this effect by

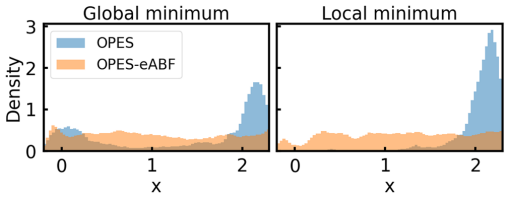


Figure 6. CV density plots for simulations initialized in the global or local minimum. Results from OPES are shown in blue and from OPES-eABF in orange. For all simulations the parameter ΔE is set to 30.4 kcal/mol.

showing obtained CV densities for simulations that start from the global (left of Figure 6), or local (right of Figure 6) minimum. As shown in blue on the left, only if the OPES simulations start from the global minimum, the biased CV density correctly converges to the well-tempered distribution defined by γ, showing two Gaussian-shaped distributions for the two metastable states. However, if simulations start from the higher-energy state, after the first transition, the global minimum is exclusively sampled, and the resulting CV density is one-sided, as shown in blue on the right. This also introduces errors in the PMF, as shown in Figure S3. In contrast, as shown in orange, for OPES-eABF the biased CV density always converges to a uniform distribution due to the additional ABF, regardless of the starting configuration.

Overall, we have shown how combining WTM and OPES with eABF yields increased robustness against input parameters. Additionally, for this simple test case, WTM-eABF and OPES-eABF yield significantly improved convergence to the analytic result, which is always reached after less than 100 ps. We mainly attribute this to the more accurate reweighting of extended-system dynamics, where sample weights are decoupled from time-dependent potentials. Furthermore, OPES-eABF is capable of eliminating weaknesses of OPES for systems with high ΔA, where the barrier factor is always well chosen for only one of the two basins, and removes the dependence of OPES on the starting configuration for such cases. In the next sections we move to more realistic test systems, where CVs are imperfect, like it is frequently the case in practice.

The Case of a Bad CV: Müller−Brown Potential. The Müller−Brown (MB) potential is a prototypical example for a system with a nonlinear reaction coordinate, which is a function of both the x and y-coordinates. A good CV can for example be obtained by using path collective variables (PCVs).[37] Figure 7 shows results for sampling the MB potential along an optimal PCV using OPES and OPES-eABF. The OPES potentials are parameterized identically. Details on the PCV, as well as a picture of the MB potential together with path nodes and sampling points are given in the SI. In line with the above results for the ADW potential, the OPES-eABF converges faster and with smaller standard deviation than the OPES.

However, in practice optimal CVs are frequently not available, and one might even be challenged with cases of bad CVs that miss important degrees of freedom. For example, in the MB potential using only the x-axis as a CV is insufficient, and cannot correctly capture the TSE. In Figure 8 we analyze how WTM, WTM-eABF, OPES, and OPES-eABF can cope with such situations, results of which are shown from the left-hand column to the right-hand column, respectively. The upper row shows a contour plot of the MB potential together with combined sampling points from 11 independent simulations. As expected, due to the poor choice of CV there is a gap in the sampling of configurations at the transition state for all sampling algorithms. As already discussed by Dietschreit et al.,[28] the analytic activation free energy $\Delta A^{\ddagger}$, as is shown in the lowest row, must therefore be underestimated, which is indicated by the blue vs green dashed lines that show the expected value from $\rho(x)$ vs the analytic $\Delta A^{\ddagger}$ from the PCV.[37] However, the reaction free energy ΔA, which is shown on the fourth row, can still correctly be obtained,[28] as shown by the alignment of the blue and green dashed lines, which again represent the analytical and expected results, respectively. Indeed, all methods except WTM converge to the analytic references within 10 ns. WTM-eABF initially tends to slightly overestimate the PMF and shows slower convergence and higher standard deviations compared to OPES and OPES-eABF. The latter two methods show similar convergence behavior, as indicated by similar error bars for the PMF, ΔA and $\Delta A^{\ddagger}$. Both methods converge to the analytic results in the first nanoseconds, over the remaining time only reducing the standard deviation, which is shown by light areas. Thus, both methods make the best of the poor choice of CV, which remains a limiting factor to the simulations. The better performance of OPES and OPES-eABF compared to WTM and WTM-eABF, can be attributed to the quasi-static nature of the OPES potential, which avoids pushing the system to high-energy transitions. The occurrence of much more transitions with WTM and WTM-eABF than with OPES aligns with this

https://doi.org/10.1021/acs.jctc.5c00395
J. Chem. Theory Comput. 2025, 21, 6434−6445

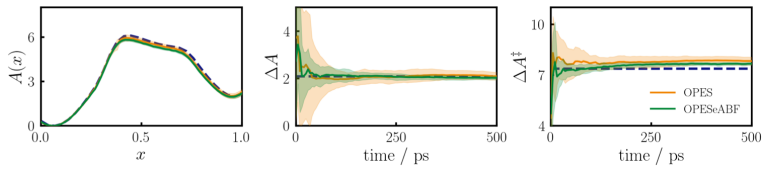


Figure 7. Sampling of the MB potential using the optimal path collective variable (PCV) as a CV and the OPES (orange) or OPES-eABF (green) sampling methods, respectively. Analytic results denoted by blue dashed lines and standard deviations from 11 independent runs by light areas.

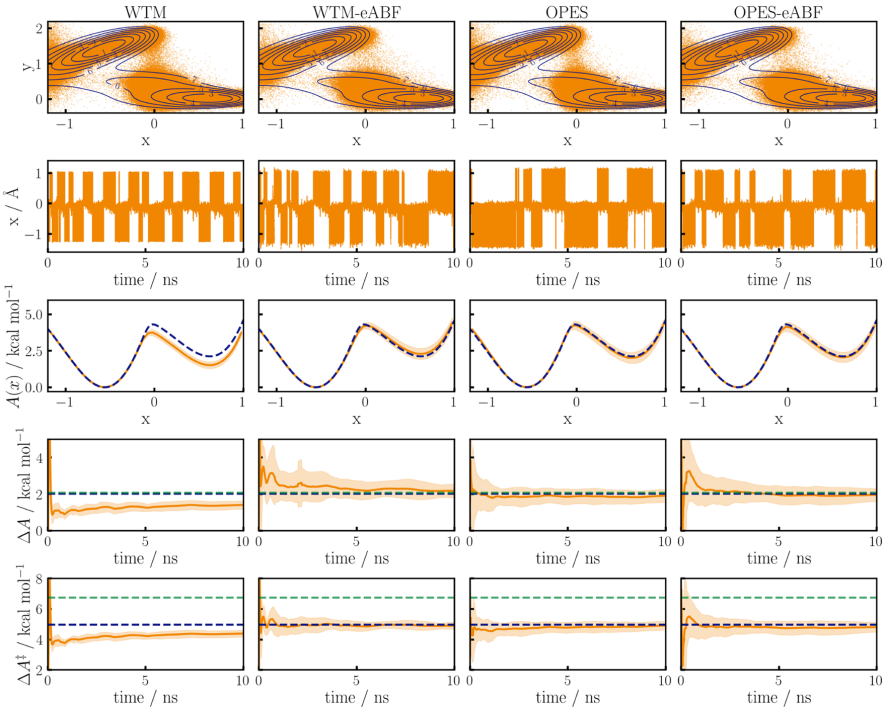


Figure 8. Sampling of a Müller−Brown potential using the x-axis as CV and four different sampling algorithms, WTM and WTM-eABF in the first two columns and OPES and OPES-eABF in the third and fourth column, respectively. In the top row, the MB potential is shown as a contour plot together with data points from 11 independent 10 ns runs. The second row contains prototypical trajectories, while the third row shows the mean final PMFs. The standard deviation from the 11 runs is indicated by light areas. In the fourth and fifth row the corresponding convergences of ΔA and $\Delta A^{\ddagger}$ are shown, respectively. Dashed blue lines indicate the reference obtained by numerically integrating the analytic probability density over the y-axis, while the dashed green lines show results for an optimal CV.[37]

observation, as can be seen in the second column of Figure 8 where prototypical trajectories are shown (remaining trajectories shown in the SI). Interestingly, OPES-eABF shows almost as many transitions as WTM-eABF without compromising accuracy. For real chemical systems, where the configurational space that needs to be sampled is much larger than for 2D potentials, this is highly important in order to enable the convergence of simulations within affordable time scales.[16]

Hence, we successfully reproduce the highly beneficial properties of OPES for sampling along bad CVs,[17] outperforming WTM(-eABF). We show that OPES-eABF inherits these strengths, but at the same time is able to increase the transition rate more effectively than OPES, combining fast exploration with accurate convergence. Hence, OPES-eABF is a promising alternative to OPES_{E}, as the latter focuses only on fast exploration while sacrificing accuracy.[17]

The Alanine Dipeptide System. Finally, we turn to alanine dipeptide in vacuum, which is one of the most popular test systems for importance sampling methods. It is well-known that the slow dynamics of alanine dipeptide are mainly governed by the Φ and Ψ dihedral angles, which discriminate between the $C7_{\text{eq}}$ and $C7_{\text{ax}}$ configurations. Figure 9 shows 2D OPES and OPES-eABF simulations, where both Φ and Ψ are employed as CVs. Snapshots of PMFs after 0.2, 0.5, and 2.0 ns are shown. Again, OPES-eABF converges significantly faster, as indicated by smoother PMFs at all stages. After 2 ns PMFs from both methods approach the reference PMF as shown in Figure 2, with OPES-eABF being more reliable in high-energy

https://doi.org/10.1021/acs.jctc.5c00395
J. Chem. Theory Comput. 2025, 21, 6434−6445

Journal of Chemical Theory and Computation pubs.acs.org/JCTC Article

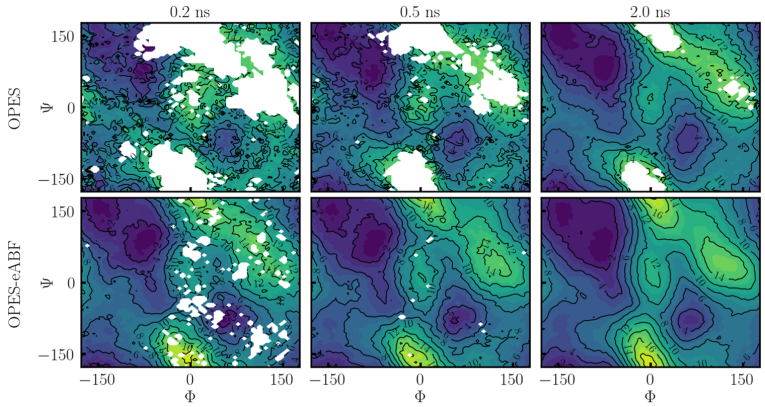


Figure 9. Sampling of the (Φ, Ψ) transitions of alanine dipeptide in vacuum (Ramachandran plot). The top row shows PMFs as obtained from OPES and the lower row from OPES-eABF. From left to right snapshots after 0.2, 0.5, and 2 ns of conformational sampling are given.

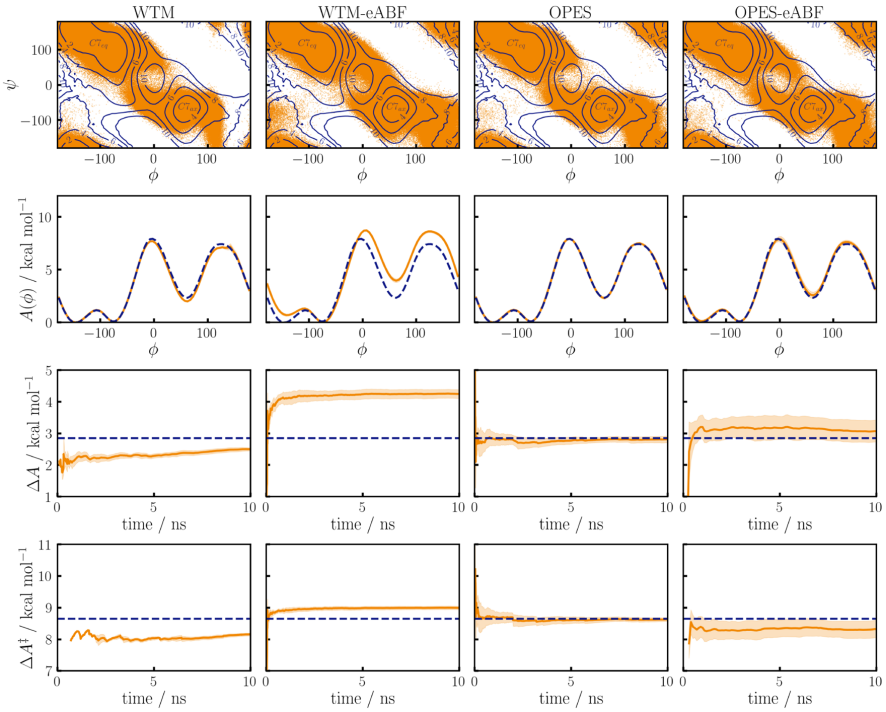


Figure 10. Sampling of alanine dipeptide in vacuum using the Φ torsion as CV and four different sampling algorithms, WTM and WTM-eABF in the first two columns and OPES and OPES-eABF in the third and fourth column, respectively. In the top row, data points from 11 independent 10 ns runs are shown in the (Φ, Ψ) plane. The third row shows the mean final PMFs, standard deviation from the 11 runs indicated by light areas. In the third and fourth row the corresponding convergences of ΔA and $\Delta A^{\ddagger}$ are shown, respectively.

regions. Along the same lines, as in OPES-eABF one samples from a uniform distribution instead of a well-tempered one, the PMF is already fully explored after 0.5 ns, including high-energy regions.

After again showing the beneficial speed of exploration and more accurate convergence of OPES-eABF as compared to OPES for the case of very good CVs, we will turn to only using the Φ coordinate as a CV, which still represents a relatively good choice, but lacks the Ψ degree of freedom. In Figure 10 we compare results from the WTM, WTM-eABF, OPES, and OPES-eABF methods, which are shown from the left-hand column to the right-hand column, respectively. To obtain a 1D

https://doi.org/10.1021/acs.jctc.5c00395
J. Chem. Theory Comput. 2025, 21, 6434−6445

reference PMF, the probability density as obtained from the 2D reference simulation is numerically integrated over the Ψ degree of freedom. On the top row the sampling of the (Φ, Ψ) plane is shown, with the reference PMF indicated by a contour plot where the $C7_{eq}$ and $C7_{ax}$ states are labeled. We focus on transitions over the TS at $\Phi \sim 0°$, which can occur through two different reaction channels, where the lower one $(\Phi < 0)$ is favored by about 1 kcal/mol. All methods except WTM-eABF predominantly sample the lower energy transition channel. This can be attributed to harsh pushing of the WTM-eABF in the Φ direction. Note that one might be able to avoid this by changing the parameterization of the WTM potential for WTM-eABF, but that OPES(-eABF) naturally avoids such effects due to the quasi-static nature of OPES, which by construction does not push on the CV. This is a fundamental property of OPES(-eABF) and not only an effect of the chosen parameters, which can be shown by using a much higher value of ΔE, still leading to dominant sampling of the lower energy transition as shown in Figure S9 of the SI. The second row of Figure 10 shows the final PMFs, while the third and fourth row show the convergence of ΔA and $\Delta A^{\ddagger}$, respectively. The final PMFs from WTM, OPES, and OPES-eABF are qualitatively similar, while WTM-eABF overestimates the PMF for the $C7_{ax}$ state, which also leads to overestimation of the ΔA and $\Delta A^{\ddagger}$. This is caused by the dominant sampling of the wrong transition channel, as discussed above. The standard deviation for $A(\Phi)$, ΔA, and $\Delta A^{\ddagger}$ is higher for OPES-eABF than for WTM and OPES, with the latter showing the fastest convergence, reproducing results from the original OPES implementation by Invernizzi and Parrinello.[14] The higher standard deviation of results from OPES-eABF can be understood by considering the broader sampling of configuration space, frequently observing both possible transitions while still overall converging to the correct result.

Overall, the results again show the better exploration capabilities of OPES-eABF compared to OPES, while high accuracy in reweighting is maintaining. For sampling the Ramachandran plot OPES-eABF emerges as significantly more efficient than OPES. However, if only Φ is chosen as CV the alanine dipeptide example also sheds light on a paradox: broader sampling of the configuration space frequently increases the statistical uncertainty of results, simply because configurations that are never visited in more local sampling cannot contribute to statistical errors. An ideal adaptive-sampling algorithm has to balance two competing goals: high accuracy in free-energy estimates can only be obtained if the system is not disturbed too much, but high efficiency in sampling, especially for the real-life scenario of suboptimal CVs, is only possible if the system is pushed to undergo transitions. We show on the example of WTM-eABF, that focusing on the latter can lead to artifacts in free-energy estimates. OPES is designed to focus on the former, resulting in almost exclusive sampling of the lower energy transition. With OPES-eABF we try to find a balance between the two extremes, as for more complicated reaction mechanisms the ability to efficiently explore multiple parallel reaction pathways can provide significant additional insight and yield more robust free-energy estimates, as the danger of missing important configurations is smaller.[41] Both reaction channels are frequently sampled, providing a full picture of the process at hand, but overall still converging to the correct result within chemical accuracy (1 kcal/mol) almost from the start.

CONCLUSIONS

In the spirit of the WTM-eABF hybrid method,[23,24] we introduce a new OPES-eABF hybrid and show that it unites multiple favorable properties of its building blocks:

- Combining WTM/OPES with the complementary ABF obscures weaknesses of the former, such that simulations become highly robust against the choice of input parameters. We observe that while OPES is highly efficient in many cases, it can be difficult to choose a good barrier factor for systems with high ΔA, as demonstrated in the ADW potential.
- Due to its quasi-static nature, OPES is very well suited for combination with eABF. Especially for safe choices of ΔE, smaller than the targeted barriers, OPES very quickly converges, leaving the remaining barrier to be cautiously removed by ABF. Therefore, the system is not pushed into high-energy transitions, that may arise in WTM or WTM-eABF if parameters are chosen too harshly.
- While OPES-eABF always converges faster and is more accurate for sampling along good CVs, both OPES and OPES-eABF show favorable convergence of PMFs and reaction free energies along incomplete or even poor CVs compared to WTM or WTM-eABF. Additionally, with OPES-eABF more transitions are observed than with OPES, representing a good balance between sampling efficiency and accuracy. Hence, OPES-eABF is a promising alternative to $OPES_E$, which is highly efficient for fast exploration, but cannot provide accurate equilibrium properties.[17]
- The extended-system decouples the physical system from time-dependent biasing potentials, may it be WTM or OPES. Therefore, the problem of time dependence of statistical weights never arises, and unbiased probabilities can accurately be recovered using MBAR.[25,26]

Hence, OPES-eABF provides a promising basis for the development of a black-box sampling tool that does not require manual parameterization. To this end, we introduce a method to automatically obtain a suitable coupling width σ_{ext} from a short unbiased MD, which is the only critical parameter for setting up the extended-system. We show for the three discussed systems that this method is robust for a wide range of applications, although there is room for improvement of the method especially for states with diffuse probability density, where the obtained coupling width may be too loose. Together with the adaptive bandwidth algorithm for OPES,[17] the only remaining parameter that is manually set is the barrier factor ΔE, which can safely be set to 20 kcal/mol for biochemical applications as the ABF will always remove remaining barriers. The ABF introduces a single additional parameter, which controls how fast the biasing force is scaled up and can always safely be set e.g., to 500 samples,[21] resulting in what resembles an out-of-the-box algorithm. Altogether, we expect that the simplicity of setting up OPES-eABF simulations will be appealing to practitioners from diverse fields. Both our implementations of OPES and OPES-eABF are publicly available in the adaptive-sampling python package.[32]

Throughout this work, we have shown the convergence of both the reaction and activation free energy, hoping to obtain both quantities with similar precision, such that complex transitions can be fully characterized from a single simulation. However, as observed in the MB potential and also discussed

https://doi.org/10.1021/acs.jctc.5c00395
J. Chem. Theory Comput. 2025, 21, 6434−6445

by Dietschreit et al.,[28] the latter requires a careful choice of CV that today is still far from trivial and requires significant experience. Despite the impressive progress made in the field of trainable CVs in recent years,[35,37,42−46] there is still much room for improvement, especially in the development of methods that are not based on manual feature selection. We envision that together with innovations in the field of chemical reaction space exploration,[47−51] it will be possible in the future to discover complicated reaction mechanisms in an automated way using adaptive importance sampling.

■ ASSOCIATED CONTENT

Data Availability Statement

The full source code of the OPES and OPES-eABF implementations and scripts to repeat all simulations are publicly available on GitHub (https://github.com/ochsenfeld-lab/adaptive_sampling).

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.5c00395.

Parameters of numerical potentials and additional simulation data for all three discussed test systems (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Christian Ochsenfeld − *Chair of Theoretical Chemistry, Department of Chemistry, University of Munich (LMU), D-81377 München, Germany; Max Planck Institute for Solid State Research, D-70569 Stuttgart, Germany;* orcid.org/0000-0002-4189-6558; Email: christian.ochsenfeld@uni-muenchen.de

Authors

Andreas Hulm − *Chair of Theoretical Chemistry, Department of Chemistry, University of Munich (LMU), D-81377 München, Germany;* orcid.org/0000-0003-1268-7578

Robert P. Schiller − *Chair of Theoretical Chemistry, Department of Chemistry, University of Munich (LMU), D-81377 München, Germany*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.5c00395

Funding

Open access funded by Max Planck Society.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Alexandra Stan-Bernhardt for fruitful discussions. Financial support was provided by the "Deutsche Forschungsgemeinschaft" (DFG, German Research Foundation) within SFB 1309-325871075 "Chemical Biology of Epigenetic Modifications". C.O. acknowledges further support as Max-Planck-Fellow at the MPI-FKF Stuttgart.

■ REFERENCES

(1) Lee, E. H.; Hsin, J.; Sotomayor, M.; Comellas, G.; Schulten, K. Discovery through the computational microscope. *Structure* **2009**, *17*, 1295−1306.

(2) Schlick, T.; Collepardo-Guevara, R.; Halvorsen, L. A.; Jung, S.; Xiao, X. Biomolecular modeling and simulation: a field coming of age. *Q. Rev. Biophys.* **2011**, *44*, 191−228.

(3) Huggins, D. J.; Biggin, P. C.; Dämgen, M. A.; Essex, J. W.; Harris, S. A.; Henchman, R. H.; Khalid, S.; Kuzmanic, A.; Laughton, C. A.; Michel, J.; et al. Biomolecular simulations: From dynamics and mechanisms to computational assays of biological activity. *WIREs Comput. Mol. Sci.* **2019**, *9*, No. e1393.

(4) Martoňák, R.; Laio, A.; Parrinello, M. Predicting crystal structures: the Parrinello-Rahman method revisited. *Phys. Rev. Lett.* **2003**, *90*, No. 075503.

(5) Martoňák, R.; Donadio, D.; Oganov, A. R.; Parrinello, M. Crystal structure transformations in SiO2 from classical and ab initio metadynamics. *Nat. Mater.* **2006**, *5*, 623−626.

(6) Zhu, Q.; Oganov, A. R.; Lyakhov, A. O. Evolutionary metadynamics: a novel method to predict crystal structures. *CrystEngComm* **2012**, *14*, 3596−3601.

(7) Chipot, C. Frontiers in Free-Energy Calculations of Biological Systems. *WIREs Comput. Mol. Sci.* **2014**, *4*, 71−89.

(8) Valsson, O.; Tiwary, P.; Parrinello, M. Enhancing important fluctuations: Rare events and metadynamics from a conceptual viewpoint. *Annu. Rev. Phys. Chem.* **2016**, *67*, 159−184.

(9) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **1977**, *23*, 187−199.

(10) Kästner, J. Umbrella sampling. *WIREs Comput. Mol. Sci.* **2011**, *1*, 932−942.

(11) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12562−12566.

(12) Barducci, A.; Bussi, G.; Parrinello, M. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Phys. Rev. Lett.* **2008**, *100*, No. 020603.

(13) Dama, J. F.; Rotskoff, G.; Parrinello, M.; Voth, G. A. Transition-tempered metadynamics: Robust, convergent metadynamics via on-the-fly transition barrier estimation. *J. Chem. Theory Comput.* **2014**, *10*, 3626−3633.

(14) Invernizzi, M.; Parrinello, M. Rethinking metadynamics: from bias potentials to probability distributions. *J. Phys. Chem. Lett.* **2020**, *11*, 2731−2736.

(15) Tiwary, P.; Parrinello, M. A time-independent free energy estimator for metadynamics. *J. Phys. Chem. B* **2015**, *119*, 736−742.

(16) Ray, D.; Rizzi, V. Enhanced Sampling with Suboptimal Collective Variables: Reconciling Accuracy and Convergence Speed. *J. Chem. Theory Comput.* **2025**, *21*, 58−69.

(17) Invernizzi, M.; Parrinello, M. Exploration vs convergence speed in adaptive-bias enhanced sampling. *J. Chem. Theory Comput.* **2022**, *18*, 3988−3996.

(18) Darve, E.; Pohorille, A. Calculating free energies using average force. *J. Chem. Phys.* **2001**, *115*, 9169−9183.

(19) Darve, E.; Rodriguez-Gomez, D.; Pohorille, A. Adaptive biasing force method for scalar and vector free energy calculations. *J. Chem. Phys.* **2008**, *128*, No. 144120.

(20) Lelièvre, T.; Rousset, M.; Stoltz, G. *Free Energy Computations: A Mathematical Perspective*; Imperial College Press, 2010.

(21) Comer, J.; Gumbart, J. C.; Hénin, J.; Lelièvre, T.; Pohorille, A.; Chipot, C. The adaptive biasing force method: Everything you always wanted to know but were afraid to ask. *J. Phys. Chem. B* **2015**, *119*, 1129−1151.

(22) Lesage, A.; Lelievre, T.; Stoltz, G.; Henin, J. Smoothed biasing forces yield unbiased free energies with the extended-system adaptive biasing force method. *J. Phys. Chem. B* **2017**, *121*, 3676−3685.

(23) Fu, H.; Zhang, H.; Chen, H.; Shao, X.; Chipot, C.; Cai, W. Zooming across the free-energy landscape: shaving barriers, and flooding valleys. *J. Phys. Chem. Lett.* **2018**, *9*, 4738−4745.

(24) Fu, H.; Shao, X.; Cai, W.; Chipot, C. Taming rugged free energy landscapes using an average force. *Acc. Chem. Res.* **2019**, *52*, 3254−3264.

(25) Shirts, M. R.; Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* **2008**, *129*, No. 124105.
(26) Hulm, A.; Dietschreit, J. C.; Ochsenfeld, C. Statistically optimal analysis of the extended-system adaptive biasing force (eABF) method. *J. Chem. Phys.* **2022**, *157*, No. 024110.
(27) Dietschreit, J. C. B.; Diestler, D. J.; Ochsenfeld, C. How to obtain reaction free energies from free-energy profiles. *J. Chem. Phys.* **2022**, *156*, No. 114105.
(28) Dietschreit, J. C. B.; Diestler, D. J.; Hulm, A.; Ochsenfeld, C.; Gómez-Bombarelli, R. From free-energy profiles to activation free energies. *J. Chem. Phys.* **2022**, *157*, No. 084113.
(29) Bussi, G.; Laio, A.; Parrinello, M. Equilibrium free energies from nonequilibrium metadynamics. *Phys. Rev. Lett.* **2006**, *96*, No. 090601.
(30) Dama, J. F.; Parrinello, M.; Voth, G. A. Well-tempered metadynamics converges asymptotically. *Phys. Rev. Lett.* **2014**, *112*, No. 240602.
(31) Fu, H.; Shao, X.; Chipot, C.; Cai, W. Extended adaptive biasing force algorithm. An on-the-fly implementation for accurate free-energy calculations. *J. Chem. Theory Comput.* **2016**, *12*, 3506−3513.
(32) Hulm, A.; Lemke, Y.; Johannes, D.; Glinkina, L.; Stan-Bernhardt, A. Adaptive Sampling, https://github.com/ochsenfeld-lab/adaptive_sampling.
(33) Welford, B. P. Note on a method for calculating corrected sums of squares and products. *Technometrics* **1962**, *4*, 419−420.
(34) Paterlini, M. G.; Ferguson, D. M. Constant temperature simulations using the Langevin equation with velocity Verlet integration. *Chem. Phys.* **1998**, *236*, 243−252.
(35) Leines, G. D.; Ensing, B. Path finding on high-dimensional free energy landscapes. *Phys. Rev. Lett.* **2012**, *109*, No. 020601.
(36) de Alba Ortíz, A. P.; Tiwari, A.; Puthenkalathil, R.; Ensing, B. Advances in enhanced sampling along adaptive paths of collective variables. *J. Chem. Phys.* **2018**, *149*, No. 072320.
(37) Hulm, A.; Ochsenfeld, C. Improved Sampling of Adaptive Path Collective Variables by Stabilized Extended-System Dynamics. *J. Chem. Theory Comput.* **2023**, *19*, 9202−9210.
(38) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; et al. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* **2017**, *13*, No. e1005659.
(39) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696−3713.
(40) Humphrey, W.; Dalke, A.; Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graphics* **1996**, *14*, 33−38.
(41) Pöverlein, M. C.; Hulm, A.; Dietschreit, J. C.; Kussmann, J.; Ochsenfeld, C.; Kaila, V. R. QM/MM Free Energy Calculations of Long-Range Biological Protonation Dynamics by Adaptive and Focused Sampling. *J. Chem. Theory Comput.* **2024**, *20*, 5751−5762, DOI: 10.1021/acs.jctc.4c00199.
(42) Branduardi, D.; Gervasio, F. L.; Parrinello, M. From A to B in free energy space. *J. Chem. Phys.* **2007**, *126*, No. 054103.
(43) Bonati, L.; Trizio, E.; Rizzi, A.; Parrinello, M. A unified framework for machine learning collective variables for enhanced sampling simulations: mlcolvar. *J. Chem. Phys.* **2023**, *159*, No. 014801, DOI: 10.1063/5.0156343.
(44) Yang, S.; Nam, J.; Dietschreit, J. C. B.; Gómez-Bombarelli, R. Learning Collective Variables with Synthetic Data Augmentation through Physics-Inspired Geodesic Interpolation. *J. Chem. Theory Comput.* **2024**, *20*, 6559−6568.
(45) Zhang, J.; Bonati, L.; Trizio, E.; Zhang, O.; Kang, Y.; Hou, T.; Parrinello, M. Descriptors-free collective variables from geometric graph neural networks. *J. Chem. Theory Comput.* **2024**, *20*, 10787−10797.
(46) Kang, P.; Trizio, E.; Parrinello, M. Computing the committor with the committor to study the transition state ensemble. *Nat. Comput. Sci.* **2024**, *4*, 451−460.
(47) Wang, L.-P.; Titov, A.; McGibbon, R.; Liu, F.; Pande, V. S.; Martínez, T. J. Discovering chemistry with an ab initio nanoreactor. *Nat. Chem.* **2014**, *6*, 1044−1048.
(48) Raucci, U.; Rizzi, V.; Parrinello, M. Discover, sample, and refine: Exploring chemistry with enhanced sampling techniques. *J. Phys. Chem. Lett.* **2022**, *13*, 1424−1430.
(49) Stan, A.; von der Esch, B.; Ochsenfeld, C. Fully automated generation of prebiotically relevant reaction networks from optimized nanoreactor simulations. *J. Chem. Theory Comput.* **2022**, *18*, 6700−6712.
(50) Stan-Bernhardt, A.; Glinkina, L.; Hulm, A.; Ochsenfeld, C. Exploring Chemical Space Using Ab Initio Hyperreactor Dynamics. *ACS Cent. Sci.* **2024**, *10*, 302−314.
(51) Weymuth, T.; Unsleber, J. P.; Türtscher, P. L.; Steiner, M.; Sobez, J.-G.; Müller, C. H.; Mörchen, M.; Klasovita, V.; Grimmel, S. A.; Eckhoff, M.; et al. SCINE-Software for chemical interaction networks. *J. Chem. Phys.* **2024**, *160*, No. 222501, DOI: 10.1063/5.0206974.

https://doi.org/10.1021/acs.jctc.5c00395
J. Chem. Theory Comput. 2025, 21, 6434−6445

Supporting Information

Combining Fast Exploration With Accurate Reweighting In the OPES-eABF Hybrid Sampling Method

Andreas Hulm,[1] Robert Schiller,[1] Christian Ochsenfeld[1,2,*]

[1]Chair of Theoretical Chemistry, Department of Chemistry,
University of Munich (LMU), Butenandtstr. 7, D-81377 München, Germany
[2]Max Planck Institute for Solid State Research, Heisenbergstr. 1, D-70569 Stuttgart, Germany
*E-Mail: christian.ochsenfeld@uni-muenchen.de

Contents

1 Asymmetric double-well potential

The asymetric double well potential is defined by

$$U^{\mathrm{ADW}}(x, y) = ax^2 - bx^3 + cx^4 + dy^2 + e\,, \tag{S1}$$

empirical parameters given in Table S1.

a	62.75
b	64.84
c	15.81
d	12.55
e	16.29

Table S1: Empirical parameters for the asymmetric double well potential (kcal/mol).

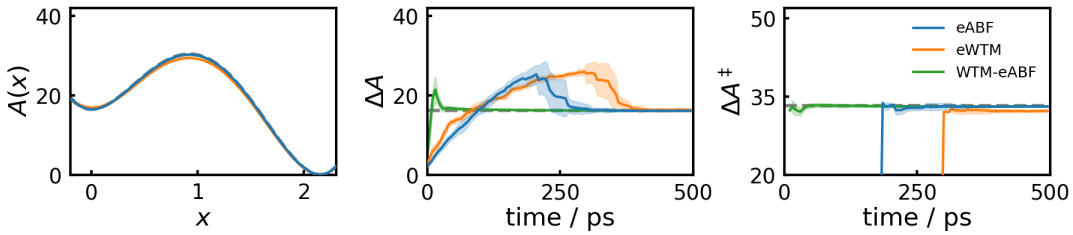


Figure S1: On the left the mean PMFs from 11 independent 500 ps extended-system runs using eABF (blue), eWTM (orange), and WTM-eABF (green) bias are shown, with standard deviations denoted by light areas. The convergence of the reaction free energy is shown in the middle and the activation free energy on the right. Dashed gray lines denote analytic results.

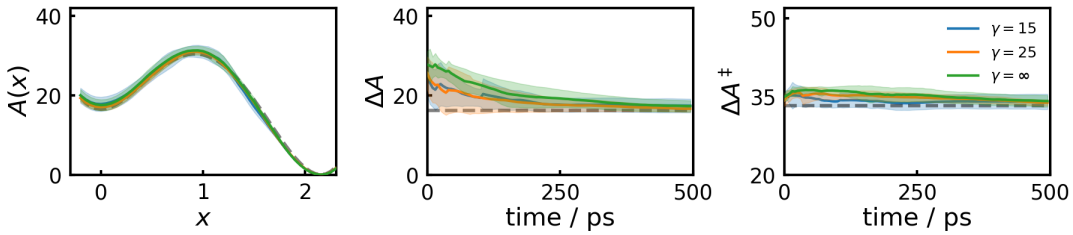


Figure S2: On the left the mean PMF from 11 independent 500 ps OPES runs with bias factors $\gamma = 15$ (blue), $\gamma = 25$ (orange) and $\gamma = \infty$ (green) is shown, with standard deviations denoted by light areas. The corresponding convergence of the reaction free energy is shown in the middle and the activation free energy on the right.

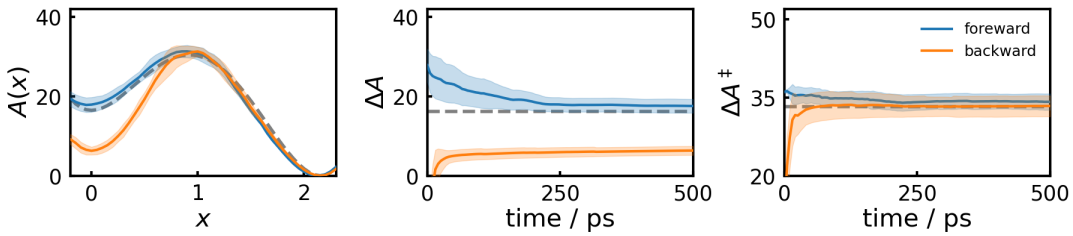


Figure S 3: On the left the mean PMFs from 11 independent 500 ps OPES runs with barrier parameter $\Delta E = 30.4$ are shown starting in the lower energy (foreward reaction, blue) or higher energy (backward reaction, orange) minimum, with standard deviations denoted by light areas. The convergence of the reaction free energy is shown in the middle and the activation free energy on the right. Dashed gray lines denote analytic results.

2 Müller-Brown potential

The Müller-Brown potential is given by

$$U^{\mathrm{MB}}(x,y) = B\sum_{i=1}^{4} A_i \exp\left[\alpha_i(x-x_i)^2 + \beta_i(x-x_i)(y-y_i) + \gamma_i(y-y_i)^2\right] \ , \tag{S2}$$

with B=1 kJ/mol and other empirical parameters given in Table S2.

i	A_i	α_i	β_i	γ_i	x_i	y_i
1	-40.0	-1.0	0.0	-10.0	1.0	0.0
2	-10.0	-1.0	0.0	-10.0	0.0	0.5
3	-34.0	-6.5	11.0	-6.5	-0.5	1.5
4	3.0	0.7	0.6	0.7	-1.0	1.0

Table S2: Empirical parameters for the Müller-Brown potential.

To obtain an optimal CV, a path is optimized using the nudged elastic band (NEB) method [1], and used for PCV simulations [2, 3]. The distance to the path is confined with a harmonic constraint with force constant 100 kcal/mol Å^2. The extended-system is stabilized against discontinuous jumps in the PCV due to path short cutting [4]. Below, path nodes are shown on the MB potential together with sampling points from 500 ps OPES and OPES-eABF simulations, with barrier factor 5 kcal/mol, as well as automatic estimation of σ_{ext} and σ_{G} from 5000 unbiased MD steps (see also Fig. 6 of the main text).

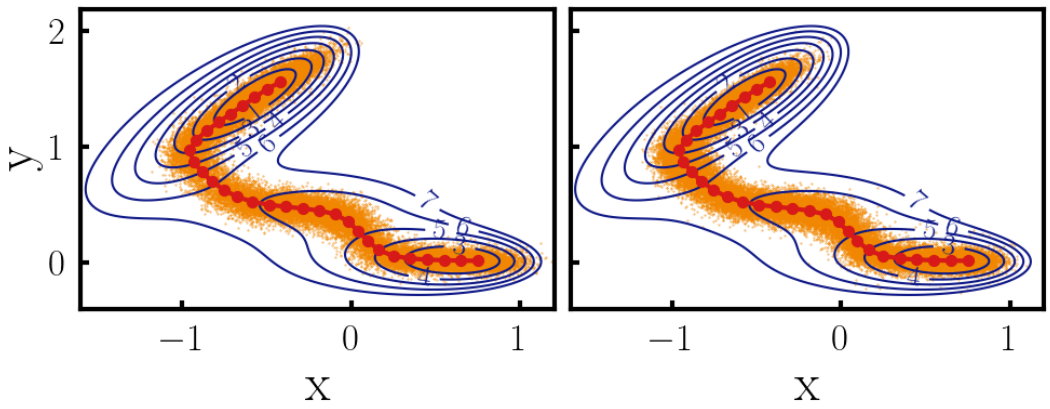


Figure S4: Path nodes (red) on the MB potential (blue), sampling points from path OPES (left) and path OPES-eABF (right) simulations shown in orange.

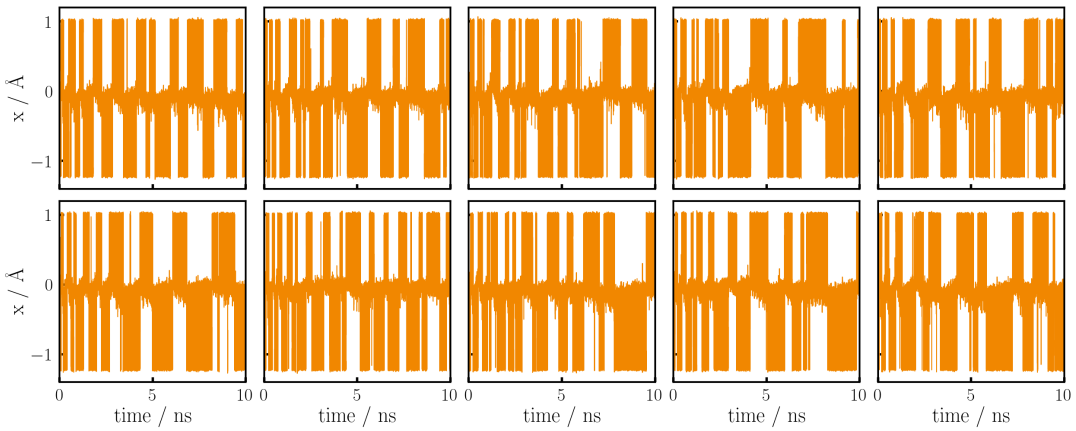


Figure S5: The remaining 10 trajectories of WTM simulations in the MB potential, that are not shown in Fig 6 of the main text.

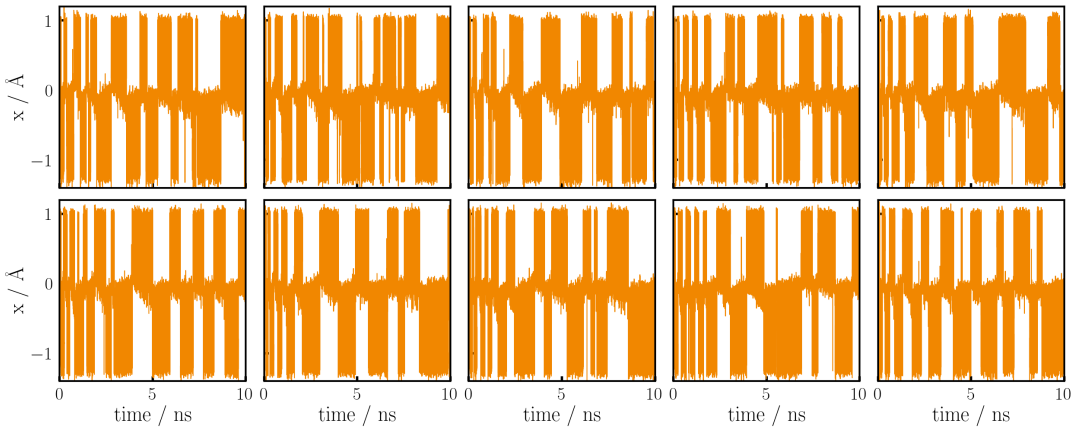


Figure S6: The remaining 10 trajectories of WTM-eABF simulations in the MB potential, that are not shown in Fig 6 of the main text.

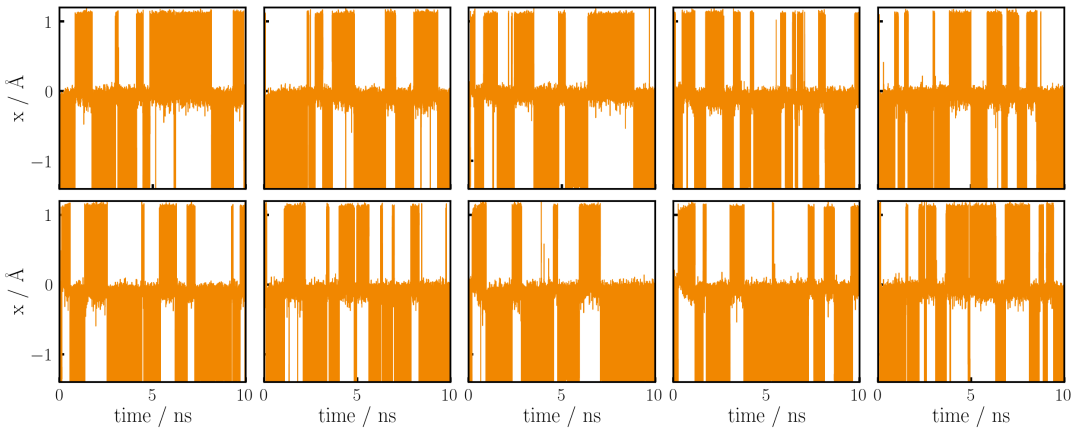


Figure S7: The remaining 10 trajectories of OPES simulations in the MB potential, that are not shown in Fig 6 of the main text.

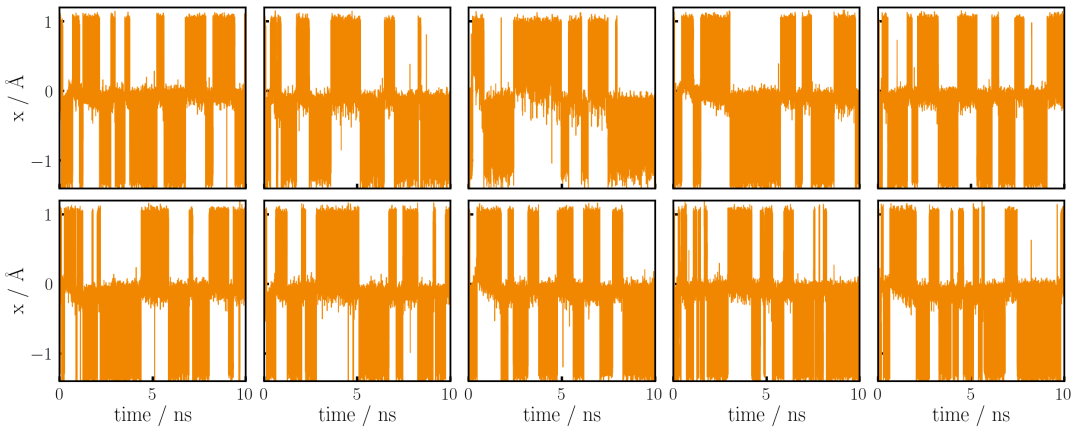


Figure S8: The remaining 10 trajectories of OPES-eABF simulations in the MB potential, that are not shown in Fig 6 of the main text.

3 Alanine dipeptide

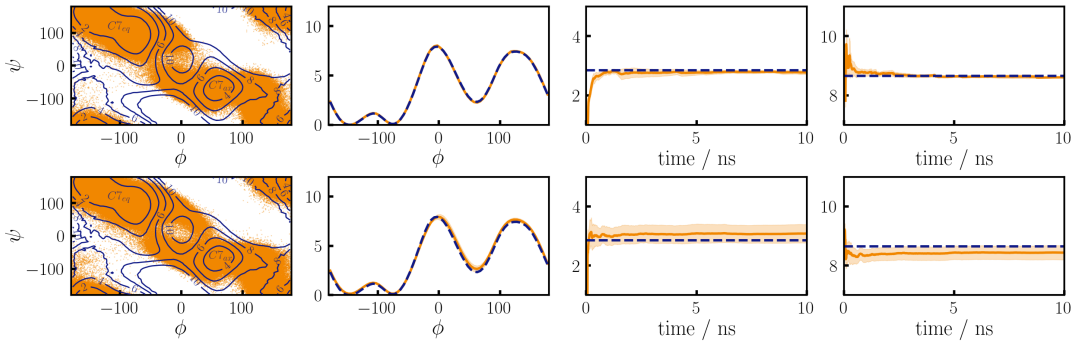


Figure S9: OPES (upper row) and OPES-eABF (lower row) simulations of alanine dipeptide along the ϕ angle. All parameters are identical to the main text, except ΔE, which is now set to 100 kJ/mol.

References

[1] Hannes Jónsson, Greg Mills, and Karsten W Jacobsen. "Nudged elastic band method for finding minimum energy paths of transitions". In: *Classical and quantum dynamics in condensed phase simulations*. World Scientific, 1998, pp. 385–404.

[2] Grisell Díaz Leines and Bernd Ensing. "Path finding on high-dimensional free energy landscapes". In: *Phys. Ref. Lett.* 109.2 (2012), p. 020601.

[3] A Pérez de Alba Ortíz et al. "Advances in enhanced sampling along adaptive paths of collective variables". In: *J. Chem. Phys.* 149.7 (2018), p. 072320.

[4] Andreas Hulm and Christian Ochsenfeld. "Improved Sampling of Adaptive Path Collective Variables by Stabilized Extended-System Dynamics". In: *J. Chem. Theory Comput.* 19.24 (2023), pp. 9202–9210.

Chapter 4

Conclusion and Outlook

In this thesis, several advancements in MD simulations of (bio)catalytic processes have been developed and applied. The biological systems studied, namely the PUS enzyme, the p97 enzyme, and the mammalian respiratory complex I, provide examples of the challenges one faces in this field of research. In short, these challenges originate from the spatial and temporal dimensions of such macromolecules, and overcoming them requires innovative and creative approaches. Focus lies on the temporal problem, which is addressed by importance sampling techniques, that can already restore ergodicity in time trajectories of only moderate length. Combined with empirical force fields that allow for the brute-force all-atom MD simulation of timescales up to milliseconds, as demonstrated by the D. E. Shaw consortium on specifically designed computer architectures [110], such methods have already had a long and fruitful history. In contrast, explicit MD simulations for macromolecular processes that require more elaborate QM/MM descriptions have only recently gained popularity, as trajectories of sufficient length have become routinely affordable.

Until today, in such large-scale QM/MM simulations, each data point is valuable because of its significant cost, and it is important to devise sampling schemes that are as data-efficient as possible. This goal is a recurring theme in this thesis, where initially in Publication **I** it is shown that the highly efficient WTM-eABF sampling algorithm (and also extended system dynamics in general) can be combined with the MBAR estimator to recover the full, bias-free, statistical information. This combination results in a versatile sampling tool, which is employed throughout the further works. In Publication **II**, WTM-eABF/MBAR is combined with adaptive PCVs to yield the path WTM-eABF scheme, which mitigates the difficulty of manually choosing appropriate CVs and enables the simulation of non-linear transitions like the reaction mechanism of PUS. A further example of the successful application of path WTM-eABF is provided in Publication **III**, where the reaction mechanism of ATP hydrolysis by the p97 enzyme is investigated. In Publication **IV**, the natural contradiction between ergodic sampling and fast convergence is addressed by combining global and local sampling strategies to mitigate the highly challenging long-range pT processes in biological proton pumps like the respiratory Complex I. Finally, in Publication **V**, an even more efficient sampling method is derived by replacing

WTM with the new OPES. All the innovations discussed above extend to the resulting OPES-eABF method, as it is also based on the extended system dynamics. Additionally, the new method requires minimal user intervention and is easy to parameterize even for non-expert practitioners.

For future projects, the presented works build a highly robust basis to devise even more autonomous sampling schemes that enable the qualitative and quantitative exploration of reaction mechanisms of catalytic systems. Especially promising is the combination of efficient reaction network discovery, as shown by hyperreactor dynamics in Publication **VII**, with the semi-automated calculation of associated free energies as provided by OPES-eABF in Publication **V**. Still, to address reaction mechanisms as complicated as many biochemical processes using such a workflow, further innovations have to be made for both the reaction space exploration and free energy refinement. The hyperreactor dynamics exploration has to be extended to QM/MM simulations by adapting the boost potentials to only affect the QM region while avoiding artifacts in the MM environment. Also, care has to be taken to provide mild conditions that prevent exploration of unrealistic reaction pathways. For OPES-eABF-based reaction refinement, the remaining difficulty consists in obtaining CVs that apply to diverse reactions without manual feature selection. PCVs, as applied in Publication **II**, are convenient in this regard, as in principle they only require a guess reaction pathway, which can be obtained by minimum energy path optimization methods as shown in Publication **III**. Note that such path optimizations are already performed in the refinement step of hyperreactor dynamics, such that OPES-eABF simulations using PCVs can be applied without additional effort. Alternatively, one might resort to the magnitude of recently developed machine-learning (ML) CVs, which provide new opportunities for all CV-based sampling methods [111–114]. While those matured over recent years, they still face two drawbacks: firstly, training data needs to be obtained before the simulation, resulting in the famous chicken-and-egg problem of ML CVs. While this can be addressed, *e.g.*, by recursive application of ML CVs starting from a guess CV, it would be highly interesting to explore if the hyperreactor dynamics trajectories are suitable for a priori training. Secondly, most ML CVs need the manual selection of a lower-dimensional feature space. Here, careful analysis of hyperreactor dynamics trajectories might provide unified ways to select CV spaces based on principal component analysis (PCA) [115] or other dimensionality reduction techniques [116].

To conclude, the developments presented in this thesis offer new opportunities for simulating highly intricate reaction mechanisms in explicit catalytic environments using importance sampling at a QM/MM level.

Bibliography

[1] D. B. Kitchen, H. Decornez, J. R. Furr, J. Bajorath, *Nat. Rev. Drug Discov.* **2004**, *3*, 935–949.

[2] J. D. Durrant, J. A. McCammon, *BMC Biol.* **2011**, *9*, 1–9.

[3] G. Sliwoski, S. Kothiwale, J. Meiler, E. W. Lowe, *Pharmacol. Rev.* **2014**, *66*, 334–395.

[4] T. Lavé, N. Parrott, H. Grimm, A. Fleury, M. Reddy, *Xenobiotica* **2007**, *37*, 1295–1310.

[5] R. J. Kazlauskas, *Curr. Opin. Chem. Biol.* **2000**, *4*, 81–88.

[6] G. Sin, J. M. Woodley, K. V. Gernaey, *Biotechnol. Prog.* **2009**, *25*, 1529–1538.

[7] Y. Ding, G. Perez-Ortiz, J. Peate, S. M. Barry, *Front. Mol. Biosci.* **2022**, *9*, 908285.

[8] A. Aksimentiev, R. Brunner, J. Cohen, J. Comer, E. Cruz-Chu, D. Hardy, A. Rajan, A. Shih, G. Sigalov, Y. Yin, K. Schulten, *Nanostructure Design: Methods and Protocols* **2008**, 181–234.

[9] M. Qiu, E. Khisamutdinov, Z. Zhao, C. Pan, J.-W. Choi, N. B. Leontis, P. Guo, *Philos. Transact. A Math. Phys. Eng. Sci.* **2013**, *371*, 20120310.

[10] A. Makarucha, N. Todorova, I. Yarovsky, *Eur. Biophys. J.* **2011**, *40*, 103–115.

[11] R. O. Dror, R. M. Dirks, J. Grossman, H. Xu, D. E. Shaw, *Annu. Rev. Biophys.* **2012**, *41*, 429–452.

[12] R. B. Fenwick, S. Esteban-Martín, X. Salvatella, *Eur. Biophys. J.* **2011**, *40*, 1339–1355.

[13] E. Fermi, P. Pasta, S. Ulam, M. Tsingou, Studies of the nonlinear problems, tech. rep., Los Alamos National Laboratory (LANL), Los Alamos, NM (United States), **1955**.

[14] B. J. Alder, T. E. Wainwright, *J. Chem. Phys.* **1959**, *31*, 459–466.

[15] A. Rahman, *Phys. Rev.* **1964**, *136*, A405.

[16] A. Hospital, J. R. Goñi, M. Orozco, J. L. Gelpí, *Adv. Appl. Bioinforma. Chem.* **2015**, 37–47.

[17] S. A. Hollingsworth, R. O. Dror, *Neuron* **2018**, *99*, 1129–1143.

[18] M. Born, *Ann. Phys.* **1927**, *84*, 457–484.

[19] S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta, P. Weiner, *J. Am. Chem. Soc.* **1984**, *106*, 765–784.

[20] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, P. A. Kollman, *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

[21] A. Warshel, M. Levitt, *J. Mol. Biol.* **1976**, *103*, 227–249.

[22] M. Karplus, M. Levitt, A. Warshel, *Nobel Media AB 2014* **2013**.

[23] J. Kussmann, C. Ochsenfeld, *J. Chem. Theory. Comput.* **2015**, *11*, 918–922.

[24] J. Kussmann, C. Ochsenfeld, *J. Chem. Theory. Comput.* **2017**, *13*, 3153–3159.

[25] H. Laqua, J. Kussmann, C. Ochsenfeld, *J. Chem. Theory. Comput.* **2018**, *14*, 3451–3458.

[26] H. Laqua, T. H. Thompson, J. Kussmann, C. Ochsenfeld, *J. Chem. Theory. Comput.* **2020**, *16*, 1456–1468.

[27] J. Kussmann, H. Laqua, C. Ochsenfeld, *J. Chem. Theory. Comput.* **2021**, *17*, 1512–1521.

[28] H. Laqua, J. C. B. Dietschreit, J. Kussmann, C. Ochsenfeld, *J. Chem. Theory. Comput.* **2022**, *18*, 6010–6020.

[29] A. Sodt, J. E. Subotnik, M. Head-Gordon, *J. Chem. Phys.* **2006**, *125*.

[30] F. Neese, F. Wennmohs, A. Hansen, U. Becker, *Chem. Phys.* **2009**, *356*, 98–109.

[31] P. Plessow, F. Weigend, *J. Comput. Chem.* **2012**, *33*, 810–816.

[32] P. Merlot, T. Kjærgaard, T. Helgaker, R. Lindh, F. Aquilante, S. Reine, T. B. Pedersen, *J. Comput. Chem.* **2013**, *34*, 1486–1496.

[33] F. Liu, J. Kong, *Chem. Phys. Lett.* **2018**, *703*, 106–111.

[34] G. M. Torrie, J. P. Valleau, *J. Comput. Phys.* **1977**, *23*, 187–199.

[35] J. Kästner, *WIREs Comput. Mol. Sci.* **2011**, *1*, 932–942.

[36] A. Laio, M. Parrinello, *Proc. Natl. Acad. Sci.* **2002**, *99*, 12562–12566.

[37] A. Barducci, G. Bussi, M. Parrinello, *Phys. Rev. Lett.* **2008**, *100*, 020603.

[38] E. Darve, A. Pohorille, *J. Chem. Phys.* **2001**, *115*, 9169–9183.

[39] E. Darve, D. Rodriguez-Gomez, A. Pohorille, *J. Chem. Phys.* **2008**, *128*.

[40] T. Lelièvre, M. Rousset, G. Stoltz, *Free energy computations: A mathematical perspective*, Imperial College Press, **2010**.

[41] L. Zheng, W. Yang, *J. Chem. Theory. Comput.* **2012**, *8*, 810–823.

[42] J. Comer, J. C. Gumbart, J. Henin, T. Lelievre, A. Pohorille, C. Chipot, *J. Phys. Chem. B.* **2015**, *119*, 1129–1151.

[43] A. Lesage, T. Lelièvre, G. Stoltz, J. Hénin, *J. Phys. Chem. B.* **2017**, *121*, 3676–3685.

[44] H. Fu, H. Zhang, H. Chen, X. Shao, C. Chipot, W. Cai, *J. Phys. Chem. Lett.* **2018**, *9*, 4738–4745.

[45] H. Fu, X. Shao, W. Cai, C. Chipot, *Acc. Chem. Res.* **2019**, *52*, 3254–3264.

[46] Y. Miao, V. A. Feher, J. A. McCammon, *J. Chem. Theory. Comput.* **2015**, *11*, 3584–3595.

[47] H. Chen, H. Fu, C. Chipot, X. Shao, W. Cai, *J. Chem. Theory. Comput.* **2021**, *17*, 3886–3894.

[48] M. R. Shirts, J. D. Chodera, *J. Chem. Phys.* **2008**, *129*, 124105.

[49] A. Hulm, A. Stan-Bernhardt, R. P. Schiller, L. Glinkina, J. C. B. Dietschreit, C. Ochsenfeld, `https://github.com/ochsenfeld-lab/adaptive_sampling`, **2022**.

[50] E. Weinan, W. Ren, E. Vanden-Eijnden, *Phys. Rev. B.* **2002**, *66*, 052301.

[51] D. Branduardi, F. L. Gervasio, M. Parrinello, *J. Chem. Phys.* **2007**, *126*, 054103.

[52] G. Diaz Leines, B. Ensing, *Phys. Rev. Lett.* **2012**, *109*, 020601.

[53] H. Jónsson, G. Mills, K. W. Jacobsen in *Classical and quantum dynamics in condensed phase simulations*, World Scientific, **1998**, pp. 385–404.

[54] P. König, N. Ghosh, M. Hoffmann, M. Elstner, E. Tajkhorshid, T. Frauenheim, Q. Cui, *J. Phys. Chem. A.* **2006**, *110*, 548–563.

[55] M. Invernizzi, M. Parrinello, *J. Phys. Chem. Lett.* **2020**, *11*, 2731–2736.

[56] B. Leimkuhler, C. Matthews, *Molecular dynamics*, *Vol. 39*, Springer, **2015**.

[57] R. Santamaria, *Molecular Dynamics*, Springer, **2023**.

[58] D. Frenkel, B. Smit, *Understanding molecular simulation: from algorithms to applications*, Elsevier, **2023**.

[59] M. G. Paterlini, D. M. Ferguson, *Chem. Phys.* **1998**, *236*, 243–252.

[60] A. Lyubartsev, A. Martsinovski, S. Shevkunov, P. Vorontsov-Velyaminov, *J. Chem. Phys.* **1992**, *96*, 1776–1783.

[61] C. J. Geyer, **1991**.

[62] J. Kästner, W. Thiel, *J. Chem. Phys.* **2005**, *123*, 144104.

[63] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, P. A. Kollman, *J. Comput. Chem.* **1992**, *13*, 1011–1021.

[64] C. H. Bennett, *J. Comput. Phys.* **1976**, *22*, 245–268.

[65] M. R. Shirts, *arXiv preprint arXiv:1704.00891* **2017**.

[66] R. W. Zwanzig, *J. Chem. Phys.* **1954**, *22*, 1420–1426.

[67] J. C. B. Dietschreit, D. J. Diestler, C. Ochsenfeld, *J. Chem. Phys.* **2022**, *156*.

[68] A. F. Voter, *J. Chem. Phys.* **1997**, *106*, 4665–4677.

[69] C. A. F. De Oliveira, D. Hamelberg, J. A. McCammon, *J. Chem. Phys.* **2007**, *127*, 175105.

[70] P. Tiwary, M. Parrinello, *Phys. Rev. Lett.* **2013**, *111*, 230602.

[71] Y. Wang, O. Valsson, P. Tiwary, M. Parrinello, K. Lindorff-Larsen, *J. Chem. Phys.* **2018**, *149*, 072309.

[72] J. McCarty, O. Valsson, P. Tiwary, M. Parrinello, *Phys. Rev. Lett.* **2015**, *115*, 070601.

[73] D. Ray, N. Ansari, V. Rizzi, M. Invernizzi, M. Parrinello, *J. Chem. Theory Comput.* **2022**, *18*, 6500–6509.

[74] D. Ray, M. Parrinello, *J. Chem. Theory Comput.* **2023**, *19*, 5649–5670.

[75] A. Kolmogoroff, *Mathematische Annalen* **1931**, *104*, 415–458.

[76] C. Dellago, P. G. Bolhuis, P. L. Geissler, *Computer Simulations in Condensed Matter Systems: From Materials to Chemical Biology Volume 1* **2006**, 349–391.

[77] H. Fu, H. Bian, X. Shao, W. Cai, *J. Phys. Chem. Lett.* **2024**, *15*, 1774–1783.

[78] A. Perez de Alba Ortiz, A. Tiwari, R. Puthenkalathil, B. Ensing, *J. Chem. Phys.* **2018**, *149*, 072320.

[79] J. Rogal, E. Schneider, M. E. Tuckerman, *Phys. Rev. Lett.* **2019**, *123*, 245701.

[80] A. France-Lanord, H. Vroylandt, M. Salanne, B. Rotenberg, A. M. Saitta, F. Pietrucci, *J. Chem. Theory. Comput.* **2024**, *20*, 3069–3084.

[81] D. Sheppard, R. Terrell, G. Henkelman, *J. Chem. Phys.* **2008**, *128*, 134106.

[82] G. Henkelman, H. Jónsson, *J. Chem. Phys.* **2000**, *113*, 9978–9985.

[83] N. Aho, G. Groenhof, P. Buslaev, *J. Chem. Theory. Comput.* **2024**, *20*, 6674–6686.

[84] D. Hamelberg, J. Mongan, J. A. McCammon, *J. Chem. Phys.* **2004**, *120*, 11919–11929.

[85] Y. Zhao, J. Zhang, H. Zhang, S. Gu, Y. Deng, Y. Tu, T. Hou, Y. Kang, *J. Phys. Chem. Lett.* **2023**, *14*, 1103–1112.

[86] L.-P. Wang, A. Titov, R. McGibbon, F. Liu, V. S. Pande, T. J. Martinez, *Nature chemistry* **2014**, *6*, 1044–1048.

[87] A. Stan, B. v. d. Esch, C. Ochsenfeld, *J. Chem. Theory. Comput.* **2022**, *18*, 6700–6712.

[88] A. Stan-Bernhardt, L. Glinkina, A. Hulm, C. Ochsenfeld, *ACS Cent. Sci.* **2024**, *10*, 302–314.

[89] T. Huber, A. E. Torda, W. F. Van Gunsteren, *J. Comput. Aid. Mol. Des.* **1994**, *8*, 695–708.

[90] S. Marsili, A. Barducci, R. Chelli, P. Procacci, V. Schettino, *J. Phys. Chem. B.* **2006**, *110*, 14011–14013.

[91] J. K. Whitmer, C.-c. Chiu, A. A. Joshi, J. J. De Pablo, *Phys. Rev. Lett.* **2014**, *113*, 190602.

[92] H. Sidky, J. K. Whitmer, *J. Chem. Phys.* **2018**, *148*, 104111.

[93] J. Henin, T. Lelievre, M. R. Shirts, O. Valsson, L. Delemotte, *Living J. Comp. Mol. Sci.* **2022**, *4*, 1583.

[94] O. Valsson, P. Tiwary, M. Parrinello, *Annu. Rev. Phys. Chem.* **2016**, *67*, 159–184.

[95] G. Bussi, A. Laio, *Nat. Rev. Phys.* **2020**, *2*, 200–212.

[96] G. Bussi, A. Laio, M. Parrinello, *Phys. Rev. Lett.* **2006**, *96*, 090601.

[97] J. F. Dama, M. Parrinello, G. A. Voth, *Phys. Rev. Lett.* **2014**, *112*, 240602.

[98] P. Tiwary, M. Parrinello, *J. Phys. Chem. B.* **2015**, *119*, 736–742.

[99] F. Giberti, B. Cheng, G. A. Tribello, M. Ceriotti, *J. Chem. Theory. Comput.* **2019**, *16*, 100–107.

[100] T. M. Schäfer, G. Settanni, *J. Chem. Theory. Comput.* **2020**, *16*, 2042–2052.

[101] J. Ono, H. Nakai, *Chem. Phys. Lett.* **2020**, *751*, 137384.

[102] A. Barducci, M. Bonomi, M. Parrinello, *WIREs Comput. Mol. Sci.* **2011**, *1*, 826–843.

[103] M. Invernizzi, M. Parrinello, *J. Chem. Theory Comput.* **2022**, *18*, 3988–3996.

[104] E. A. Carter, G. Ciccotti, J. T. Hynes, R. Kapral, *Chem. Phys. Lett.* **1989**, *156*, 472–477.

[105] W. K. den Otter, *J. Chem. Phys.* **2000**, *112*, 7283–7292.

[106] G. Ciccotti, R. Kapral, E. Vanden-Eijnden, *ChemPhysChem* **2005**, *6*, 1809–1814.

[107] J. Henin, *J. Chem. Theory. Comput.* **2021**, *17*, 6789–6798.

[108] J. Henin, G. Fiorin, C. Chipot, M. L. Klein, *J. Chem. Theory. Comput.* **2010**, *6*, 35–47.

[109] H. Fu, X. Shao, C. Chipot, W. Cai, *J. Chem. Theory. Comput.* **2016**, *12*, 3506–3513.

[110] D. E. Shaw, R. O. Dror, J. K. Salmon, J. Grossman, K. M. Mackenzie, J. A. Bank, C. Young, M. M. Deneroff, B. Batson, K. J. Bowers, E. Chow, M. P. Eastwood, D. J. Ierardi, J. L. Klepeis, J. S. Kuskin, R. H. Larson, K. Lindorff-Larsen, P. Maragakis, M. A. Moraes, S. Piana, Y. Shan, B. Towles in Proceedings of the conference on high performance computing networking, storage and analysis, **2009**, pp. 1–11.

[111] L. Bonati, E. Trizio, A. Rizzi, M. Parrinello, *J. Chem. Phys.* **2023**, *159*, 014801.

[112] S. Yang, J. Nam, J. C. B. Dietschreit, R. Gómez-Bombarelli, *J. Chem. Theory Comput.* **2024**, *20*, 6559–6568.

[113] J. Zhang, L. Bonati, E. Trizio, O. Zhang, Y. Kang, T. Hou, M. Parrinello, *J. Chem. Theory Comput.* **2024**, *20*, 10787–10797.

[114] P. Kang, E. Trizio, M. Parrinello, *Nat. Comput. Sci.* **2024**, *4*, 451–460.

[115] I. T. Jolliffe, *Principal component analysis for special types of data*, Springer, **2002**.

[116] W. Jia, M. Sun, J. Lian, S. Hou, *Complex Intell. Syst.* **2022**, *8*, 2663–2693.