Advanced Techniques for the Computer Simulation and Analysis of Biomolecular Systems

Johannes C. B. Dietschreit

München 2020

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

Advanced Techniques for the Computer Simulation and Analysis of Biomolecular Systems

Johannes Carl Bertold Dietschreit

aus

Berlin

2020

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, 15.12.2020 _____

(Johannes Dietschreit)

Dissertation eingereicht am:	30.09.2020
1. Gutachter:	Prof. Dr. Christian Ochsenfeld
2. Gutachter:	Prof. Dr. Regina de Vivie-Riedle
Mündliche Prüfung am:	23.11.2020

To my family who gave me everything I needed,
to my mentors who taught me what I know,
and to my friends who were there for me when it mattered.

Acknowledgement

Here, I would like to mention and thank those who have enabled me to write this thesis, who have motivated and inspired me over the years.

I would like to thank **Prof. Dr. Christian Ochsenfeld** for the opportunity of doing my doctorate thesis in his group and for the exciting topics that were offered to me.

Further, my thanks go to **Prof. Dr. Regina de Vivie-Riedle** for being my second examiner.

I would like to thank **Dr. Laurens Peters** for the productive time we have had and together with my friend **Gökçen Savaşçı** for their help when I needed it the most. I also want to acknowledge **Beatriz von der Esch**, **Eli Naydenova**, and **Dr. Sigurd Vogler** for our joint work on projects, and the good time we had at work as well as otherwise.

Additionally, I would like to thank all members of the **Ochsenfeld group** for the welcoming atmosphere and inspiring discussions over the years.

I also want to acknowledge **Dennis J. Diestler** for being a friend and a mentor over all these years.

Finally, I want to thank all members of my family and my friends for their continuing support.

Abstract

The Helmholtz free energy is one of the central quantities of classical thermodynamics, as it governs important chemical properties such as equilibria or reaction kinetics. It is, therefore, a desirable quantity to measure, predict, and understand. Unsurprisingly, many methods exist to compute free energy differences between two states of a system. In this thesis, the density of states integration method (DSI) is developed; it detects which subsystems mainly contribute to the free energy difference. The method utilizes the velocity density of states function (VDoS) of each atom to calculate its contribution to the vibrational free energy. It is possible without any approximation to assign fractions of the vibrational free energy to meaningful subsystems, where the local free energy difference is the sum over all atoms comprising that subsystem. In this way, large local changes can be identified (free energy hot-spots), which is crucial for the understanding of free energy differences. The validity and usefulness of DSI is shown via several examples and comparison with state of the art free energy methods.

In addition to the development of DSI, this thesis also focuses on free energy barriers in the context of investigating the reaction mechanism of Sirtuin 5, a lysine deacylase class III. The relationship between the configuration of the enzyme's active site and the height of the reaction barrier is studied by computing minimal energy paths for the catalyzed reaction starting from many different (educt) configurations. Using the power of machine learning, atom-atom distances influencing the activation barrier are identified, allowing for a comprehensive understanding of the interplay of the substrate and residues within the active site of Sirtuin 5. Subsequently, we set out to compute the free energy as a function of the reaction coordinate instead of a minimum energy path.

Another theme of this thesis is the computation of spectroscopic observables in a cost effective manner while simultaneously including important features of the experimental setup. The inclusion of solvent molecules and finite temperature effects has a decisive effect on the accuracy of the computed observables. In this context, we highlight the importance of sampling atomic configurations (with and without explicit solvent) and the non-negligible influence of electron correlation on the accuracy of computed observables. Simulation protocols are developed that enable sampling, the inclusion of correlation methods, and large quantum mechanical subsystems at a low computational cost.

List of Publications

This is a cumulative dissertation, comprising five articles in peer-reviewed journals (I-III, V-VI) and one manuscript (IV). Their complete contents can be found in Chapter 3 together with their corresponding supporting information. In the following, all articles are stated together with the author's contribution to each of them.

I L. D. M. Peters, **J. C. B. Dietschreit**, J. Kussmann, and C. Ochsenfeld
"Calculating free energies from the vibrational density of states function: Validation and critical assessment"
J. Chem. Phys. **2019**, *150*, 194111
Contribution by the Author: *Conjoint conception with L. D. M. Peters, majority of the derivation, analysis of the simulation data, and writing the manuscript.*

II **J. C. B. Dietschreit**, L. D. M. Peters, J. Kussmann, and C. Ochsenfeld
"Identifying Free Energy Hot-Spots in Molecular Transformations"
J. Phys. Chem. A **2019**, *123*, 2163-2170
Contribution by the Author: *Conjoint conception with L. D. M. Peters, classical MD simulations, analysis of simulation data, as well as writing the manuscript.*

III B. von der Esch, **J. C. B. Dietschreit**, L. D. M. Peters, and C. Ochsenfeld
"Finding Reactive Configurations: A Machine Learning Approach for Estimating Energy Barriers Applied to Sirtuin 5"
J. Chem. Theory Comput. **2019**, *15*, 6660-6667
Contribution by the Author: *Conjoint conception and writing the manuscript.*

IV **J. C. B. Dietschreit**, B. von der Esch, and C. Ochsenfeld
"QM/MM Free Energy Investigation of the Initial Step of the Desuccinylation Reaction Catalyzed by Sirtuin 5 Points Towards a Conserved Mechanism among Sirtuins"
Contribution by the Author: *Participation in designing the research question, method implementation, umbrella sampling simulations, and writing the manuscript.*

V **J. C. B. Dietschreit**, A. Wagner, T. A. Le, P. Klein, H. Schindelin, T. Opatz, B. Engels, U. A. Hellmich, and C. Ochsenfeld
"Predicting ^{19}F NMR chemical shifts: A combined computational and experimental study of a trypanosomal oxidoreductase-inhibitor complex"
Angew. Chem. Int. Ed. **2020**, *59*, 12669-12673
Contribution by the Author: *Participation in designing the research question, all ^{19}F NMR calculations, and writing the manuscript.*

VI S. Vogler, **J. C. B. Dietschreit**, L. D. M. Peters, and C. Ochsenfeld
"Important Components for Accurate Hyperfine Coupling Constants: Electron Correlation, Dynamic Contributions, and Solvation Effects"
Mol. Phys. **2020**, e1772515
Contribution by the Author: *Participation in designing the research question, computations regarding solvent effects, data evaluation, plotting, and writing the manuscript.*

Further publications:

XI C. Glas, **J. C. B. Dietschreit**, N. Wössner, L. Urban, E. Ghazy, W. Sippl, M. Jung, C. Ochsenfeld, and F. Bracher
"Identification of the subtype-selective Sirt5 inhibitor balsalazide through systematic SAR analysis and rationalization via theoretical investigations"
Eur. J. Med. Chem. **2020**, *206*, 112676

X J. Egli, T. Schnitzer, **J. C. B. Dietschreit**,
C. Ochsenfeld, and H. Wennemers
"Why Proline? Influence of Ring-Size on the Collagen Triple Helix"
Org. Lett. **2020**, *22*, 348-351

IX E. Naydenova, **J. C. B. Dietschreit**, and C. Ochsenfeld
"Reaction Mechanism for the N-Glycosidic Bond Cleavage of 5-Formylcytosine by Thymine DNA Glycosylase"
J. Phys. Chem. B **2019**, *123*, 4173-4179

VIII **J. C. B. Dietschreit**, D. J. Diestler, and E.-W. Knapp
"Chemically Realistic Tetrahedral Lattice Models for Polymer Chains: Application to Polyethylene Oxid"
J. Chem. Theory Simul. **2016**, *12*, 2388-2400

VII **J. C. B. Dietschreit**, D. J. Diestler, and E.-W. Knapp
"Models for self-avoiding polymer chains on the tetrahedral lattice"
Macromol. Theory Simul. **2014**, *23*, 452-463

Contents

Chapter 1

Introduction

The overall goal of this work is to comprehend the macroscopic behaviour of complex (bio-)molecular systems in terms of their microscopic properties. For this purpose, statistical thermodynamics are employed. In practice, complex systems are handled by either Monte Carlo (MC) [1] or Molecular Dynamics (MD) [2–5] simulations. These techniques use a number of algorithms that generate sequences (trajectories) of atomic configurations. They can be applied to either gaseous, liquid, or solid phase systems; simulations are no longer limited to a few atoms, and can today be performed for systems of impressive size [6].

Crucial ingredients in the modelling of a molecular system are i) the choice of mechanics used, meaning either classical or quantum mechanics, and ii) which further approximations are made with regard to the potential energy function. An atomic model requires a quantum mechanical (QM) description. It has been shown, however, that, for many biochemical problems, molecular mechanics (MM) (i.e., classical mechanics) are a reasonable choice. However, accurate observables and reaction energies require quantum mechanics, as, for example, in the case of bond breaking, which is difficult to describe classically. Especially biomolecular systems consist of so many atoms that it is computationally infeasible to model them completely through QM. A compromise can be reached by partitioning the system into QM and MM regions, where only those atoms are included in the QM region that are important for the computation of a specific property. This combined approach, dubbed QM/MM [7], has earned its inventors the Nobel prize [8].

MD simulation has become a workhorse and is employed in many different fashions in this thesis, ranging from only MM-MD, over QM/MM-MD, to exclusively *ab initio* MD (AIMD). The Born-Oppenheimer approximation [9] has been invoked when the system includes a QM part (i.e., the nuclei are treated classically). Hence, Newton's equations of motion have been solved to propagate the nuclear coordinates in time, and not the Schrödinger equation [10].

The research of this thesis is divided into two parts. One part uses existing protocols (a sequence of computational steps) or designs new ones to compute observables accurately in an economical manner. The employed modelling process aims to include as many features of the experiment as possible, while at the same time keeping the computational costs low. Such experiment-oriented protocols are important as they strengthen the link between experimental and computational chemistry.

The other part of the research focuses on the Helmholtz free energy, the central quantity of thermodynamics. Since it governs important chemical processes, e.g., equilibria and reaction kinetics, it is an important quantity to calculate. A novel method was developed that can be used to divide a system into subsystems for which the free energy is estimated. Linking contributions to the free energy to defined subsystems offers valuable information for theoreticians and experimentalists alike, and helps them interpret the size of the overall free energy.

Though the free energy is extremely useful, it is also very time consuming to compute, as its value corresponds to a multidimensional integral over all variables that constitute a system. MC or MD simulations can be used to sample the system, but the Helmholtz free energy often converges either slowly or to the wrong result, therefore, algorithms have been developed to speed up the convergence of such simulations [11, 12]. AIMD or QM/MM-MD simulations can still be so costly that free energies have to be avoided all together. One often resorts to searching for minimal energy configurations, which can still be computationally expensive, but do not require sampling. However, the large number of degrees of freedom in a biomolecular system make such minimisations difficult, ergo global minima become hard, if not impossible, to find; one has access to only many local minima. This means that large numbers of atoms make sampling at high levels of theory too costly and minimisations at any level of theory almost meaningless. In this work, we have harnessed the power of machine learning algorithms to ameliorate the problem of investigating complex systems. Machine learning is a quickly growing field, which focuses on monitoring many variables simultaneously and extracting information from them (see section 2.4).

This cumulative dissertation is based on five publications and one manuscript, and it is structured as follows. Chapter 2 presents the theory which forms the basis of all these papers, Chapter 3 contains the publications, which present the results of this dissertation, and Chapter 4 concludes this work together with an outlook. The manuscripts are summarized briefly below.

Publication I introduces a method by which the vibrational part of the free energy can be assigned to subsystems within a system. The method, which utilizes the vibrational density of states (VDoS) function, can also be used to compute free energy differences between different states of the system. It builds upon a paper by Berens *et al.* [13]. The VDoS function is multiplied with a weighting function (derived from the harmonic oscillator model) and integrated to obtain the vibrational free energy. Therefore, we have coined the term density of states integration (DSI) for this method. Subsystems that predominantly contribute to these differences are identified as free energy hot-spots. In this publication, we derive the algorithm and compare its numerical accuracy and convergence behaviour with existing and frequently used free energy techniques, namely exponential averaging (EXP) [14] and Bennett's acceptance ratio (BAR) [15]. The analysis shows that free energy differences are described correctly espe-

cially when the system behaves harmonically. Problems arise when vibrations include transitions over potential energy maxima. However, the results are still qualitatively correct and help to interpret the total change in free energy, which can be reliably computed by other algorithms (e.g., EXP or BAR).

Publication II uses the DSI and showcases its strength for two examples. The first analyzes a standard MM problem, the binding of an inhibitor to the bromo- and extra-terminal domain [16, 17]. The vibrational free energy hot-spots enable easy visualization of the effects of inhibitor binding. The protein residues, mainly affected by the presence of the inhibitor in the binding pocket, can be clearly identified. Interactions stabilizing and destabilizing the complex can be distinguished. The second example demonstrates a famous quantum effect in organic chemistry, namely the anomeric effect [18]. Its roots are controversial [19–22]. Our hot-spots method allows to identify all atoms which take part in this effect and how they do so. It is thus a powerful tool to analyze the locality of effects. Additionally, the results for both examples align well with chemical intuition.

Publication III focuses on the utilization of QM/MM modelling in the description of enzymatic reactions. The first reaction step of the desuccinylation reaction catalyzed by Sirtuin 5 [23, 24] is investigated. First, MM-MD is used to sample configurations of the fully solvated enzymatic system together with its reactant (a short peptide) and co-factor NAD+. Many frames are taken from the MD trajectory as starting configurations for the minimal energy path along the reaction coordinate (adiabatic mapping). HF-3c/minix [25], a low-cost approximation of solving the Schrödinger equation, is used. The computed reaction barriers scatter over a large range of energies, as is expected when using many different configurations as educts for reaction path simulations. In a small benchmark, we computed the reaction barrier for several starting configurations not only with HF-3c but also with higher level methods. The benchmark shows that HF-3c overestimates the energy barrier height in comparison to the higher level methods, but in a predictable manner. Machine learning is used to connect the value of the activation barrier to the configuration within the active site and to determine geometrical parameters important for a near attack conformation [26–28], a conformation which a molecular system has to assume in order for a reaction to happen. It confirms the findings by Ryde [29] that barrier heights are broadly distributed and very large numbers of configurations (MM trajectory frames) are needed to obtain a reliable estimate of the effective free energy barrier.

In **Manuscript IV** we build upon **Publication III** and investigate the first reaction step catalyzed by Sirtuin 5 further by using advanced sampling methods. Umbrella sampling [12] is combined with MBAR [30], i.e., biased QM/MM-MD simulations along the reaction coordinate are performed and afterwards the free energy surface is extracted by removing the influence of the bias *a posteriori*. It is found that in Sirtuin 5 the first step of the dicarboxylic acid deacylation proceeds via a concerted S_N2 mechanism, as it does for the deacetylation in homologous enzymes [31, 32]. The comparison of the previously obtained minimal energy paths and the calculated free energy barrier underlines that conclusions on a reaction mechanism or a reaction barrier height based on minimal energy paths have to be taken with care. Subsequently, a reweighting of the free energies from HF3c/minix to the popular B3LYP-D3/def2-svp [33–38] is attempted to obtain a more reasonable barrier height. The reweighting fails, however, due to lacking overlap of the configuration spaces associated with the two methods. A solution

will have to be found in future work as otherwise higher level QM-based free energies or corrections will remain inaccessible.

In **Publication V**, the chemical shift of a ^{19}F nucleus in a covalent inhibitor targeting the oxidoreductase tryparedoxin of Trypanosoma brucei [39, 40] has been calculated. The paper presents a multi-disciplinary study in which experimental NMR measurements and theoretical computations are used to determine the predominant conformation of the inhibitor-enzyme complex in solution. The starting point of the computational investigation forms a co-crystal structure in which two different binding poses are present [41]. MM-MD simulations are used to sample configurations of the solvated system and the NMR shieldings are computed in a QM/MM ansatz based on the MM trajectories. This study is the largest to date in terms of numbers of QM atoms and configurations.

Publication VI deals with another spectroscopic observable, namely hyperfine coupling constants (HFCC). They are an important property of radicals. Here, methods previously developed in the Ochsenfeld group [42, 43] were used to enable fast and accurate HFCC computations. In this publication we provide a protocol by which observables, not only HFCCs, can be computed in a cost effective manner without oversimplifying the experimental conditions. The impact of three contributions to the observable was analyzed. It is shown that electron correlation, finite temperature sampling, and modelling of the molecular environment (solvation) are crucial to a reliable comparison with experiment. Those three contributions can be included separately in an additive manner, circumventing expensive calculations that include all three simultaneously.

Chapter 2

Theory

In the first section of this chapter, the basic equations of statistical mechanics will be reviewed, many of which can be found in standard textbooks [44–46]. The second section will present a detailed analysis of motions of molecules and how free energy expressions can be derived from them. It will also outline the theory behind vibrational hot-spots, the vibrational density of states integration method (DSI). The third section concentrates on numerical algorithms to estimate free energy differences. All applications presented in this thesis and all formulas derived in this chapter assume equilibrium conditions. Non-equilibrium methods based on the Jarzinsky equality [47, 48] and the Crooks fluctuation theorem [49, 50] will not be discussed.

The fourth section will give a brief introduction to machine learning. The fifth and final section highlights the importance of vibrational motions in computing observables.

2.1 Basics of Statistical Thermodynamics

2.1.1 Principal Idea and Ergodicity

The macrostate of a system is defined through macroscopic variables such as volume, pressure, temperature, energy, or number of particles. In contrast, a microstate describes the microscopic configuration of the system, meaning the instantaneous value of all coordinates and momenta. Over time, the system moves from one microstate to another exploring all accessible states, which are constrained by the macrostate. This means that a measurement in the lab, which is usually performed on a long time scale relative to the thermal fluctuations between microstates, measures an average value over many microstates. In a thought experiment, the same average result could be achieved by performing many measurements of the observable O, which probe only one microstate

each.

$$O_{\text{observed}} = \frac{1}{M}\sum_{i=1}^{M} O_i \tag{2.1}$$

Here, O_i denotes the value of the observable in measurement i, and M is the number of measurements. If the number of these hypothetical measurements is large enough, one can transform this sum into a weighted average where the weights, P_i, correspond to the relative occurrence of the microstates

$$O_{\text{observed}} = \sum_{i\in\mathcal{S}} P_i O_i = \langle O \rangle \ , \tag{2.2}$$

where $\mathcal{S}$ is the set of all microstates that fulfill the constraints of the system's macrostate. This set is called *ensemble* and eq. (2.2) therefore *ensemble average*. Thus, monitoring the time evolution of a system over a long time period yields the same result as the ensemble average of microstates:

$$\sum_{i\in\mathcal{S}} P_i O_i \overset{!}{=} \lim_{t\to\infty} \frac{1}{t}\sum_{\tau=0}^{t} O(\tau) \tag{2.3}$$

The relation in eq. (2.3) is termed *ergodicity*. It is generally assumed to be valid in experiments as well as simulations, but it cannot – except in rare cases – be proven [51].

Unfortunately, there exist many cases in which ergodicity is not fulfilled, e.g., high energy barriers cannot be passed and prohibit the system to visit all microstates. Generally, simulation methods to overcome such barriers are called enhanced sampling techniques and are important to ensure the generation of proper ensembles. They will be briefly discussed in Section 2.3.5.

2.1.2 Microcanonical Ensemble

The microcanonical ensemble describes the collection of all microstates in a system where the macroscopic variables N (number of particles), V (volume), and E (total energy) are fixed. As the total energy is constant, all states have the same probability

$$P_i = \frac{1}{\Omega(N,V,E)\mathrm{d}E} \ . \tag{2.4}$$

$\Omega \mathrm{d}E$ is the number of microstates with energy $E \pm \frac{\mathrm{d}E}{2}$, which also fulfill the additional constrains V and N. In this case, the entropy S is proportional to the number of available states

$$S = k_{\mathrm{B}} \ln\left(\Omega \mathrm{d}E\right) \ . \tag{2.5}$$

2.1.3 Canonical Ensemble

Under conditions relevant to most experimental setups, the temperature is fixed, but the energy can fluctuate through exchange with the surroundings. An ensemble with fixed N, V, and T (absolute temperature) is termed the *canonical ensemble*. As not all

microstates of the ensemble have the same energy, they occur with different probabilities

$$P_i \propto e^{-\beta E_i} \ , \tag{2.6}$$

with E_i the energy of microstate i and $\beta = \frac{1}{k_\mathrm{B}T}$ the inverse thermal energy, with k_B being the Boltzmann constant. This form of the weight P_i is called the *Boltzmann factor*. As the probability function has to be normalized, one finds

$$P_i = \frac{e^{-\beta E_i}}{\sum\limits_{j \in \mathcal{S}} e^{-\beta E_j}} = Q^{-1} e^{-\beta E_i} \ . \tag{2.7}$$

The sum of all Boltzmann weights Q is called the *canonical partition function*. It can be expressed as the Laplace transform of the microcanonical density of states function [52].

$$Q(N, V, T) = \int \mathrm{d}E \ \Omega(N, V, E) \ e^{-\beta E} \tag{2.8}$$

Here Q was replaced by an integral under the assumption that the energy levels are distributed continuously, but it can also expressed as a sum over discrete energy levels. Hence, the value of $\Omega(N, V, E)\mathrm{d}E$ equals the degeneracy of an energy level. Boltzmann weighted averages are performed in the same manner as indicated in eq. (2.2) for the microcanonical ensemble.

The partition function is the central quantity in statistical thermodynamics from which the various thermodynamic properties can be derived. One of the most important properties is the (Helmholtz) free energy:

$$A = \langle E \rangle - TS = -\beta^{-1} \ln Q \tag{2.9}$$

The left hand side of the equation is the thermodynamic definition of the Helmholtz energy, the right hand side connects it to the partition function. The ensemble average of the energy is called the internal energy, sometimes denoted by U. To get a more explicit expression for the entropy, one can use the partial derivative of the free energy with respect to T, based on a fundamental thermodynamic relation, the total differential of A

$$\mathrm{d}A = -S \ \mathrm{d}T - p \ \mathrm{d}V + \mu \ \mathrm{d}N \ ,$$

where p is the total pressure and μ the chemical potential.

$$
\begin{aligned}
S &= -\left(\frac{\partial A}{\partial T}\right)_{N,V} \\
&= k_B \ln Q + k_B T \left(\frac{\partial \ln Q}{\partial T}\right)_{N,V} \\
&= k_B \left(\ln Q + \frac{T}{Q} \sum_{i \in \mathcal{S}} \frac{E_i}{K_B T^2} e^{-\beta E_i}\right) \\
&= k_B \left(\ln Q + \sum_{i \in \mathcal{S}} \beta E_i P_i\right) \\
&= k_B \left(\sum_{i \in \mathcal{S}} P_i \ln Q - \sum_{i \in \mathcal{S}} P_i \ln e^{-\beta E_i}\right) \\
S &= -k_B \sum_{i \in \mathcal{S}} P_i \ln P_i
\end{aligned} \tag{2.10}
$$

The last line is the Gibbs entropy formula, which also holds for the microcanonical ensemble (compare eq. (2.5)).

2.1.4 Problems in Defining Microstates

For large systems, it can be cumbersome to work with microstates, as their number is extremely high. Additionally, in the case of classical systems, the microscopic configuration of coordinates and momenta is of little interpretational value. They are nonetheless of interest, as the transition from one microstate to another is at the core of kinetic theories such as Markov State Models [53–55]. Therefore, one has to define the microstates. This is commonly done by projecting out many degrees of freedom from the high dimensional coordinate and momenta space, and subsequently characterising regions in the lower dimensional space as microstates.

For small systems, a partitioning into microstates can often be done empirically. For example, in the case of a small dipeptide (doubly capped amino acid) in solution, one can partition the plane spanned by the two central dihedral angles ϕ and ψ (Ramachandran plane) into three ranges α, β, and L_α. The exchange between those three ranges captures the main kinetics of the backbone, which is important in studying protein motions. However, this completely neglects any vibrational motion or movement of the side chains or surrounding solvent [56, 57]. Thus, when defining the states empirically, one has to focus on those degrees of freedom relevant to the research question.

This approach can usually not be used for a large system, as the number of degrees of freedom is exceedingly large. Here, clustering approaches are useful [58–60]. All in all, determining sensible microstates is very difficult and subject of its own research area. Hence, it will not be discussed any further here.

Instead of expressing the Boltzmann distribution as a function of microstates, we will recast P_i as weight of an energy level E_i with degeneration g_i (number of microstates with the same energy).

$$
P_i = \frac{g_i}{Q} e^{-\beta E_i} \tag{2.11}
$$

2.1.5 Phase Space

In general, the motions of a classical system are governed by Newton's equations of motion

$$-\nabla U = m\vec{a}$$

where U is the potential energy (U is used to avoid confusion with the volume V), which depends only on particle positions.

As all simulations performed in this thesis were either purely classical or Born-Oppenheimer MD, the motion of a particle can be fully described by its position $\vec{r}$ and its conjugate momentum $\vec{p}$. The total energy, $\mathcal{H}$, is thus a function of the set of all N positions, $\vec{r}^N$, and momenta, $\vec{p}^N$.

$$\mathcal{H}\left(\vec{r}^N, \vec{p}^N\right) = K\left(\vec{p}^N\right) + U\left(\vec{r}^N\right) \tag{2.12}$$

Here, K denotes the kinetic energy (K is used instead of T to avoid confusion with the absolute temperature T) which depends solely on the momenta.

The space spanned by the set of positions and momenta is called *phase space*. It contains all possible configurations and momenta of the system. From here on, all equations will be formulated in terms of integrals over phase space.

$$Q = \sum_{i \in \mathcal{S}} \ldots \quad \Rightarrow \quad Q \propto \int \mathrm{d}\vec{r}^N \int \mathrm{d}\vec{p}^N \; e^{-\beta \mathcal{H}\left(\vec{r}^N, \vec{p}^N\right)} \tag{2.13}$$

The value of the prefactor depends on the composition of the system. For a system of N indistinguishable particles it is $\frac{1}{N! h^{3N}}$ [44–46].

The integral on the right hand side of eq. (2.13) makes it much clearer than the sum on the left hand side that the exploration of the entire phase space during one computer simulation is virtually impossible. Even very long simulations on modern high performance computers do not explore all of phase space, as the potential energy surface (PES) is often rough and prohibits uniform sampling. A system spends most of the time in those regions of phase space which correspond to a local minimum on the PES. That is, however, not a violation of eq. (2.3), as any average is dominated by those parts of phase space with a low total energy, since they have the largest weights.

Note that the partition function can be expressed in terms of microstates, energy levels, or phase space. It is sometimes convenient to switch between those expressions, as one can be more informative than the other.

2.1.6 Integration of Kinetic Energy

As a Boltzmann factor is an exponential of the total energy, it factorizes since the energy can be written as a sum of kinetic and potential energies, and thus we can separate the multidimensional integral into coordinate and momentum parts. The integral over coordinates contains the potential energy, which can rarely be evaluated analytically. The kinetic energy, however, has always the same form and the corresponding integral can be done analytically.

$$
\begin{aligned}
Q &= \frac{1}{N!h^{3N}} \int \mathrm{d}\vec{p}^N \; e^{-\beta K\left(\vec{p}^N\right)} \int \mathrm{d}\vec{r}^N \; e^{-\beta U\left(\vec{r}^N\right)} \\
&= \frac{1}{N!h^{3N}} \int_{-\infty}^{\infty} \mathrm{d}p_x^N \int_{-\infty}^{\infty} \mathrm{d}p_y^N \int_{-\infty}^{\infty} \mathrm{d}p_z^N \; e^{-\beta\left(\sum_{i=1}^{N} \frac{p_{i,x}^2+p_{i,y}^2+p_{i,z}^2}{2m_i}\right)} \int \mathrm{d}\vec{r}^N \; e^{-\beta U\left(\vec{r}^N\right)} \\
&= \frac{1}{N!} \prod_{i=1}^{N} \left(\frac{2\pi m_i}{h^2\beta}\right)^{3/2} \int \mathrm{d}\vec{r}^N \; e^{-\beta U\left(\vec{r}^N\right)} \\
&= \frac{1}{N!} \prod_{i=1}^{N} \Lambda_i^{-3} \int \mathrm{d}\vec{r}^N \; e^{-\beta U\left(\vec{r}^N\right)}
\end{aligned} \tag{2.14}
$$

Λ_i is called *thermal de Broglie wavelength* of particle i with mass m_i. If no particles (atoms) are destroyed, created, or change their mass, this remains the same when we manipulate the potential. Thus, we will focus on the integral over configuration space $\left\{\vec{r}^N\right\}$.

$$
Q \propto \int \mathrm{d}\vec{r}^N \; e^{-\beta U\left(\vec{r}^N\right)} \tag{2.15}
$$

The configurational integral is often abbreviated with Z.

2.1.7 Free Energy Differences

Equation (2.14) shows that evaluating Q exactly is very difficult [61]. Even if the system is constrained, the phase space spanned is still enormous. Hence, obtaining the free energy from eq. (2.9), is impractical.

Free energy differences are sometimes more easily obtained:

$$
\Delta A = -\beta^{-1} \ln \frac{Q_1}{Q_0} \tag{2.16}
$$

ΔA denotes the difference in free energy between two systems 0 and 1, Q_0 and Q_1 are the respective partition functions. This definition is very general. For example, the two systems could be the same, except for different constraints like an added potential, they could refer to different parts of phase space along a reaction coordinate, or even to different systems. Here, we focus only on changes within the canonical ensemble, meaning the total number of particles N always stays constant. The fraction in eq. (2.16) is numerically more stable than the difference of two separately computed free energies.

If only the potential energy function changes the fraction reduces to

$$
\begin{aligned}
\frac{Q_1}{Q_0} &= \frac{\frac{1}{N!h^{3N}} \int \mathrm{d}\vec{p}^N \; e^{-\beta K\left(\vec{p}^N\right)} \int \mathrm{d}\vec{r}^N \; e^{-\beta U_1\left(\vec{r}^N\right)}}{\frac{1}{N!h^{3N}} \int \mathrm{d}\vec{p}^N \; e^{-\beta K\left(\vec{p}^N\right)} \int \mathrm{d}\vec{r}^N \; e^{-\beta U_0\left(\vec{r}^N\right)}} \\
&= \frac{\int \mathrm{d}\vec{r}^N \; e^{-\beta U_1\left(\vec{r}^N\right)}}{\int \mathrm{d}\vec{r}^N \; e^{-\beta U_0\left(\vec{r}^N\right)}} \; .
\end{aligned} \tag{2.17}
$$

If the mass of one or more particles changes, the corresponding integrals over the kinetic energy no longer cancel, which was for example the case in **Publication I**, where we corrected the fraction of configuration integrals accordingly.

$$\frac{Q_1}{Q_0} = \frac{\prod_i^{3N_\mathrm{A}} \sqrt{\frac{2\pi m_i^1}{\beta}} \int \mathrm{d}\vec{r}^N \; e^{-\beta U_1(\vec{r}^N)}}{\prod_i^{3N_\mathrm{A}} \sqrt{\frac{2\pi m_i^0}{\beta}} \int \mathrm{d}\vec{r}^N \; e^{-\beta U_0(\vec{r}^N)}} \tag{2.18}$$

All thermal wavelengths of particles that retained their mass cancel out; only those that differ between the two states 0 and 1 remain.

It is easier to calculate free energy differences for similar systems (or states) than the free energy itself. The integral over those parts of phase space with a high energy are going to be similarly small for both states, such that only the ratio of the sum of large Boltzmann factors is important. This ratio converges faster, as those are the more frequently visited regions of phase space. Mathematical manipulations that aide the numerical stability of this fraction will be presented in Section 2.3.

2.2 The Vibrational Free Energy Hot-Spots Method

Here, the theory behind the analysis method that we term "vibrational free energy hot-spots", which has been published in **Publications I** and **II**, is derived. The method is also dubbed density of states integration (DSI). For a complete discussion, we will also look at translational and rotational expressions which are neglected in DSI and make comparisons with the two-phase thermodynamic approach [62, 63] in Section 2.2.8.

2.2.1 Simple Systems

The simplest systems that can be described analytically with statistical mechanics are monatomic ideal gases and solids, as the former exhibit only translational and the latter only vibrational motions. A more complex system is a dilute molecular gas. The molecules not only translate, but they also rotate and vibrate internally. As the gas is presumed to be dilute, intermolecular interactions are neglected. Hence, the Hamiltonian is a sum of energies associated with those three motions. Additionally, each molecule has an electronic and nuclear energy associated with corresponding degrees of freedom. Since intermolecular interactions are absent, we can write the partition function as

$$Q = Q_{\text{trans}} \times Q_{\text{rot}} \times Q_{\text{vib}} \times Q_{\text{elec}} \times Q_{\text{nuc}} \ . \tag{2.19}$$

The electronic partition function is relatively simple, as it is a sum over the electronic states

$$Q_{\text{elec}} = \sum_{i=0}^{\infty} g_i e^{-\beta\epsilon_i} \ , \tag{2.20}$$

where ϵ_i is the energy of electronic state i and g_i is its degeneracy. This sum can be reformulated with respect to the ground state energy ϵ_0.

$$Q_{\text{elec}} = e^{-\beta\epsilon_0} \left(g_0 + \sum_{i=1}^{\infty} g_i e^{-\beta\Delta\epsilon_i} \right) \tag{2.21}$$

For most molecules and especially for all of those considered here, the energy gap $\Delta\epsilon_i$ is very large in comparison with the thermal energy, and thus, all terms in the sum in eq. (2.21) are close to zero. Therefore, only the exponential of the ground state energy and its degeneracy remain. For closed shell molecules, the ground state is not degenerate ($g_0 = 1$).

The nuclear energy levels are even further apart than the electronic ones, such that the above arguments apply here especially. The influence of degenerate ground states is usually neglected. From here on, we will not consider the nuclear partition function further.

Following the above arguments, the Helmholtz energy can be written as

$$A = \epsilon_0 + A_{\text{trans}} + A_{\text{rot}} + A_{\text{vib}} \tag{2.22}$$

In the next three sections, we will examine the latter three contributions in detail. We note here in passing that a liquid cannot be reduced to such terms easily, as it is dense. The translation and rotations of a molecule are hampered through contact with neighbouring molecules and intermolecular potentials cannot be completely neglected either. We will return to this discussion at the end of Section 2.2.

2.2.2 Translation of Non-interacting Particles

In the solid phase, translation does not occur; in the liquid phase it is strongly hindered. The only phase for which translation can be described analytically is the gas phase. Here, the model of non-interacting spheres is used, i.e., all particles move unhindered through space. The Hamiltonian for such a particle contains only the kinetic energy. Using the results from Section 2.1.6, we can write the partition function of a single particle as

$$\begin{aligned} Q_{\text{trans}} &= \frac{1}{h^3} \int \mathrm{d}\vec{p}\; e^{-\beta T(\vec{p})} \int \mathrm{d}\vec{r}\; e^{-\beta U(\vec{r})} \\ &= \frac{1}{h^3} \sqrt{\frac{2\pi m}{\beta}}^3 \int_V \mathrm{d}\vec{r}\; e^0 \\ Q_{\text{trans}} &= \frac{V}{\Lambda^3}\ , \end{aligned} \tag{2.23}$$

where Λ is again the thermal de Broglie wavelength and V is the volume available to the particle. Boltzmann statistics, as employed in this work, are only valid for particles where Q_{trans} has a value much greater than one. The same result can be derived from the model of a quantum particle in a three-dimensional box [44–46].

2.2.3 Rotation

The arguments concerning the restriction of translation also apply to rotation. In the dilute gas phase, the analytical model is the rigid rotor in an otherwise field free environment. For a spherical rotor, the quantum mechanical problem yields the energy levels

$$E_J = J(J+1)B\ , \tag{2.24}$$

where J is the quantum number and B the rotational constant, here defined in the dimension of energy.

$$B = \frac{\hbar^2}{2I}$$

I is the moment of inertia. The partition function of one molecule is the sum over Boltzmann factors of energy levels E_J

$$Q_{\text{rot}} = \sum_{J=0}^{\infty} (2J+1) e^{-\beta J(J+1)B}\ , \tag{2.25}$$

and the prefactor corresponds to the degeneracy of those levels. At room temperature, the difference between exponents becomes rather small, so that the sum can be converted into an integral.

$$\begin{aligned} Q_{\text{rot}} &= \int_0^{\infty} \mathrm{d}J\ (2J+1) e^{-\beta J(J+1)B} \\ &= \int_0^{\infty} \mathrm{d}J(J+1)\ \ e^{-\beta J(J+1)B} \\ Q_{\text{rot}} &= (\beta B)^{-1} \end{aligned} \tag{2.26}$$

For a non-spherical particle, where the moment of inertia tensor has three different components on its diagonal, the solution becomes

$$Q_{\mathrm{rot}} = \frac{\sqrt{\pi}}{\sigma}\left(\frac{T^3}{\Theta_{\mathrm{A}}\Theta_{\mathrm{B}}\Theta_{\mathrm{C}}}\right)^{1/2} . \tag{2.27}$$

σ is the symmetry number of the molecule and Θ, the rotational temperature, is defined as

$$\Theta_{\mathrm{A}} = \frac{\hbar^2}{2I_{\mathrm{A}}k_{\mathrm{B}}} ,$$

where I_{A} is the principal moment of inertia about axis A.

2.2.4 Vibration

Vibrations are most easily treated by the harmonic oscillator model. The potential energy of the harmonic oscillator is

$$V_{\mathrm{HO}} = \sum_i \frac{k_i}{2} q_i^2 , \tag{2.28}$$

where k_i are the force constants and q_i the displacements from the minimum position. In molecules, q is called a normal mode as the harmonic modes consist of collective motions which are distinct from motions of single atoms, and are therefore not denoted with $\Delta\vec{r}$. Each harmonic oscillator has a characteristic angular frequency which is directly linked to the force constant.

$$\omega = \sqrt{\frac{k}{\mu}}$$

μ is the mass associated with the motion.

For the classical harmonic oscillator, we obtain Q according to eq. (2.13)

$$\begin{aligned} Q_{\mathrm{HO}}^{\mathrm{CL}} &= \frac{1}{h}\int \mathrm{d}p \int \mathrm{d}q \; e^{-\beta\mathcal{H}(p,q)} \\ &= \Lambda^{-1}\int_{-\infty}^{\infty} \mathrm{d}q e^{-\beta\frac{k}{2}q^2} \\ &= \sqrt{\frac{2\pi\mu}{h^2\beta}}\sqrt{\frac{2\pi}{\beta k}} \\ &= \frac{2\pi}{h\beta}\sqrt{\frac{\mu}{k}} \\ &= \frac{2\pi}{h\beta\omega} \\ Q_{\mathrm{HO}}^{\mathrm{CL}} &= (\beta h\nu)^{-1} \end{aligned} \tag{2.29}$$

Hence, the free energy of a classical harmonic oscillator is

$$A_{\mathrm{HO}}^{\mathrm{CL}} = \beta^{-1}\ln\beta h\nu . \tag{2.30}$$

Thus, the free energy increases monotonously with the frequency and changes more slowly with increasing frequencies (higher force constants).

$$\frac{\mathrm{d}A_{\mathrm{HO}}^{\mathrm{CL}}}{\mathrm{d}\nu} \propto \frac{1}{\nu}$$

However, for weak force constants with $h\nu$ lower than the thermal energy, the free Helmholtz energy falls off steeply towards $-\infty$ for decreasing frequency.

In the quantal case, the solution of the time-independent Schrödinger equation yields

$$E_v = h\nu\left(\frac{1}{2} + v\right) \ , \qquad v = 0, 1, 2, \ldots \tag{2.31}$$

where v is the quantum number. Therefore, the partition function is

$$\begin{aligned} Q_{\mathrm{HO}}^{\mathrm{QM}} &= \sum_{v=0}^{\infty} e^{-\beta h\nu\left(\frac{1}{2}+v\right)} \\ &= e^{-\frac{\beta h\nu}{2}} \sum_{v=0}^{\infty} e^{-\beta h\nu v} \\ &= e^{-\frac{\beta h\nu}{2}} \frac{1}{1 - e^{-\beta h\nu}} \\ Q_{\mathrm{HO}}^{\mathrm{QM}} &= \left(e^{\frac{\beta h\nu}{2}} - e^{-\frac{\beta h\nu}{2}}\right)^{-1} \ , \end{aligned} \tag{2.32}$$

where the sum was evaluated by means of the formula for a geometric series. The free energy of the quantum harmonic oscillator is thus

$$A_{\mathrm{HO}}^{\mathrm{QM}} = \beta^{-1} \ln\left(e^{\frac{\beta h\nu}{2}} - e^{-\frac{\beta h\nu}{2}}\right). \tag{2.33}$$

Here, we notice that for large frequencies, the first term in the logarithm dominates and the free energy increases linearly with increasing frequency, not logarithmically. For small frequencies, the free energy behaves identically to the classical oscillator. But the cross over to negative free energies is only slightly shifted to $h\nu = 2\ln\left(\frac{1+\sqrt{5}}{2}\right)\beta^{-1} \approx 0.96\beta^{-1}$ and not simply $h\nu = \beta^{-1}$ as in the classical case.

For the discussion of numerical stability, it is important to underline that, in the high frequency regime, a change in the force constant or frequency of the oscillator has a small effect (especially in the classical case) on the free energy. However, in the low frequency regime, even a small change can have a huge effect.

2.2.5 Vibrational Density of States

Vibrational Density of States Function

A non-linear molecule with N_{A} atoms has 3 translational, 3 rotational, and $3N_{\mathrm{A}} - 6$ vibrational degrees of freedom. If all normal frequencies are known, the vibrational partition function becomes the product of $3N_{\mathrm{A}} - 6$ single harmonic oscillator functions

$$Q_{\mathrm{vib}} = \prod_{i=1}^{3N_{\mathrm{A}}-6} q_{\mathrm{vib}}(\nu_i) \ , \tag{2.34}$$

where ν_i is the eigenfrequency of harmonic mode i and q its partition function. According to eq. (2.9), the total free energy is a sum of the free energies of all modes.

$$\beta A_{\text{vib}} = -\sum_{i=1}^{3N_\text{A}-6} \ln q_{\text{vib}}(\nu_i) \tag{2.35}$$

This sum can be turned into an integral over frequencies by using a function $D(\nu)$ that selects the frequencies present in the system.

$$\beta A_{\text{vib}} = -\int_0^\infty \mathrm{d}\nu \; D(\nu) \ln q_{\text{vib}}(\nu) \tag{2.36}$$

$D(\nu)$ is called the *vibrational density of states* function (VDoS). It should not be confused with the density of states function Ω, mentioned in the Section 2.1.2 on the microcanonical ensemble. In the case of $3N_\text{A} - 6$ normal modes, $D(\nu)$ consists of a sum of Dirac delta functions which transforms eq. (2.36) back to eq. (2.35).

$$D(\nu) = \sum_{i=1}^{3N_\text{A}-6} \delta(\nu - \nu_i) \tag{2.37}$$

The delta functions at negative frequencies have been omitted, as we integrate only over positive values. Eq. (2.36) was originally used for solids, where the vibrations are, to a good approximation, harmonic. The same applies to molecular vibrations in the dilute gas phase. In solution, however, motions are constrained, causing vibrational motion to become anharmonic. Thus, $D(\nu)$ becomes more complex than a sum of delta functions.

For a real system the frequencies of the vibrational modes are not known *a priori*. In order to estimate A_{vib} or Q_{vib} one has to determine D. It has been shown that one can extract the VDoS from the Fourier transform of the velocity autocorrelation function [64]

$$D(\nu) = \int_{-\infty}^{\infty} \mathrm{d}t \; C(t) \; e^{-i2\pi\nu t} \; , \tag{2.38}$$

where the autocorrelation function is usually defined as:

$$C(t) = \frac{\langle v(t)v(0)\rangle}{\langle v(0)^2\rangle} \tag{2.39}$$

Following this definition, the autocorrelation function is normalized for the time lag $t = 0$ and dimensionless. The averaging indicated by the brackets is performed over different simulations or origins in time along the simulation:

$$\langle v(t)v(0)\rangle = \langle v(t+\tau)v(\tau)\rangle_\tau = \frac{1}{t_{\text{total}} - t} \int_0^{t_{\text{total}}-t} v(t+\tau)v(\tau)\mathrm{d}\tau$$

As we were interested in the relative amplitudes of the motions within a molecule and the motions of different species of atoms, we followed the derivation of Berens et al. [13] and used the definition

$$\begin{aligned} D(\nu) &= 2\beta \sum_{i=1}^{N_\text{A}} m_i \; \Re\left\{ \int_{-\infty}^{\infty} \mathrm{d}t \; \langle \vec{v}_i(t+\tau)\vec{v}_i(\tau)\rangle_\tau \, e^{-i2\pi\nu t} \right\} \\ &= 4\beta \sum_{i=1}^{N_\text{A}} m_i \; \int_0^{\infty} \mathrm{d}t \; \langle \vec{v}_i(t+\tau)\vec{v}_i(\tau)\rangle_\tau \cos(2\pi\nu t) \; , \end{aligned} \tag{2.40}$$

where m_i is the mass of atom i and $\Re$ is the real part of the Fourier transform. In all our studies, we performed the average of eq. (2.40) over at least 10 simulations. According to the equipartition theorem, the vibrational density of states function, defined in eq. (2.40), has to fulfill

$$\int_0^\infty \mathrm{d}\nu \; D(\nu) = 3N_\mathrm{A} \tag{2.41}$$

irrespective of whether a harmonic approximation is invoked or not. A detailed discussion and derivation can be found in Ref. [13], where the above expressions were developed to estimate the difference between a classical and quantum mechanical description of vibrations and to correct corresponding free energy values for nuclear quantum effects.

As a side note, our numerical experiments had a finite time resolution of 1 fs (frequency with which data were saved) which limits ν to $5 \cdot 10^{14}$ Hz or 16 678 cm^{-1}. This upper limit is sufficient for capturing all vibrational motions in molecules, as the fastest intramolecular vibrations are below 4 000 cm^{-1} [65].

There are two more frequently used methods (*vide infra*) to determine $D(\nu)$, even though their connection to the VDoS is rarely acknowledged explicitly. Both approximate the potential energy with harmonic functions and yield a result according to eq. (2.37).

$$U(\vec{r}^N) \approx U_0 + \sum_{i,j} k_{ij} \Delta x_i \Delta x_j \tag{2.42}$$

Here x_i is any Cartesian or normal mode component, Δx_i a displacement from the minimum, and U_0 the value of the potential energy at the minimum.

Normal Mode Analysis

The most commonly used method is the *normal mode analysis* (NMA) [66, 67], which calculates the Hessian Matrix

$$H_{ij} = \frac{\partial^2 U}{\partial x_i \partial x_j}$$

at the global minimum configuration. The diagonalization of $\mathbf{H}$ yields the eigenfrequencies ν_i, which are then used to calculate the vibrational free energy according to eq. (2.35). With increasing molecule size, the determination of the minimum configuration as well as the construction and diagonalization of $\mathbf{H}$ become prohibitively expensive.

Quasi-harmonic Analysis

The other method, which is in spirit similar to the DSI determination of the VDoS, is called the *quasi-harmonic analysis* (QHA) [68, 69]. It uses information from MD simulations to determine the vibrational frequencies, and thus circumvents the search for a minimum. The frequencies ν_i are obtained by diagonalizing the mass weighted covariance matrix

$$\left[\mathbf{M}^{\frac{1}{2}} \boldsymbol{\sigma} \mathbf{M}^{\frac{1}{2}} - \beta^{-1} \boldsymbol{\nu}\right] \mathbf{M}^{\frac{1}{2}} \Delta \mathbf{x} = 0 \; , \tag{2.43}$$

where $\mathbf{M}$ contains the atomic masses on the diagonal and $\boldsymbol{\sigma}$ is the covariance matrix, defined as

$$\sigma_{ij} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle \; . \tag{2.44}$$

The indicated ensemble averages are performed over all trajectory frames.

If the system is truly harmonic, all three methods yield exactly the same result. NMA calculates the effect of small vibrations around the minimum and is numerically exact for small temperatures. In the case of high temperatures, it neglects the effect of occurring anharmonicities. QHA uses the distribution of coordinates to calculate the frequencies. In the harmonic case, these distributions are Gaussian functions. Any anharmonic behaviour smears out those distributions, and therefore leads to an underestimation of the vibrational frequencies. Eqs. (2.40) and (2.36) do not suffer from this underestimation of frequencies; there, $D(\nu)$ itself is uncoupled from a harmonic approximation. The harmonic approximation enters the calculation of the vibrational free energy through the weighting function representing the partition function of the harmonic oscillator. One can interpret eq. (2.36) as a weighted combination of an infinite number of harmonic oscillators, which is not restricted to $3N_\mathrm{A}$ distinct ones. In this way, it captures deviations from perfect harmonic behaviour much better. However, movements across local maxima of the PES cannot be correctly described by any of those methods.

2.2.6 Partitioning the Vibrational Density of States Function

All aforementioned methods are similar in the way that they make use of eq. (2.36), but there is one important difference. NMA and QHA calculate normal modes, which are collective motions of the atoms in a molecule. They only allow a separation of the vibrational free energy into the energies of the modes. The definition in eq. (2.40), however, enables the determination of atomic contributions to the vibrational free energy, as it is a superposition of atomic functions

$$D(\nu) = \sum_{j=1}^{N_\mathrm{A}} D_j(\nu) \ , \tag{2.45}$$

where $D_j(\nu)$ is the VDoS of atom j. This allows the vibrational partition function to be written without any loss of generality as

$$\begin{aligned} Q_\mathrm{vib} &= \prod_{j}^{N_\mathrm{A}} \prod_{i} q(\nu_i)^{D_j(\nu_i)} \\ &= \prod_{j}^{N_\mathrm{A}} Q_\mathrm{vib}(j) \ , \end{aligned} \tag{2.46}$$

because

$$\prod_{j}^{N_\mathrm{A}} q(\nu_i)^{D_j(\nu_i)} = q(\nu_i)^{D(\nu_i)} \ . \tag{2.47}$$

Therefore, the vibrational free energy can be expressed as a sum over single atom contributions

$$\begin{aligned} A_{\text{vib}} &= -\beta^{-1} \ln Q_{\text{vib}} \\ &= -\beta^{-1} \ln \left[\prod_{j}^{N_{\text{A}}} Q_{\text{vib}}(j) \right] \\ &= -\beta^{-1} \sum_{j}^{N_{\text{A}}} \ln Q_{\text{vib}}(j) \\ &= \sum_{j}^{N_{\text{A}}} A_{\text{vib}}(j) \ . \end{aligned} \tag{2.48}$$

This partitioning of the vibrational free energy is very flexible. Without any additional approximation, it can be defined for meaningful sets of atoms, e.g., all atoms of a functional group or of an amino acid residue in a protein. This ansatz can be used as an aide in the interpretation and especially the localization of changes occurring in the system.

Note that this method focuses on the vibrational and not the total free energy, which cannot be partitioned without sometimes severe approximations [70–73].

2.2.7 Definition of Hot-Spots

As aforementioned, free energy differences are numerically easier to obtain than absolute values [61]. However, the above presented vibrational density of states method does in principle estimate absolute vibrational free energies. When interpreting large changes in the vibrational free energy computed via the VDoS, it is assumed that these changes co-localize with regions within the system that mainly contribute to the total free energy change.

When performing the density of states integration, the main obstacle are slow modes which have small frequency values and converge slower than fast modes, as longer simulation times are needed to characterize them correctly. As pointed out above, the harmonic partition function changes rapidly for frequencies with $h\nu\beta < 1$. Therefore, numerical noise and convergence problems have an especially large impact in this region of the spectrum. Computing the vibrational free energy difference between two states can only overcome the numerical instability partially. Most of the numerical noise cancels out, but deviations in the density of states spectrum due to insufficient convergence will still have a large impact on the final result. These problems were studied in detail in **Publication I**. In general, we compute the vibrational free energy change as the integral over the vibrational density of state difference spectrum.

$$\Delta D(\nu) = D_1(\nu) - D_0(\nu) \tag{2.49}$$

$$\Delta A_{\text{vib}} = -\beta^{-1} \int_0^\infty \mathrm{d}\nu \ \Delta D(\nu) \ln q_{\text{vib}}(\nu) \tag{2.50}$$

Very recently, an approach has been developed that uses the velocity autocorrelation function instead of the VDoS, which is then also combined and integrated with a

weighting function [74]. This ansatz circumvents the double integration of the velocity autocorrelation function (Fourier transform followed by integration with weighting function) and uses weighting functions for the Helmholtz free energy, which do not approach $-\infty$ for $t \to 0$. It might, therefore, present a possible improvement to our protocol with regards to numerical stability.

As we developed the DSI not for the calculation of free energy difference values, but for their interpretation, we use a property of vibrational modes that makes this method very useful, namely the locality of a mode. Fast modes are always strongly localized, e.g., bond vibrations almost exclusively include the two atoms forming the bond. Also bond angle and many dihedral angle vibrations can be considered fast and local with respect to the number of involved atoms and frequency. Much slower modes, often called lattice vibrations, occur where the whole or large parts of a molecule moves, and the mode is thus said to be delocalized. Those modes are very hard to interpret, both because of their complex behaviour and their contributions to the vibrational free energy is smeared over many atoms.

We have defined hot-spots as small regions of a molecule where the vibrational free energy per atom changes significantly between two states. Those hot-spots are easier to interpret as they are local and numerically more stable, since they correspond to changes in high frequency modes.

Differences in the internal energy or entropy can be retrieved in exactly the same manner as for the free energy, namely integrating the product of VDoS and a weighting function. To do so, one has to substitute $-\beta^{-1} \ln q_{\mathrm{vib}}(\nu)$ in the integral by the respective derivatives, $\left(-\frac{\partial \ln q_{\mathrm{vib}}(\nu)}{\partial \beta}\right)$ and $\left(\frac{\partial \beta^{-1} \ln q_{\mathrm{vib}}(\nu)}{\partial T}\right)$, as $U = \frac{\partial \beta A}{\partial \beta}$ and $S = -\frac{\partial A}{\partial T}$. This puts all these three quantities on the same footing. Usually entropy and internal energy differences are not as easily available as differences in the free energy itself [75, 76]; their values can be very informative about whether a process is more energetically or entropically driven [61].

2.2.8 Differences between DSI and the Two-Phase Model

An approach that covers all kinds of internal motions has been used by the group of Goddard [62, 63]. They designed a model which estimates the full partition function (translation, rotation, and vibration) in the same manner, by computing and integrating density of states spectra for each kind of motion D_{trans}, D_{rot}, and D_{vib}. It is called the "two-phase thermodynamics" model (2PT). First, they separate off the center of mass movement, one $\vec{v}_{\mathrm{trans}}(t)$ per molecule. Afterwards, they determine the angular velocity of each molecule as $\vec{\omega}(t)$ and subtract it from their velocity vectors as well. The remaining atomic velocities contain only vibrations. The separation of velocity components is necessary, since including translational and rotational motions in the vibrational density of states function would artificially increase low-frequency modes. Consequently, they determined three density spectra, one vibrational D_{vib} as described in eq. (2.40), one translational D_{trans} where they use the total mass of the molecule instead of an atom, and the rotational density D_{rot} according to

$$D_{\mathrm{rot}}(\nu) = 2\beta \sum_{A=1}^{3} I_A \, \Re \left\{ \int_{-\infty}^{\infty} \mathrm{d}t \;\; \langle \omega_A(t+\tau)\omega_A(\tau) \rangle_\tau \, e^{-i2\pi\nu t} \right\} \, ,$$

where I_A is again the principal moment of inertia about the A axis and ω_A is the corresponding angular frequency.

As described above, the best analytical models can be constructed for solids, which only exhibit vibrations, and gases that are so dilute that translation and rotation can be described in a field free environment. Lin et al. [62, 63] approximate the liquid state as a linear combination of a gas-like and a solid-like component. They split D_{trans} and D_{rot} into solid and gas parts, according to a "fluidicity" parameter f [62, 63]. All solid-like components and D_{vib} are multiplied by the partition function for the harmonic oscillator (eq. (2.29) or eq. (2.32)) and integrated as described above in eq. (2.36). Similarly, the gas-like translational density spectrum is multiplied by the partition function for the ideal gas (eq. (2.23)) and the rotational density spectrum is multiplied by the partition function for the rigid rotor (eq. (2.25)). The reference energy V_0 or ϵ_0 (compare eq. (2.22) or eq. (2.42)) is defined as the difference between the total energy and the vibrational energy corrected by the rotational and translational fluidicities.

$$V_0 = E^{\text{MD}} - 3\beta^{-1}N_{\text{A}}(1 - 0.5f_{\text{trn}} - 0.5f_{\text{rot}})$$

This model was employed to estimate thermodynamic properties for molecular liquids [63, 77]. We have not used the solid-gas model for two reasons: First, we do not aim to compute absolute thermodynamic properties, and second, the translational and rotational degrees of freedom do not allow for any form of localization. Any microscopic analysis can only be performed per molecule, e.g., studying differences in entropy within a mixture.

A local resolution was achieved by the group of Heyden, who studied the solvation entropy of small organic solutes [78] and proteins [79]. They introduced a grid of volume boxes ("voxels") centred around the solute and created spatially resolved VDoS spectra which they treated with the 2PT formalism to obtain a spatially resolved translational and rotational entropy function. They computed the difference between the local entropy close to the solute and the bulk value to get a spatial resolution of the solvation entropy. This helped them to distinguish strongly and weakly bound water molecules [80]. However, this local resolution can only be performed for the solvent not the solute, as long as translation and rotation are included. Due to the size of each voxel, there is only enough data for the solvent.

2.3 Coordinate-based Free Energies

Contrary to the vibrational free energy hot-spots, the methods presented in this section do not depend on the velocities, but rather on the configurations created by means of MD or MC simulations. In most of the following equations concerning free energy differences, it is assumed that the kinetic energy contributions cancel exactly. In case they do not cancel, a correction according to Section 2.1.6 has to be added.

Two kinds of free energies will be discussed in this section as they have been applied in this research. The first is the difference in free energy between two systems (0 and 1) that have different potential energy functions.

$$\beta\Delta A = -\ln\frac{Q_1}{Q_0} = -\ln\frac{\int \mathrm{d}\vec{r}^N \; e^{-\beta U_1(\vec{r}^N)}}{\int \mathrm{d}\vec{r}^N \; e^{-\beta U_0(\vec{r}^N)}} \tag{2.51}$$

Here ΔA denotes the difference $\Delta A = A_1 - A_0$. In the following, all differences with a Δ-sign denote the "forward" direction, i.e., $\Delta U = U_1 - U_0$ or $\Delta\mathcal{H} = \mathcal{H}_1 - \mathcal{H}_0$.

The second kind is the free energy as a function of a reaction coordinate ξ, which is generally a function of many atomic coordinates. For generality, it is expressed here as a function of all coordinates, $\xi = \xi(\vec{r}^N)$. Therefore, it is often referred to as collective variable. $A(\xi)$ is given as

$$\beta A(\xi) = -\ln \; Z^{-1} \int \mathrm{d}\vec{r}^N \; \delta\left(\xi(\vec{r}^N) - \xi\right) \; e^{-\beta U(\vec{r}^N)} \; . \tag{2.52}$$

2.3.1 Free Energy Perturbation Theory

Introduction

The central idea of perturbation theory can be summarized as follows. One starts with a system that can be solved exactly or with sufficient accuracy, which is called the reference or unperturbed system. The target system is then cast as a perturbation to the reference. The effect of the perturbation is expanded in a series of the perturbation parameter which is usually small. It is expected that this series converges quickly and allows truncation after the first few terms [14, 81, 82].

However, computer simulations have rendered such an expansion unnecessary and the equation behind free energy perturbation (FEP) has very little resemblance with the original perturbation formalism. The name has been kept nonetheless. In recent years, it has also been called *exponential averaging* (EXP) to avoid the term perturbation. It is a widely used technique [83–85] that also forms the basis for many other free energy difference formulas, some of which will be discussed later. To state the importance of this method: "FEP is not only the oldest but also one of the more useful, general-purpose strategies for calculating free energy differences." [61]

Derivation

The target Hamiltonian $\mathcal{H}_1$ differs from the unperturbed system $\mathcal{H}_0$ by the amount of the perturbation $\Delta\mathcal{H}$

$$\mathcal{H}_1 = \mathcal{H}_0 + \Delta\mathcal{H} \tag{2.53}$$

Following eq. (2.16), one can derive:

$$\begin{aligned} e^{-\beta\Delta A} = \frac{Q_1}{Q_0} &= \frac{\int\int \mathrm{d}\vec{r}^N \; \mathrm{d}\vec{p}^N \; e^{-\beta\mathcal{H}_1}}{\int\int \mathrm{d}\vec{r}^N \; \mathrm{d}\vec{p}^N \; e^{-\beta\mathcal{H}_0}} \\ &= \frac{\int \mathrm{d}\vec{r}^N \; e^{-\beta U_0} e^{-\beta\Delta U}}{\int \mathrm{d}\vec{r}^N \; e^{-\beta U_0}} \\ &= \int \mathrm{d}\vec{r}^N \; P_0(\vec{r}^N) e^{-\beta\Delta U} \\ &= \langle e^{-\beta\Delta U} \rangle_0 \end{aligned}$$

The last line is the central equation of FEP. It states that the change in free energy ΔA can be computed as the ensemble averaged exponential of the energy difference over configurations from system 0 alone. As computer simulations do not require analytically solvable reference, one can also obtain the free energy difference from simulations of system 1:

$$\beta\Delta A = -\ln \langle e^{-\beta\Delta U} \rangle_0 \tag{2.54}$$

$$\beta\Delta A = \ln \langle e^{\beta\Delta U} \rangle_1 \tag{2.55}$$

Similar expressions can be derived for any observable for which one wants to obtain an ensemble average in a system that is not explicitly simulated.

Interpretation and Limits

Both expressions in eq. (2.54) or eq. (2.55) are exact and should yield the same result. In practice, however, that is rarely the case. The two ensemble averages do not necessarily show the same convergence behaviour. It is informative to switch to the energy representation instead of the phase space formulation [12].

$$e^{-\beta\Delta A} = \int \mathrm{d}\Delta U \; P_0\left(\Delta U\right) \; e^{-\beta\Delta U} \tag{2.56}$$

This one-dimensional integral runs over the difference in potential energy between systems 0 and 1. The integrand contains an exponential of the energy difference and the distribution in ensemble 0 of the difference. For large systems, which contain many particles that undergo many random movements, $P_0(\Delta U)$ can, because of the central limit theorem [86], be assumed to form a Gaussian distribution. The result is depicted in Figure 2.1. The majority of the area under the curve of the integrand in eq. (2.56) lies in a region where P_0 is very small, i.e., the value of the integrand depends on a region of P_0 which is not well known. Thus, the integral is reliable only if the systems are sufficiently similar and P_0 is very narrow. Methods that extract information about the insufficiently sampled regions from the well known part of P_0 are available [87–90].

Often, equations (2.54) and (2.55) do not converge to the same values, which has to do with where the system's important regions are located in configuration space relative to one another. If those highly populated areas overlap, the forward and

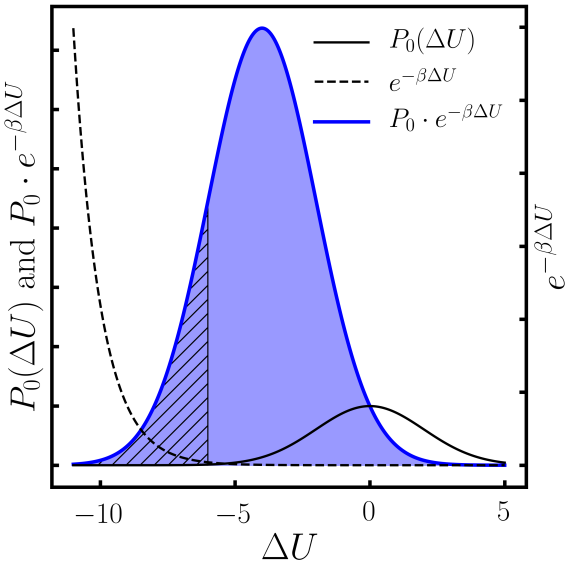


Figure 2.1: Depiction of the integrand in eq. (2.56) and its components. The perturbation ΔU is given in units of β^{-1}. The distribution $P_0(\Delta U)$ is a normalized Gaussian with mean zero and a variance of $4\beta^{-2}$. The blue integrand is the product of the two black curves. The shaded area lies outside of three times the confidence interval of P_0, and is thus obtained with very little accuracy.

backward perturbations are similar (Fig. 2.2a). If the two are disjoint, neither forward nor backward perturbations will produce a reliable result (Fig. 2.2b). In the case where one envelops the other, the transformation from the enveloping distribution to the enveloped one will be far more accurate than any result derived from the opposite transformation (Fig. 2.2c). However, for real systems it is usually unknown how the important regions are located with respect to one another.

Determining changes in entropy and enthalpy, which constitute the free energy change, are not as simple to derive as the free energy itself. The equations are not as numerically stable as the pure free energy [76]. Therefore, it is not as simple to split the full free energy into enthalpy and entropy as can be done with the DSI result.

2.3.2 Stratification or Staging

As mentioned above, free energy differences suffer from two problems that are both related to the Boltzmann distributions of states 0 and 1. If the distributions are sufficiently similar, the distribution of energy differences and the overlap of important regions are both well behaved. Using that the free energy is a state function, one can choose stages between the endpoints of the transformation (states 0 and 1) that mitigate these problems [91]. It is common to sample systems of a mixed Hamiltonian.

$$\mathcal{H}(\lambda) = (1 - \lambda)\mathcal{H}_0 + \lambda\mathcal{H}_1 \tag{2.57}$$

The Hamiltonian $\mathcal{H}(\lambda)$ interpolates between the end states and enables the creation of ensembles that connect the end points. A linear formulation is not necessary, but popular [61]. The intermediates do not have to be physically meaningful. The Hamiltonian

$\mathcal{H}(\lambda)$ fulfills the following properties:

$$\mathcal{H}(0) = \mathcal{H}_0$$
$$\mathcal{H}(1) = \mathcal{H}_1$$

The difference between two adjacent interpolated Hamiltonians is:

$$\Delta\mathcal{H}_i = \mathcal{H}(\lambda_{i+1}) - \mathcal{H}(\lambda_i) = \Delta\lambda_i \Delta\mathcal{H} \tag{2.58}$$

Note that $\Delta\mathcal{H}$ is the difference between the two end-point Hamiltonians. Using those intermediate stages, the total free energy change can be calculated. If there are n stages including the two end states, we get by using only the potential energy

$$\beta\Delta A = \sum_{i=i}^{n-1} \beta\Delta A_i = -\sum_{i=i}^{n-1} \ln \langle e^{-\beta\Delta U_i} \rangle_{\lambda_i} = -\sum_{i=i}^{n-1} \ln \langle e^{-\beta\Delta\lambda_i \Delta U} \rangle_{\lambda_i} \ . \tag{2.59}$$

There is no ideal way of choosing n and $\Delta\lambda_i$. For n, one has to balance numerical accuracy and computational effort, and the change in λ is ideally chosen in a way that there are more stages close to the end points than in the intermediate range, as the convergence to the end points is the more critical part.

2.3.3 Using Forward and Backward Perturbations

As pointed out in the Section 2.3.1 on free energy perturbation, there are two ways of calculating the free energy difference, a forward or a backward manner (eqs. (2.54) and (2.55)), which are in principle equivalent, but do not always yield the same result. However, it seems reasonable to extract information from both forward and backward perturbations.

On first glance, the easiest way to do so appears to be the mean of a forward and backward perturbation.

$$\beta\Delta A = \frac{1}{2}\left(\ln \langle e^{\beta\Delta U} \rangle_1 - \ln \langle e^{-\beta\Delta U} \rangle_0\right)$$

The above average is, however, often worse than one of the perturbations alone, since the distributions $P(\Delta U)$ are not the same for forward and backward perturbations, so that one result is typically more reliable than the other.

An improvement in the use of both forward and backward information is *simple overlap sampling* (SOS) [92]. Here, one uses the simulation data of two adjacent stratification stages to calculate the free energy difference via a not-simulated additional intermediate.

$$\beta\Delta A_i = -\ln\left(\frac{\langle e^{-\beta\Delta U_i/2} \rangle_0}{\langle e^{\beta\Delta U_i/2} \rangle_1}\right) \tag{2.60}$$

In this way, one can make the distribution $P(\Delta U)$ narrower without an extra simulation for the artificial midpoint. There are other formulas that make use of the forward and backward perturbations. The next section will deal with one of the better known methods.

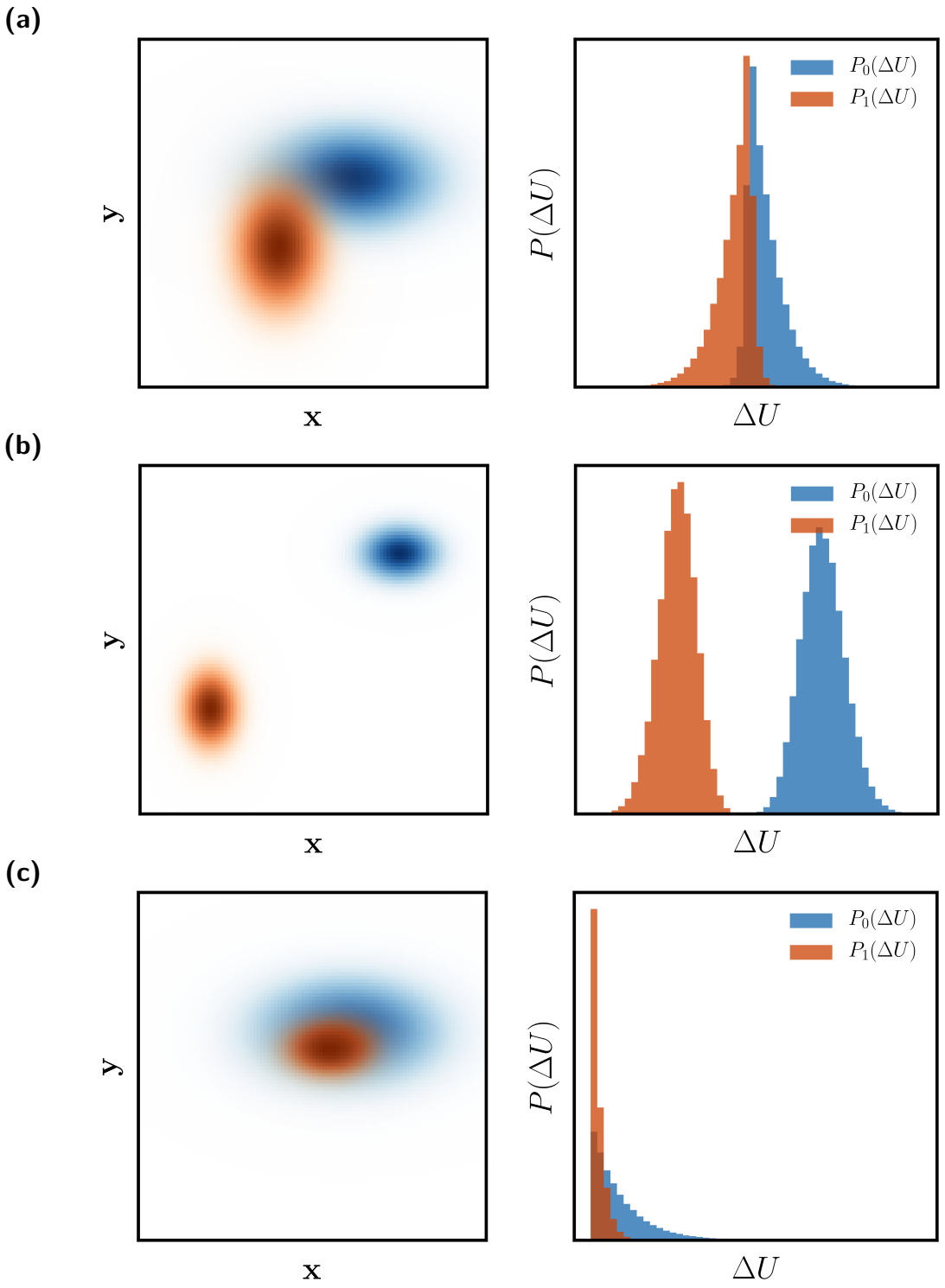


Figure 2.2: Panels in the left column depict the Boltzmann distributions of system 0 (blue) and system 1 (orange) in the x,y-plane. The right column shows the corresponding distributions of the potential energy difference $P(\Delta U)$. (a) The distributions overlap and FEP will give similar results for either forward or backward perturbation. (b) The important areas of the Boltzmann distributions are far removed from on another and the distributions $P(\Delta U)$ have no overlap. Either perturbation will predict the wrong result. (c) Boltzmann distribution 0 envelops 1, thus the forward perturbation will be very accurate whereas the backward perturbation misses important regions of system 0.

2.3.4 Bennett's Acceptance Ratio

Bennett's Acceptance Ratio (BAR), derived in 1976 [15], uses the information on both forward and backward perturbations. The formula aims at minimizing the variance in the free energy estimate. In the following derivation, Bennett's reasoning is used, but with more explicit intermediate steps, as the original paper itself is very brief.

The free energy difference defined in eq. (2.16) is a function of the ratio of partition functions ($\Delta A \propto \ln \frac{Q_0}{Q_1}$), which in eq. (2.17) is reduced to an integral over configuration space (using only the potential energy function). The fraction can be expanded with an integral over Boltzmann factors to turn it into a ratio of ensemble averages instead

$$\frac{Q_0}{Q_1} = \frac{Q_0 \int \mathrm{d}\vec{r}^N \; W(\vec{r}^N) e^{-\beta U_0(\vec{r}^N) - \beta U_1(\vec{r}^N)}}{Q_1 \int \mathrm{d}\vec{r}^N \; W(\vec{r}^N) e^{-\beta U_0(\vec{r}^N) - \beta U_1(\vec{r}^N)}} = \frac{\langle W e^{-\beta U_0} \rangle_1}{\langle W e^{-\beta U_1} \rangle_0} \; , \tag{2.61}$$

where W can be any finite function that has no roots. This is necessary to ensure that the integrals used in numerator and denominator do not vanish. Without loss of generality we can assume $W(\vec{r}^N) > 0$.

For brevity's sake, the argument of the potential energy functions U_0 and U_1, as well as that of the function W, will be omitted in the following equations. In numerical evaluations of the ensemble averages as sums over trajectory frames, the index naming one configuration i will be added as an argument to the function, e.g., $U_0(\vec{r}_i^N) = U_0(i)$.

Assuming that simulations of the states 0 and 1 have been carried out with trajectories of length n_0 and n_1, respectively, the exact estimate (integral over configuration space) of the free energy difference between states 0 and 1 using eq. (2.61) is

$$\beta \Delta A_{\text{exact}} = \ln \frac{\langle W e^{-\beta U_0} \rangle_1}{\langle W e^{-\beta U_1} \rangle_0} \; , \tag{2.62}$$

whereas the numerical value of this difference based on the trajectories of length n_0 and n_1 is obtained with

$$\beta \Delta A_{\text{num}} = \ln \left(\frac{\frac{1}{n_1} \sum_j^{n_1} W(j) e^{-\beta U_0(j)}}{\frac{1}{n_0} \sum_i^{n_0} W(i) e^{-\beta U_1(i)}} \right) \; . \tag{2.63}$$

The dimensionless expected squared deviation between ΔA_{exact} and ΔA_{num} is:

$$\begin{aligned}
\langle (\beta \Delta \Delta A)^2 \rangle = \sigma^2 &= \langle \beta^2 \left(\Delta A_{\text{num}} - \Delta A_{\text{exact}} \right)^2 \rangle \\
\sigma^2 &= \left\langle \left[\ln \left(\frac{\frac{1}{n_1} \sum_j^{n_1} W(j) e^{-\beta U_0(j)}}{\frac{1}{n_0} \sum_i^{n_0} W(i) e^{-\beta U_1(i)}} \right) - \ln \left(\frac{\langle W e^{-\beta U_0} \rangle_1}{\langle W e^{-\beta U_1} \rangle_0} \right) \right]^2 \right\rangle \\
&= \left\langle \left[\ln \left(\frac{\frac{1}{n_1} \sum_j^{n_1} W(j) e^{-\beta U_0(j)}}{\langle W e^{-\beta U_0} \rangle_1} \right) - \ln \left(\frac{\frac{1}{n_0} \sum_i^{n_0} W(i) e^{-\beta U_1(i)}}{\langle W e^{-\beta U_1} \rangle_0} \right) \right]^2 \right\rangle
\end{aligned} \tag{2.64}$$

If the sample size is large, the numerical averages are close to the exact ensemble averages, i.e., the argument of the logarithm is close to 1, thus $\ln x \approx x - 1$.

$$\begin{aligned}
\sigma^2 =& \left\langle \left[\frac{\frac{1}{n_1}\sum_j^{n_1} W(j) e^{-\beta U_0(j)}}{\langle W e^{-\beta U_0}\rangle_1} - 1 - \frac{\frac{1}{n_0}\sum_i^{n_0} W(j) e^{-\beta U_1(i)}}{\langle W e^{-\beta U_1}\rangle_0} + 1 \right]^2 \right\rangle \\
=& \left\langle \frac{\frac{1}{n_1^2}\left(\sum_j^{n_1} W(j)\ e^{-\beta U_0(j)}\right)^2}{\langle W e^{-\beta U_0}\rangle_1^2} - 2\frac{\frac{1}{n_0 n_1}\left(\sum_i^{n_0} W(i)\ e^{-\beta U_1(i)}\right)\left(\sum_j^{n_1} W(j)\ e^{-\beta U_0(j)}\right)}{\langle W e^{-\beta U_1}\rangle_0 \langle W e^{-\beta U_0}\rangle_1} \right. \\
&\left. + \frac{\frac{1}{n_0^2}\left(\sum_i^{n_0} W(i)\ e^{-\beta U_1(i)}\right)^2}{\langle W e^{-\beta U_1}\rangle_0^2} \right\rangle \\
=& \left\langle \frac{\frac{1}{n_1^2}\left(\sum_j^{n_1} W^2(j)\ e^{-2\beta U_0(j)} + \sum_j^{n_1}\sum_{j',j'\neq j}^{n_1-1} W(j)\ W(j')\ e^{-\beta U_0(j)} e^{-\beta U_0(j')}\right)}{\langle W e^{-\beta U_0}\rangle_1^2} \right. \\
&- 2\frac{\frac{1}{n_0 n_1}\left(\sum_i^{n_0} W(i)\ e^{-\beta U_1(i)}\right)\left(\sum_j^{n_1} W(j)\ e^{-\beta U_0(j)}\right)}{\langle W e^{-\beta U_1}\rangle_0 \langle W e^{-\beta U_0}\rangle_1} \\
&\left. + \frac{\frac{1}{n_0^2}\left(\sum_i^{n_0} W^2(i)\ e^{-2\beta U_1(i)} + \sum_i^{n_0}\sum_{i',i'\neq i}^{n_0-1} W(i)\ W(i')\ e^{-\beta U_1(i)} e^{-\beta U_1(i')}\right)}{\langle W e^{-\beta U_1}\rangle_0^2} \right\rangle
\end{aligned}$$

Now the expected value operator ($\langle\ \rangle$) is applied to every term individually.

$$\begin{aligned}
\sigma^2 =& \frac{\frac{1}{n_1}\left\langle W^2 e^{-2\beta U_0}\right\rangle_1 + \frac{n_1(n_1-1)}{n_1^2}\left\langle W e^{-\beta U_0}\right\rangle_1^2}{\langle W e^{-\beta U_0}\rangle_1^2} - 2\frac{\langle W e^{-\beta U_1}\rangle_0 \langle W e^{-\beta U_0}\rangle_1}{\langle W e^{-\beta U_1}\rangle_0 \langle W e^{-\beta U_0}\rangle_1} \\
&+ \frac{\frac{1}{n_0}\left\langle W^2 e^{-2\beta U_1}\right\rangle_0 + \frac{n_0(n_0-1)}{n_0^2}\left\langle W e^{-\beta U_1}\right\rangle_0^2}{\langle W e^{-\beta U_1}\rangle_0^2} \\
\sigma^2 =& \frac{\left\langle W^2 e^{-2\beta U_0}\right\rangle_1}{n_1 \langle W e^{-\beta U_0}\rangle_1^2} + \frac{\left\langle W^2 e^{-2\beta U_1}\right\rangle_0}{n_0 \langle W e^{-\beta U_1}\rangle_0^2} - \frac{1}{n_0} - \frac{1}{n_1}
\end{aligned} \tag{2.65}$$

This expression can be used to minimize the squared deviation with respect to the choice of function W. The ensemble averages from eq. (2.65) are written out explicitly as functional $\sigma^2[W]$.

$$\begin{aligned}
\sigma^2[W] =& \frac{\int \mathrm{d}\vec{r}^N\ W^2 e^{-2\beta U_0} e^{-\beta U_1} Q_1^{-1}}{n_1 \left(\int \mathrm{d}\vec{r}^N\ W e^{-\beta U_0} e^{-\beta U_1} Q_1^{-1}\right)^2} + \frac{\int \mathrm{d}\vec{r}^N\ W^2 e^{-2\beta U_1} e^{-\beta U_0} Q_0^{-1}}{n_0 \left(\int \mathrm{d}\vec{r}^N\ W e^{-\beta U_1} e^{-\beta U_0} Q_0^{-1}\right)^2} - \frac{1}{n_0} - \frac{1}{n_1} \\
=& \frac{\int \mathrm{d}\vec{r}^N\ \left(\frac{Q_0}{n_0} e^{-\beta U_1} + \frac{Q_1}{n_1} e^{-\beta U_0}\right) W^2 e^{-\beta U_0} e^{-\beta U_1}}{\left(\int \mathrm{d}\vec{r}^N\ W e^{-\beta U_1} e^{-\beta U_0}\right)^2} - \frac{1}{n_0} - \frac{1}{n_1}
\end{aligned} \tag{2.66}$$

To obtain the minimal squared deviation, one needs the functional derivative of $\sigma^2[W]$ with respect to W. Using a test function ϕ, the stationary point can be determined

according to:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\epsilon}\sigma^2[W+\epsilon\phi]\bigg|_{\epsilon=0} &= \frac{\mathrm{d}}{\mathrm{d}\epsilon}\frac{\int \mathrm{d}\vec{r}^N \left(\frac{Q_0}{n_0}e^{-\beta U_1}+\frac{Q_1}{n_1}e^{-\beta U_0}\right)(W+\epsilon\phi)^2 e^{-\beta U_0}e^{-\beta U_1}}{\left(\int \mathrm{d}\vec{r}^N\ (W+\epsilon\phi)e^{-\beta U_1}e^{-\beta U_0}\right)^2} - \frac{1}{n_0} - \frac{1}{n_1}\bigg|_{\epsilon=0} \\
&= \frac{\int \mathrm{d}\vec{r}^N \left(\frac{Q_0}{n_0}e^{-\beta U_1}+\frac{Q_1}{n_1}e^{-\beta U_0}\right) 2We^{-\beta U_0}e^{-\beta U_1}\phi}{\left(\int \mathrm{d}\vec{r}^N\ We^{-\beta U_1}e^{-\beta U_0}\right)^4}\left(\int \mathrm{d}\vec{r}^N\ We^{-\beta U_1}e^{-\beta U_0}\right)^2 \\
&\quad - \frac{\int \mathrm{d}\vec{r}^N \left(\frac{Q_0}{n_0}e^{-\beta U_1}+\frac{Q_1}{n_1}e^{-\beta U_0}\right) W^2 e^{-\beta U_0}e^{-\beta U_1}}{\left(\int \mathrm{d}\vec{r}^N\ We^{-\beta U_1}e^{-\beta U_0}\right)^4} \\
&\quad \times 2\int \mathrm{d}\vec{r}^N\ We^{-\beta U_1}e^{-\beta U_0}\int \mathrm{d}\vec{r}^N\ \phi e^{-\beta U_1}e^{-\beta U_0} \quad \stackrel{!}{=} \mathbf{0}
\end{aligned} \tag{2.67}$$

$$\begin{aligned}
&\int \mathrm{d}\vec{r}^N \left(\frac{Q_0}{n_0}e^{-\beta U_1}+\frac{Q_1}{n_1}e^{-\beta U_0}\right) We^{-\beta U_0}e^{-\beta U_1}\phi \ \times\ \int \mathrm{d}\vec{r}^N\ We^{-\beta U_1}e^{-\beta U_0} = \\
&\qquad \int \mathrm{d}\vec{r}^N \left(\frac{Q_0}{n_0}e^{-\beta U_1}+\frac{Q_1}{n_1}e^{-\beta U_0}\right) W^2 e^{-\beta U_0}e^{-\beta U_1} \ \times\ \int \mathrm{d}\vec{r}^N\ \phi e^{-\beta U_1}e^{-\beta U_0}
\end{aligned} \tag{2.68}$$

At the stationary point of the functional $\sigma^2[W]$, eq. (2.68) has to be fulfilled for any arbitrary function ϕ. The products on both sides of the equation contain one term with and one term without ϕ. As ϕ is arbitrary, the terms containing ϕ and the terms without ϕ have to be pairwise equal. Hence,

$$W \propto \frac{1}{\frac{Q_0}{n_0}e^{-\beta U_1}+\frac{Q_1}{n_1}e^{-\beta U_0}}\ . \tag{2.69}$$

This choice of W minimizes the squared error of an estimated free energy difference. To get the working equation, which is referred to as BAR, one needs to insert the optimal W into eq. (2.61).

$$\begin{aligned}
\frac{Q_0}{Q_1} &= \frac{\langle We^{-\beta U_0}\rangle_1}{\langle We^{-\beta U_1}\rangle_0} = \frac{\left\langle \frac{e^{-\beta U_0}}{\frac{Q_0}{n_0}e^{-\beta U_1}+\frac{Q_1}{n_1}e^{-\beta U_0}}\right\rangle_1}{\left\langle \frac{e^{-\beta U_1}}{\frac{Q_0}{n_0}e^{-\beta U_1}+\frac{Q_1}{n_1}e^{-\beta U_0}}\right\rangle_0} \\
&= \frac{\left\langle \frac{1}{\frac{Q_0}{n_0}e^{-\beta\Delta U}+\frac{Q_1}{n_1}}\right\rangle_1}{\left\langle \frac{1}{\frac{Q_0}{n_0}+\frac{Q_1}{n_1}e^{\beta\Delta U}}\right\rangle_0} = \frac{\left\langle \frac{\frac{n_1}{Q_1}}{\frac{Q_0 n_1}{Q_1 n_0}e^{-\beta\Delta U}+1}\right\rangle_1}{\left\langle \frac{\frac{n_0}{Q_0}}{\frac{Q_1 n_0}{Q_0 n_1}e^{\beta\Delta U}+1}\right\rangle_0} \\
&= \frac{\left\langle \frac{1}{e^{\ln\frac{Q_0 n_1}{Q_1 n_0}-\beta\Delta U}+1}\right\rangle_1}{\left\langle \frac{1}{e^{-\ln\frac{Q_0 n_1}{Q_1 n_0}+\beta\Delta U}+1}\right\rangle_0}\frac{Q_0 n_1}{Q_1 n_0}
\end{aligned} \tag{2.70}$$

The function inside the ensemble average brackets is the Fermi function $f(x) = (e^x+1)^{-1}$. The logarithm of the above expression yields the free energy difference:

$$0 = \ln \frac{\left\langle f\left(-\beta\Delta U - \ln\frac{n_0}{n_1} + \beta\Delta A\right)\right\rangle_1}{\left\langle f\left(\beta\Delta U + \ln\frac{n_0}{n_1} - \beta\Delta A\right)\right\rangle_0} + \ln\frac{n_1}{n_0} \tag{2.71}$$

$$0 = \ln \frac{\sum_j^{n_1} f\left(-\beta\Delta U(j) - \ln\frac{n_0}{n_1} + \beta\Delta A\right)}{\sum_i^{n_0} f\left(\beta\Delta U(i) + \ln\frac{n_0}{n_1} - \beta\Delta A\right)} \tag{2.72}$$

Eq. (2.71) is the analytical result and eq. (2.72) the numerical equivalent, which has to be used for evaluating trajectory data. Note that the definition of the free energy difference is implicit. Thus, BAR has to be solved iteratively.

Eq. (2.71) has been derived by Shirts et al. [93] by a maximum-likelihood formulation of the free energy problem. Therefore, given a certain data set, the BAR method yields the most likely result and has the smallest possible squared error. It is practically and numerically superior to free energy perturbation [85, 94, 95] as well as thermodynamic integration [82]. In **Publication I**, we used the BAR estimator to determine the accuracy of the vibrational hot-spots method. Even though the hot-spots are not designed to yield precise free energy differences, it is important that the results show the correct behaviour in order to use it as an interpretation tool.

2.3.5 Free Energy along a Reaction Coordinate and Umbrella Sampling

Free Energy as function of ξ

Computing the free energy as a function of a variable ξ is conceptually a relatively simple task. In practice, the Boltzmann distribution $P(\xi)$ is approximated as a histogram along the reaction coordinate.

$$P(\xi) = \frac{\sum_i^n \delta\left(\xi(\vec{r}_i^N) - \xi\right)}{n} \tag{2.73}$$

Numerically, the delta function is evaluated with a certain histogram bin width around ξ where it is equal to 1, and n is again the number of simulation frames. The free energy is then simply:

$$A(\xi) = -\beta^{-1} \ln P(\xi) \tag{2.74}$$

Umbrella Sampling

In most applications, ξ describes a process connecting several regions in configuration space. Those are often minima on the potential energy surface. However, minima are separated by free energy barriers, such that exchange between them is rare. Therefore, trajectories spend very little time in the transition area or do not venture into that region at all during a simulation of normal duration. Hence, a histogram generated from one trajectory contains very few or no data for the transition region or both minima, which in turn results in bad estimates of the free energy in those regions.

Techniques to overcome the problem of insufficient sampling of high energy regions are dubbed *importance sampling* [11, 96]. One class of importance sampling methods aims to modify the potential energy surface in such a way that all regions of ξ are sampled with the same frequency. Well known methods are, for example, exponential flooding [97] and metadynamics [98]. They introduce a time-dependent potential that "learns" from the previous simulation and aims to flatten the free energy surface along ξ so that the dynamics become diffusive. A time-independent approach is utilized by the Umbrella Sampling method [12]. Here, one adds a biasing potential to increase the duration of stay in a certain region of configuration space. Ideally, this potential has the same effect as the time-dependent potentials and flattens the potential energy surface significantly, allowing for more diffusive dynamics. A good choice of biasing potential requires knowledge of the underlying energy surface. It has become very popular to choose harmonic biasing potentials [99–101]

$$B_i(\xi) = \frac{k_i}{2} \left(\xi - \xi_i\right)^2 \ , \tag{2.75}$$

where the index i indicates a given biasing potential *window* along the reaction path and k_i is the corresponding force constant. These windows have the same effect as staging and allow for a parallelization of the computational effort.

Removing Biasing Potentials

The Boltzmann distribution of the reaction coordinate in the biased ensemble $P_B(\xi)$ is given as

$$P_B(\xi) = \frac{\int \mathrm{d}\vec{r}^N \ \delta\left(\xi(\vec{r}^N) - \xi\right) e^{-\beta(U_0+B)}}{\int \mathrm{d}\vec{r}^N \ e^{-\beta(U_0+B)}} = \frac{1}{Q_B} \int \mathrm{d}\vec{r}^N \ \delta\left(\xi(\vec{r}^N) - \xi\right) e^{-\beta(U_0+B)} \tag{2.76}$$

U_0 denotes here the original potential energy function.

The unbiased distribution can recovered by [91, 102]

$$\begin{aligned} P_0(\xi) =& \frac{1}{Q_0} \int \mathrm{d}\vec{r}^N \ \delta\left(\xi(\vec{r}^N) - \xi\right) e^{-\beta U_0} \\ =& \frac{1}{Q_0} \int \mathrm{d}\vec{r}^N \ \delta\left(\xi(\vec{r}^N) - \xi\right) e^{-\beta U_0} e^{-\beta B} e^{+\beta B} \\ =& \frac{e^{+\beta B(\xi)}}{Q_0} \int \mathrm{d}\vec{r}^N \ \delta\left(\xi(\vec{r}^N) - \xi\right) e^{-\beta U_0} e^{-\beta B} \\ =& P_B(\xi) \frac{Q_B}{Q_0} e^{+\beta B(\xi)} \end{aligned} \tag{2.77}$$

and inserting this into eq. (2.74) yields

$$A_0(\xi) = A_B(\xi) - B(\xi) - \beta^{-1} \ln \frac{Q_B}{Q_0} \ . \tag{2.78}$$

The last term is the free energy difference between the biased and unbiased system. It corrects the absolute free energy values for the influence of the biasing potential. If

only one biasing potential has been used, this correction can be neglected, as only the relative values of $A_0(\xi)$ along ξ are of interest. However, when several different biasing potentials have been used, it is important to determine these corrections to obtain one continuous free energy surface, as each simulation contributes only fragments.

In principle, this free energy difference can be calculated via FEP or BAR. It is desirable, however, to use the information from all simulations which overlap along ξ and to compute the alignment free energy differences $-\beta^{-1} \ln \frac{Q_B}{Q_0}$ simultaneously to have similar precision. The following section derives such algorithms.

2.3.6 WHAM and MBAR

Nowadays, there are a number of methods that can extract the original free energy surface $A_0(\xi)$ from different biased simulations, which have potentially been carried out at different temperatures. WHAM [103] and MBAR [30] need equilibrium data, which are ideally not correlated. Umbrella integration combines umbrella sampling with thermodynamic integration by assuming that the values of the reaction coordinate within one umbrella window are in fact Gaussian distributed [100, 104]. DHAM [105] and DHEMed [106] are based on a kinetic approach within the Markov State Model framework, and TRAM [107], as it estimates multiensemble Markov models, is the generalization to all of them.

As this thesis does not use Markov models, the focus in this section is on WHAM and MBAR.

WHAM

The first algorithm to combine the data from several umbrella window simulations was the *Weighted Histogram Analysis Method* (WHAM). It is able to combine data not only from simulations with different biasing potentials, but also from simulations at different temperatures.

The original WHAM formulation assumes that the overall potential U in one simulation is a linear combination of the original potential energy function and biasing potentials. Here we will denote the potential energy function of simulation i simply by U_i. The inverse thermal energy of run i is β_i. All the data can be reweighted for any new β or new potential U. The total number of simulations is denoted by S. The WHAM equations [103] are therefore:

$$P_{U,\beta}\left(\mathbf{U},\xi\right) = \frac{\sum_{i}^{S} n_i\left(\mathbf{U},\xi\right)\ e^{-\beta U}}{\sum_{i}^{S} n_i\ e^{\beta_i A_i - \beta_i U_i}} \tag{2.79}$$

$$\beta_i A_i = -\ln \sum_{\mathbf{U},\xi} P_{U_i,\beta_i}\left(\mathbf{U},\xi\right) \tag{2.80}$$

In eq. (2.79), β and U designate new values. A_i denotes the free energy of simulation i and n_i its total number of frames. $n_i\left(\mathbf{U},\xi\right)$ is the bin value for simulation i in a

multidimensional histogram over values of both the coordinate ξ and potential energy functions U_i. The free energies, A_i, of the different simulations are obtained via eq. (2.80) and their difference corresponds to the third term on the right hand side in eq. (2.78). Their absolute value is meaningless, but their relative magnitude is important. As the free energies appear in both equations, WHAM has to be solved iteratively, e.g., in a self-consistent manner.

While WHAM presented the first set of equations that enable a simultaneous evaluation of different umbrella window simulations, the multidimensional histogram over potential energies and the reaction coordinates make it numerically complicated and unstable. The histogram over potential energies is unnecessary if one is interested only in the reaction coordinate ξ [108]. A formulation is warranted that focuses on determining the A_i before any reweighting takes place. That can easily be accomplished by inserting eq. (2.79) into eq. (2.80). From here on, the distinction of different β_i made above will be dropped as all simulations in this thesis have been carried out at room temperature. Hence,

$$e^{-\beta A_i} = \sum_j^S \sum_k^{n_j} \frac{e^{-\beta U_i(j,k)}}{\sum_l^S n_l \; e^{\beta A_l - \beta U_l(j,k)}} \;, \tag{2.81}$$

where $U_i(j,k)$ is the value of the potential energy function i in the k^{th} frame of the j^{th} simulation. This equation already appeared in the original WHAM paper and was re-derived by Shirts and Chodera [30], who called it MBAR (Multistate Bennett Acceptance Ratio). It is also referred to as binless WHAM to highlight the fact that no multidimensional histogram $P(\mathbf{U}, \xi)$ is needed.

MBAR

A short derivation of the connection between binless WHAM and Benett's acceptance ratio follows. The first step is to recast the function W from eq. (2.69).

$$\begin{aligned} W \propto & \frac{1}{\frac{Q_0}{n_0} e^{-\beta U_1} + \frac{Q_1}{n_1} e^{-\beta U_0}} = \frac{\frac{n_1 n_0}{Q_0 Q_1}}{\frac{n_0}{Q_0} e^{-\beta U_0} + \frac{n_1}{Q_1} e^{-\beta U_1}} \\ \propto & \frac{1}{n_0 \; e^{\beta A_0 - \beta U_0} + n_1 \; e^{\beta A_1 - \beta U_1}} \end{aligned} \tag{2.82}$$

If this is inserted again into eq. (2.61) one can see that

$$\frac{e^{-\beta A_0}}{e^{-\beta A_1}} = \frac{\frac{1}{n_1} \sum_j^{n_1} \frac{e^{-\beta U_0(j)}}{n_0 \; e^{\beta A_0 - \beta U_0(j)} + n_1 \; e^{\beta A_1 - \beta U_1(j)}}}{\frac{1}{n_0} \sum_i^{n_0} \frac{e^{-\beta U_1(i)}}{n_0 \; e^{\beta A_0 - \beta U_0(i)} + n_1 \; e^{\beta A_1 - \beta U_1(i)}}}$$

$$\sum_i^{n_0} \frac{\frac{n_1 \; e^{-\beta A_0}}{n_0 \; e^{-\beta A_1}} \; e^{-\beta U_1(i)}}{\sum_{m=0}^1 n_m \; e^{\beta A_m - \beta U_m(i)}} = \sum_j^{n_1} \frac{e^{-\beta U_0(j)}}{\sum_{m=0}^1 n_m \; e^{\beta A_m - \beta U_m(j)}}$$

$$\sum_i^{n_0} \frac{\frac{n_1 \; e^{-\beta A_0}}{n_0 \; e^{-\beta A_1}} \; e^{-\beta U_1(i)}}{\sum_{m=0}^1 n_m \; e^{\beta A_m - \beta U_m(i)}} + \sum_i^{n_0} \frac{e^{-\beta U_0(i)}}{\sum_{m=0}^1 n_m \; e^{\beta A_m - \beta U_m(i)}} = \sum_{k=0}^1 \sum_l^{n_k} \frac{e^{-\beta U_0(k,l)}}{\sum_{m=0}^1 n_m \; e^{\beta A_m - \beta U_m(k,l)}}$$

$$\frac{e^{-\beta A_0}}{n_0}\sum_i^{n_0}\frac{n_0\ e^{\beta A_0-\beta U_0(i)}+n_1\ e^{\beta A_1-\beta U_1(i)}}{\sum_{m=0}^{1} n_m\ e^{\beta A_m-\beta U_m(i)}}=\sum_{k=0}^{1}\sum_l^{n_k}\frac{e^{-\beta U_0(k,l)}}{\sum_{m=0}^{1} n_m\ e^{\beta A_m-\beta U_m(k,l)}}$$

$$e^{-\beta A_0}=\sum_{k=0}^{1}\sum_l^{n_k}\frac{e^{-\beta U_0(k,l)}}{\sum_{m=0}^{1} n_m\ e^{\beta A_m-\beta U_m(k,l)}}$$

which is exactly the MBAR equation for two states. For more than two potentials, W takes the generalized form [109]

$$W_{i,j}=\frac{n_i\ e^{\beta A_i}}{\sum_l^S n_l\ e^{\beta A_l-\beta U_l}}\ , \tag{2.83}$$

when it is applied to the expanded ratio

$$\frac{Q_i}{Q_j}=\frac{\langle W_{i,j}e^{-\beta U_i}\rangle_j}{\langle W_{i,j}e^{-\beta U_j}\rangle_i}\ .$$

Just like the WHAM equations, the MBAR equations have to be solved iteratively. The most direct way is to do so self-consistently, which can, however, show slow convergence close to the optimal set of A_i's [30]. Therefore, the implicit MBAR equations are often recast as minimization problem.

Optimization of Free Energy Constants A_i

To speed up convergence, optimization algorithms such as the Newton-Raphson [110] or Broyden-Fletcher–Goldfarb-Shanno (BFGS) [111–114] methods can be used. In order to make use of these algorithms, the equations above have to be reformulated in such a manner that the desired result coincides with the root of a function. Eq. (2.81) can easily be reformulated to:

$$g_i=n_i-\sum_j^S\sum_k^{n_j}\frac{n_i\ e^{\beta A_i-\beta U_i(j,k)}}{\sum_l^S n_l\ e^{\beta A_l-\beta U_l(j,k)}}=0 \tag{2.84}$$

The multiplication by n_i ensures symmetric derivatives. In this way, S functions g_i are obtained of which one aims to find the root simultaneously. In other words, one seeks to reduce the vector $\mathbf{g}$ to $\mathbf{0}$, where $(\mathbf{g})_i=g_i$. Shirts et al. [30] proposed to use Newton-Raphson, in order to avoid a third order tensor Hessian needed for BFGS.

The idea of finding the optimal A_i's by minimizing an objective function (e.g., the g_i above) instead of a self-consistent procedure has also been applied to the binned WHAM estimator by Zhu and Hummer [115], who then used BFGS. In another publication, solving the set of equations by direct inversion of the iterative subspace (DIIS) [116] has been proposed [117].

2.3.7 Reweighting

Sometimes it is not possible to generate enough configurations needed for the application of free energy algorithms at the desired level of theory (the sampling of configuration

space has to be thorough as the algorithms have been derived under the assumption of equilibrium conditions). This is especially true when using quantum mechanical methods in the context of AIMD or QM/MM-MD. Neighbouring configurations along a trajectory are usually highly correlated and add little new information to those averages. After data generation the same average values can be obtained with fewer data, but unfortunately the many simulation steps in between data points are necessary for the MD or MC algorithms to work. The number of potential energy function and gradient evaluations might be too large to use a high level of QM theory, even though one desires to know the free energy surface, $A(\xi)$, on that level.

However, it is in general possible to sample a system at a lower level of theory and estimate through a second, less demanding step the high level result. We have applied this idea to the reaction mechanism of Sirt5 in **Manuscript IV**.

First, one needs to solve the unbinned WHAM or MBAR equations for the level of theory which the umbrella simulations were performed at, in order to obtain the relative free energies A_i. The free energy surface, $A(\xi)$, can be computed for the low-level potential energy function U_0 using

$$\beta A_0(\xi) = -\ln \sum_j^S \sum_k^{n_j} \frac{\delta\left(\xi(j,k)-\xi\right)\ e^{-\beta U_0(j,k)}}{\sum_l^S n_l\ e^{\beta A_l - \beta U_l(j,k)}} = -\ln \sum_j^S \sum_k^{n_j} \frac{\delta\left(\xi(j,k)-\xi\right)}{\sum_l^S n_l\ e^{\beta A_l - \beta B_l(j,k)}}\ . \tag{2.85}$$

The expression on the right hand side can be obtained by writing the potential energy function of each umbrella window as $U_l = U_0 + B_l$. The term U_0 will therefore cancel in numerator and denominator. In general, one can use any potential energy function in the numerator, which results in the corresponding "reweighted" free energy surface. Therefore, the reweighting equation reads for an arbitrary (desirably high-level) potential energy function U_1 [118]:

$$\beta A_1(\xi) = -\ln \sum_j^S \sum_k^{n_j} \frac{\delta\left(\xi(j,k)-\xi\right)\ e^{-\beta U_1(j,k)}}{\sum_l^S n_l\ e^{\beta A_l - \beta U_l(j,k)}} = -\ln \sum_j^S \sum_k^{n_j} \frac{\delta\left(\xi(j,k)-\xi\right)\ \ e^{-\beta \Delta U(j,k)}}{\sum_l^S n_l\ e^{\beta A_l - \beta B_l(j,k)}}\ , \tag{2.86}$$

where again $\Delta U = U_1 - U_0$. In practice, the Boltzmann distributions associated with U_0 and U_1 have to have sufficient overlap as depicted in Fig. 2.2a, since otherwise the exponential average becomes ill-conditioned and shows poor convergence behaviour (see discussion about the limitations of FEP).

2.4 Machine Learning

2.4.1 Introduction

The term *Machine Learning* refers to a growing field of statistical data analysis that has become ubiquitous, e.g., in the form of handwriting or speech recognition, shopping recommendations, risk management, or even clever computers that have mastered games [119]. It covers a wide range of applications and fields; chemistry is no exception. With the availability of freely accessible experimental databases, predictive models have been built that can, for example, aid retro-synthesis [120] or predict new crystal structures [121]. It has also been used to enhance computational chemistry, where machine learning is used to infer solutions of the Schrödinger equation from previous calculations. In this way new functionals, basis set effects, observables, potential energy surfaces, and more have been learned (see for example Ref. [122, 123] and references therein).

Some believe that machine learning can, with enough data and computational power, become more powerful than human researchers [123]: "With machine learning, given enough data and a rule-discovery algorithm, a computer has the ability to determine all known physical laws (and potentially those that are currently unknown) without human input."

2.4.2 The Different Branches of Machine Learning

In general, the field of machine learning can be divided into three main categories, namely supervised learning, unsupervised learning, and reinforcement learning. Reinforcement learning can be described as explorative learning, where the algorithm takes new steps (action) based on what it has seen and then needs a feedback, called reward, to make its next decision.

Unsupervised learning uses unlabelled data (data with no specific output value) and seeks to find patterns, which can be used to sort (clustering) or reduce the data to their important components (signal preparation). The widely-known principal component analysis [124] is such a dimensionality reduction technique. Clustering techniques such as k-means [58] or DBSCAN [59], can be used to group similar data together. It can, e.g., be used to identify groups of configurations representing wells in the potential energy surface, which can be identified as microstates (see Section 2.1.4).

Supervised learning is an important category of machine learning for the natural sciences. Here, the algorithm is supplied with labelled data and generates a model that connects the features (input) with the label (output). However, there are a number of different models and algorithms to choose from, regression analysis, decision trees, and artificial neural networks, to name a few.

2.4.3 Selecting an Algorithm

Several key questions have to be addressed when building a supervised learning model. One has to choose a suitable kind of data with respect to the research question, generate or acquire enough data, choose a suitable feature that represents the data, select a model, train it, and finally evaluate it. The first two aspects are rather obvious.

Feature and model selection are key to any successful modelling. Good features are very important, ideally the representation has to be unique, e.g., for single molecules a Coulomb matrix [125, 126] is often used, as it is insensitive to rotation and translation. Other feature descriptions for chemistry are Bag of Bonds [127], XYZ-Coordinates, or SMILES [128]. The fewer features used for input description the better.

When selecting an algorithm, one has to pay attention to the flexibility of the model. An inflexible model is associated with a high bias, meaning it cannot adjust enough to the true function, whereas, a very flexible model might generate a high variance, as it creates a more complex function than the true underlying function. A suitable trade-off between bias and variance needs to be found. As a consequence, with increasing model complexity, more data is needed to fit the function well.

In mathematical terms, one seeks to find a function f of the feature vector $\mathbf{x}$ that correctly predicts the label y. The deviation between the correct value of y and $f(\mathbf{x})$ is measured by a loss function $L(y, f(\mathbf{x}))$. A popular loss function is the squared error $L(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2$. Then to optimize f, the expected error of a function, which is called its risk R, has to be estimated and reduced. The true risk, which is the weighted average of the loss function over all feature-label pairs is unknown. Thus, the empirical risk is calculated

$$R_n(f) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i)) \tag{2.87}$$

as an average over n samples. There is one important caveat: If f is complex enough it can always reduce the empirical risk by overfitting the data. Therefore, the complexity of f has to be kept as low as possible, which is related to the famous theorem of Occam's razor [129].

A common approach is to divide the training data into a learning and a test set, which enables one to check for overfitting. One optimizes f on the learning set and then evaluates its performance on the test set. This is especially important if f contains additional parameters that are not optimized in a normal learning procedure (minimization of the risk). Such parameters are called hyperparameters and are often used to keep the model complexity low. In an extra step an optimal setting of the hyperparameters has to be found. To avoid an unfortunate split of the data in training and test set, by, for example, putting crucial data exclusively into the test set, one can use bootstrapping or cross-validation.

2.4.4 The Algorithm Used in Publication III

In **Publication III**, we learned the relationship between active site configuration of the protein Sirtuin 5 and the height of the activation barrier for the reaction this enzyme catalyzes. 150 minimum energy paths from educt to product were computed, where the configuration of the local educt minimum was used as feature vector and the activation barrier height as label. The educt configurations were described by using atom-atom distances, which were then reduced to a minimum by preselection through cross-correlation analysis and requiring a minimal correlation between atom distances and barrier height.

We chose the elastic net regression model [130] to fit our data. It is a type of linear regression model, for which f takes the form [131]

$$f(\mathbf{x}) = \sum_{j=1}^{d} \alpha_j x_j = \alpha^{\mathrm{T}}\mathbf{x} \ , \tag{2.88}$$

where d is the dimension of the feature vector and α_j and x_j are the components of the coefficient and feature vector, respectively. In the elastic net model, least-squares regression [132] is combined with the penalty functions of Lasso [133] (L_1 norm or Manhatten norm of α, i.e., the sum of the absolute entries of α) and Ridge [134] (L_2 norm or Euclidean norm of α) regression.

$$L(y, f(\mathbf{x})) = ||y - f(\mathbf{x})||_2^2 + \lambda_1||\alpha||_1 + \lambda_2||\alpha||_2^2 \tag{2.89}$$

The penalty of Lasso regression enforces variable selection, whereas the Euclidean norm penalizes high coefficients. The effect of the additional penalty functions is called regularization and helps to prevent overfitting. The machine learning process then solves

$$\underset{\alpha\in\mathbb{R}^d}{\arg\min} \left\{ \sum_{i}^{n} ||y - f(\mathbf{x}_i)||_2^2 + \lambda_1||\alpha||_1 + \lambda_2||\alpha||_2^2 \right\} \tag{2.90}$$

for given λ_1 and λ_2, which are the two hyperparameters of this model.

In addition, we tested other regression models as reported in **Publication III**, but the algorithm described above was the simplest that described the given data best.

2.5 Inclusion of Experimental Conditions

2.5.1 Introduction

The principal aim of theoretical chemistry is to build predictive models and help explain experimental results. It is therefore important to compute quantities that are measurable in the lab. In **Publications V** and **VI**, two spectroscopic observables were computed that have a long computational history [135, 136], namely nuclear magnetic resonance (NMR) shieldings and electron paramagnetic resonance (EPR) hyperfine coupling constants (HFCC). The study of NMR chemical shifts focused on a biological system consisting of an enzyme and covalent inhibitor. Simulations of different states were used to determine the actual conformational state of the inhibitor-protein complex in solution. The EPR study looked at a set of different organic radicals. A scheme was devised to reduce computational cost, while at the same time reproducing the experimental conditions more closely than standard procedures. This is the point that unites both studies: calculating an observable under model conditions that mimic the experiment.

In principle, observables are computed as ensemble averages according to eq. (2.2). However, both P_i and O_i can potentially be sources of error. One can generally distinguish between an observable hypersurface and a configurational hypersurface, as both energy and observable are a function of the atom coordinates. Consequently, the level of theory used to generate the configurational ensemble does not have to be the same as the one used for estimating the value of the observable for each configuration of atoms.

Traditionally, quantum-chemical benchmark studies put a large emphasis on the accuracy of O_i; a less severe approximation to the Schrödinger equation yields a better result for a certain configuration of the studied system [137, 138]. In **Publication VI**, we also noted in accordance with previous findings that inclusion of electron correlation is an important factor. However, the focus of this thesis is rather on P_i and modelling experimental conditions than on electronic structure theory.

2.5.2 How to Approximate the Boltzmann Distribution

In the past, computational studies usually focused on the minimum energy configuration of a molecule, since it would have the largest Boltzmann factor, and computed the observable for this one configuration. However, experiments are usually carried out at finite temperatures, so that the molecule samples also other conformations than the global minimum. Here, two different kinds of explorations have to be distinguished: First the vibrational motion of the molecule which is always present even at 0 K, and second the population of local minima that have a higher energy than the global minimum.

The effect of thermal vibrations can be quite significant as studies on NMR and EPR have shown (see, e.g., Refs. [139–141] and references therein). The impact of molecular motion can be as significant as electron correlation effects.

The inclusion of vibrations through the method by Ruud et al. [140] requires second derivatives of the observable with respect to all vibrational modes, which can become costly already for medium sized molecules. The chemical shift itself is a second derivative of the energy, and the number of vibrational modes increases with the system size.

For systems larger than a few atoms, not only vibrations are important, but as well the dependence of an observable on conformation. Therefore, other studies have laid a focus on identifying all local minima of the potential energy surface for which the observable is then computed [142].

Thus, to obtain a proper ensemble average vibrational motions have to be included as well as all conformations accessible at the chosen temperature. To this end, one can either combine conformational search algorithms with vibrational averaging, or use sampling methods such as MD or MC to explore the configurational hypersurface, which includes simultaneously conformational sampling and the vibrational displacement of atoms. As structure generation and observable computation are disconnected, a less costly level of theory appropriate for the system of interest can be chosen for structure generation. Employed levels range from MM (e.g., Ref. [143] or **Publication V**), to *ab initio* or mixed QM/MM (e.g., Refs. [144–146] or **Publication VI**).

2.5.3 Environmental Effects

Most experiments are not performed in either vacuum or the gas phase, but rather in solution or matrix. The environment interacts with the studied molecule, and thus influences the value of the observable. Environmental effects are, for example, especially important for chemical shifts of polarizable hydrogen atoms [147]. The environment can be described either implicitly through polarizable continuum models as done in **Publication X** or explicitly through inclusion of solvent molecules.

In [142] was stated that the search for all minima is superior to sampling, as very long trajectories might be needed to overcome high rotational barriers, and thus to reach ergodicity. However, the inclusion of explicit solvent molecules makes the search for local minima difficult, if not impossible. The sampling not only explores multiple minima on the potential energy surface, but also includes vibrations. Therefore, in **Publications V** and **VI** all mentioned effects - configurations, vibrations, and environment - were included through MD simulations that sample the configuration space while at the same time allowing inclusion of solvent molecules.

2.5.4 Reducing Computational Effort

In summary, an ideal scenario uses a protocol that generates a set of configurations that closely resembles the Boltzmann distribution and uses an approximation to the Schrödinger equation that does not neglect important electronic effects. Such a scenario can obviously become costly. Hence, it is important to devise schemes that include the contributions (sampling minima and vibrations, interactions with the environment, and electron correlation), which are at the same time cost effective. In **Publications VI** we have devised an additive protocol, which uses several levels of theory for both structure generation and observable computation. It can be summarized with an equation:

$$\langle O \rangle = O_{\mathrm{low-level}} + \Delta_{\mathrm{method}} + \Delta_{\mathrm{corr}} + \Delta_{\mathrm{dyn}} + \Delta_{\mathrm{solv}} \tag{2.91}$$

It is possible to start out with the configuration of the global minimum and the corresponding observable value $O_{\mathrm{low-level}}$, both obtained at low levels of theory, and then

correct it stepwise. The minimum is re-optimized with a higher level of theory and the observable is computed with the same low-level method as before for the new minimum configuration, the observable difference between the minimum energy configurations is denoted with Δ_{method}. Dynamics are performed with the low-level structure method and the observable is computed for many frames to approximate the ensemble average. The difference between the value corresponding to the low-level minimum and the average over simulation frames is Δ_{dyn}. The lower level of theory used for computing observables can be corrected to a higher level including electron correlation for selected configurations (Δ_{corr}). Solvation effects can be treated similarly (Δ_{solv}). In this way, one can circumvent using exclusively high-level theories for structure generation and observable computation. Some of those ideas have also been considered in Refs. [145, 148]. Such schemes help to obtain reliable estimates of the Boltzmann distribution as well as of the observable, and thus values that can be more reliably compared with experiment.

Chapter 3

Publications

3.1 Publication I: Calculating free energies from the vibrational density of states function: Validation and critical assessment

Laurens D. M. Peters, Johannes C. B. Dietschreit,
Jörg Kussmann, and Christian Ochsenfeld
"Calculating free energies from the vibrational density of states function:
Validation and critical assessment"
J. Chem. Phys. **2019**, *150*, 194111

Abstract: We explore and show the usefulness of the density of states function for computing vibrational free energies and free energy differ- ences between small systems. Therefore, we compare this density of states integration method (DSI) to more established schemes such as Bennett's Acceptance Ratio method (BAR), the Normal Mode Analysis (NMA), and the Quasiharmonic Analysis (QHA). The strengths and shortcomings of all methods are highlighted with three numerical examples. Furthermore, the free energy of the ionization of ammonia and the mutation from serine to cysteine are computed using extensive ab initio molecular dynamics simulations. We conclude that DSI improves upon the other frequency-based methods (NMA and QHA) regarding the treatment of anharmonicity and yielding results comparable to BAR in all cases without the need for alchemical transformations. Low-frequency modes lead to larger errors indicating that long simulation times might be required for larger systems. In addition, we introduce the use of DSI for the localization of the vibrational free energy to specific atoms or residues, leading to insights into the underlying process, a unique feature that is only offered by this method.

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

Calculating free energies from the vibrational density of states function: Validation and critical assessment

Cite as: J. Chem. Phys. 150, 194111 (2019); doi: 10.1063/1.5079643
Submitted: 1 November 2018 • Accepted: 21 April 2019 •
Published Online: 20 May 2019

Laurens D. M. Peters,[1,2,a] Johannes C. B. Dietschreit,[1,2,a] Jörg Kussmann,[1,2] and Christian Ochsenfeld[1,2,b]

AFFILIATIONS

[1]Chair of Theoretical Chemistry, Department of Chemistry, University of Munich (LMU), Butenandtstr. 7, D-81377 München, Germany
[2]Center for Integrated Protein Science (CIPSM) at the Department of Chemistry, University of Munich (LMU), Butenandtstr. 5–13, D-81377 München, Germany

[a]**Contributions:** L. D. M. Peters and J. C. B. Dietschreit contributed equally to this work.
[b]**Electronic mail:** christian.ochsenfeld@uni-muenchen.de

ABSTRACT

We explore and show the usefulness of the density of states function for computing vibrational free energies and free energy differences between small systems. Therefore, we compare this density of states integration method (DSI) to more established schemes such as Bennett's Acceptance Ratio method (BAR), the Normal Mode Analysis (NMA), and the Quasiharmonic Analysis (QHA). The strengths and shortcomings of all methods are highlighted with three numerical examples. Furthermore, the free energy of the ionization of ammonia and the mutation from serine to cysteine are computed using extensive *ab initio* molecular dynamics simulations. We conclude that DSI improves upon the other frequency-based methods (NMA and QHA) regarding the treatment of anharmonicity and yielding results comparable to BAR in all cases without the need for alchemical transformations. Low-frequency modes lead to larger errors indicating that long simulation times might be required for larger systems. In addition, we introduce the use of DSI for the localization of the vibrational free energy to specific atoms or residues, leading to insights into the underlying process, a unique feature that is only offered by this method.

I. INTRODUCTION

Free energy differences are closely connected to experimental thermodynamic data (e.g., binding affinities, reaction energies, and activation barriers of molecular transformations) as they incorporate contributions from the internal energy as well as from entropy.[1–7] Methods to calculate free energies or their differences are, therefore, of great interest in computational chemistry. Roughly, they can be divided into two groups: (1) frequency- and (2) energy-based methods.

In the first group, the free energy is calculated from frequencies of molecular vibrations (or rotations).[8] These frequencies can be obtained from the second derivative of the energy with respect to the nuclear coordinates at the minimum energy geometry or from the covariance matrix taken from a molecular dynamics[9–11] (MD) or Monte Carlo[12,13] (MC) simulation referring to the Normal Mode Analysis[14,15] (NMA) and the Quasiharmonic Analysis[16,17] (QHA), respectively. Energy-based methods calculate free energy differences from sampled energies along MC or MD simulations, applying exponential averaging theory[18] (EXP), thermodynamic integration[19] (TI), or Bennett's acceptance ratio method[20] (BAR).

All mentioned methods have, despite their great success and broad fields of application,[21–28] well-known shortcomings. The use of NMA, for example, requires the search for the minimum energy geometry (or geometries). It is, therefore, usually applied to small- or medium-sized molecules, using quantum-mechanical (QM) methods. Free energy methods, using data from simulations (QHA, EXP, TI, BAR), usually require a large number of steps to converge, which is connected to the universal problem of sampling the phase space sufficiently to estimate the ratio of partition functions.

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

This challenge has been tackled in many publications,[1–3,29] applying, e.g., alchemical transformations or enhanced sampling techniques. As this may still require long simulation times, the levels of theory for these calculations range from molecular-mechanical (MM) over semiempirical to combined quantum-mechanical/molecular-mechanical (QM/MM) methods, depending on the size of the simulated system and the problem at hand. Additional shortcomings are the harmonic approximation in NMA and QHA and the neglect of vibrational quantum effects[30] in EXP, TI, and BAR.

An alternative approach has been proposed by Berens *et al.*[31] by calculating free energies and free energy differences as a weighted integral over the density of states function, which is determined from sampled nuclear velocities along MD simulations. Although already developed in 1983, this method [named integration of the density of states method (DSI) in the following] has only been used occasionally for absolute entropy calculations[32,33] or the calculation of solvation entropies.[34–36] Therefore, its convergence with respect to the simulation time and the number of independent trajectories has not been investigated in detail. Nevertheless, we expect the sampling problem to be as crucial in DSI as in the other simulation-based methods as the vibrational partition function can again only be approximated (in this case via the density of states function).

In this work, we compare DSI to the established methods mentioned above and validate its use for (1) free energy calculations of molecular transformations and (2) localization of the free energy to specific atoms or residues. First, we briefly recapitulate the theory of EXP, BAR, NMA, QHA, and DSI (Sec. II A) and introduce our novel ansatz to obtain atomic contributions of vibrational free energy changes using DSI (Sec. II B). Having listed the computational details in Sec. III, we compare the free energy methods in Sec. IV. For this purpose, we (1) investigate three numerical examples, (2) determine the ionization potential of ammonia, and (3) calculate the free transformation energy from serine to cysteine *in vacuo* from *ab initio* MD simulations[37,38] using HF-3c.[39] Conclusions are drawn in Sec. V.

II. THEORY

The free energy (A) and the free energy difference between two systems 0 and 1 ($\Delta A_{0\to1}$) can be calculated from the partition function (Q) as

$$A = -\beta^{-1} \ln Q, \tag{1}$$

$$\Delta A_{0\to1} = -\beta^{-1} \ln \frac{Q_1}{Q_0}, \tag{2}$$

where β is equal to $1/(k_B T)$ with k_B being the Boltzmann constant and T being the absolute temperature. It is generally assumed that A can be separated into contributions of translation (A_{trans}), rotation (A_{rot}), and vibration (A_{vib}) as well as the energy of the electronic ground state (E)

$$\begin{aligned} A &= E + A_{trans} + A_{rot} + A_{vib} \\ &= E - \beta^{-1} \ln\{Q_{trans} \times Q_{rot} \times Q_{vib}\}, \end{aligned} \tag{3}$$

where Q_{trans}, Q_{rot}, and Q_{vib} are the corresponding partition functions. In this article, we restrict ourselves to the calculation of vibrational free energies. In the simulations, this is realized by removing the center of mass translation and the overall rotation of the system in each step of the molecular dynamics simulation.

A. Review of free energy methods

1. Energy-based methods

The partition function of a canonical ensemble (NVT) is defined as

$$Q \propto \int d\mathbf{x} \exp\{-\beta U(\mathbf{x})\}. \tag{4}$$

$U(\mathbf{x})$ is the potential energy at a given nuclear structure $\mathbf{x}$, whereas the kinetic energy terms are part of the proportionality constant. Equation (2) can thus be transformed into

$$\Delta A_{0\to1} = -\beta^{-1} \ln \frac{\int d\mathbf{x} \exp\{-\beta U_1(\mathbf{x})\}}{\int d\mathbf{x} \exp\{-\beta U_0(\mathbf{x})\}}. \tag{5}$$

$U_0(\mathbf{x})$ and $U_1(\mathbf{x})$ are the potential energy functions of systems 0 and 1, respectively. In exponential averaging theory (EXP), the difference between the potential energies of the two systems $\Delta U = U_1 - U_0$ is calculated so that[18]

$$\Delta A_{0\to1} = -\beta^{-1} \ln\langle \exp\{-\beta \Delta U(\mathbf{x})\}\rangle_0, \tag{6}$$

where $\langle B(\mathbf{x})\rangle_0$ denotes the ensemble average of $B(\mathbf{x})$ over configurations sampled from the reference system 0. In many cases, the underlying distribution of ΔU is too wide for an efficient calculation of ΔA so that a coupling parameter $\lambda \in [0; 1]$ is introduced, which gradually transforms system 0 into system 1 and thus creates a better overlap of the distributions[2,18,40]

$$U_\lambda(\mathbf{x}) = (1-\lambda) U_0(\mathbf{x}) + \lambda U_1(\mathbf{x}). \tag{7}$$

This transformation, which can be, for example, a chemical reaction or an artificial (so-called alchemical) transformation, is then separated into M sufficiently small steps of size $\Delta\lambda_i$, and the free energy difference of each step is calculated individually leading to

$$\Delta A_{0\to1} = -\beta^{-1} \sum_{i=0}^{M-1} \ln\langle \exp\{-\beta \Delta U_i(\mathbf{x})\}\rangle_{\lambda_i}, \tag{8}$$

with

$$\Delta U_i(\mathbf{x}) = U_{\lambda_{i+1}}(\mathbf{x}) - U_{\lambda_i}(\mathbf{x}) = (\lambda_{i+1} - \lambda_i) \Delta U(\mathbf{x}) = \Delta\lambda_i \Delta U(\mathbf{x}) \tag{9}$$

and

$$\sum_{i=0}^{M-1} \Delta\lambda_i = 1. \tag{10}$$

In the additive scheme of Eq. (8), forward ($\Delta A_{0\to1}$) and backward ($-\Delta A_{1\to0}$) calculations of the free energy differ in almost all cases, again due to the different distributions.[41] This error can be reduced by increasing the sampling of the system or by applying the double-wide sampling scheme.[42]

A more sophisticated approach to obtain the "best" free energy from forward and backward calculations has been derived by Bennett in 1976 [see Eq. (11)].[20] It minimizes the variance of ΔA and is equivalent to its maximum likelihood estimator, as shown by Shirts and Pande in 2003,[43]

$$0 = \ln\left[\frac{\sum_F^{N_{0\to1}} f(M + \beta\Delta U_{0\to1}^F - \beta\Delta A_{0\to1})}{\sum_B^{N_{1\to0}} f(-M + \beta\Delta U_{1\to0}^B + \beta\Delta A_{0\to1})}\right], \tag{11}$$

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

$$M = \ln \frac{N_{0\to1}}{N_{1\to0}}. \tag{12}$$

f is the Fermi function $f(x) = \frac{1}{1+\exp(x)}$; $\Delta U^F_{0\to1}$ and $\Delta U^B_{1\to0}$ are independent forward and backward perturbations, respectively; and $N_{0\to1}$ and $N_{1\to0}$ are the corresponding numbers of frames. The resulting Bennett's Acceptance Ratio method (BAR) is known to be more robust than EXP or thermodynamic integration (TI) schemes.[44–49]

2. Frequency-based methods

Frequency-based methods assume that the potential energy function can be approximated by a sum of $N_F - 6$ harmonic oscillators (harmonic approximation)

$$U(\mathbf{x}) = \sum_{ij}^{N_F-6} k_{ij}(x_i - x_i^0)(x_j - x_j^0), \tag{13}$$

where N_F is the number of degrees of freedom of the system and k_{ij} are the force constants. In the Normal Mode Analysis (NMA),[14,15] the Hessian matrix (the second derivative of the energy with respect to the nuclear coordinates) at the minimum energy configuration ($\mathbf{x}^0$),

$$H_{ij} = \frac{\partial^2 E}{\partial x_i \partial x_j}, \tag{14}$$

is diagonalized, yielding the normal modes ν_i, which are then used to calculate the vibrational free energy either classically (CL) or quantum-mechanically (QM)

$$A^{\mathrm{CL}}_{\mathrm{vib}} = \beta^{-1} \sum_i \ln[\beta h \nu_i], \tag{15}$$

$$A^{\mathrm{QM}}_{\mathrm{vib}} = \beta^{-1} \sum_i \ln\left[\frac{1-\exp(-\beta h\nu_i)}{\exp(-\frac{1}{2}\beta h\nu_i)}\right], \tag{16}$$

where h is the Planck constant. While NMA works well for small systems, great care has to be taken in the case of large systems. Here, NMAs have to be performed at all (relevant) local minima and the results have to be weighted by the Boltzmann factor of the respective minimum. This task becomes harder with increasing system size.

The search for minima is not required in the Quasiharmonic Analysis (QHA)[16,17] as it is performed after a molecular dynamics or Monte Carlo simulation. Assuming ergodicity and that all x_i are Boltzmann distributed, ν_i can be obtained by diagonalizing the mass weighted covariance matrix

$$\left[\mathbf{M}^{\frac{1}{2}}\boldsymbol{\sigma}\mathbf{M}^{\frac{1}{2}} - \beta^{-1}\boldsymbol{\nu}\right]\mathbf{M}^{\frac{1}{2}}\Delta\mathbf{x} = 0, \tag{17}$$

where $\mathbf{M}$ is the kinetic energy matrix and $\boldsymbol{\sigma}$ is the covariance matrix,

$$\sigma_{ij} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle. \tag{18}$$

$\langle\,\rangle$ symbolizes the average over all trajectory frames. The frequencies obtained this way can then again be transformed into vibrational free energies with Eqs. (15) and (16).

Frequencies obtained with QHA are always equal or lower to those of the NMA since it approximates all possible minima along one coordinate as well as its anharmonicity with one single harmonic oscillator. This leads to a lower curvature of the potential energy surface than actually present.

3. Integration of the density of states method

The vibrational partition function (Q_{vib}) can also be calculated as a product of the partition functions $q(\nu)$ of the single vibrational modes with frequency ν. It is assumed that these partition functions can be written as classic (q^{CL}) or quantum (q^{QM}) harmonic oscillators

$$q^{\mathrm{CL}}(\nu) = \frac{1}{\beta h \nu}, \tag{19}$$

$$q^{\mathrm{QM}}(\nu) = \frac{\exp(-\frac{1}{2}\beta h\nu)}{1-\exp(-\beta h\nu)}. \tag{20}$$

The logarithm of the vibrational partition function can thus be calculated from the following integral:[31]

$$\ln Q_{\mathrm{vib}} = \int_0^\infty d\nu D(\nu) \ln\{q(\nu)\}. \tag{21}$$

D is the density of states function, which singles out the specific frequencies of the investigated system, while an infinite number of harmonic oscillators (classical or quantum) is considered. D itself is determined as the mass-weighted Fourier transform of the nuclear velocity autocorrelation

$$D(\nu) = 2\beta \sum_{j=1}^{N_A} m_j \int dt \langle \mathbf{v}_j(\tau)\mathbf{v}_j(\tau+t)\rangle_\tau e^{-i2\pi\nu t}. \tag{22}$$

m_j and $\mathbf{v}_j$ denote the mass and velocity vector of the nucleus j, respectively, while N_A is the number of atoms. Integration over the entire density of states functions yields the number of degrees of freedom

$$N_F = \int_0^\infty d\nu D(\nu). \tag{23}$$

Insertion of Eq. (21) into Eq. (1) yields the following expression for the free energy:

$$A = E + \beta^{-1} \int_0^\infty d\nu D(\nu) W_A(\nu). \tag{24}$$

W_A is, depending on the inserted q, the classical (W^{CL}_A) or quantum (W^{QM}_A) weighting function

$$W^{\mathrm{CL}}_A(\nu) = \ln[\beta h\nu], \tag{25}$$

$$W^{\mathrm{QM}}_A(\nu) = \ln\left[\frac{1-\exp(-\beta h\nu)}{\exp(-\frac{1}{2}\beta h\nu)}\right]. \tag{26}$$

If one assumes that D consists of delta functions at the frequencies of the normal modes (ν_i) of a system, Eq. (24) with W^{CL}_A and W^{QM}_A is equal to Eqs. (15) and (16), respectively. The difference between the integration of the density of states method (DSI) and the other frequency-based methods is, thus, that instead of $N_F - 6$ harmonic oscillators

$$Q_{\mathrm{vib}} = \prod_i^{N_F-6} q(\nu_i), \tag{27}$$

an arbitrary large number of harmonic oscillators (depending on the simulation time) weighted by the density of states function

$$Q_{\mathrm{vib}} = \prod_i q(\nu_i)^{D(\nu_i)} \tag{28}$$

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

are assumed to describe the system (harmonic approximation).

From Eq. (24), one can easily obtain the expression for free energy differences

$$\Delta A_{0\to1} = \Delta E + \beta^{-1}\int_0^\infty d\nu \Delta D(\nu) W_A(\nu), \tag{29}$$

where $\Delta D = D_1 - D_0$ and $\Delta E = E_1 - E_0$.

B. Atomic contributions to the density of states functions

Our ansatz to determine atomic contributions to the vibrational free energy uses the fact that the density of states function from Eq. (22) can be rewritten as a superposition of atomic functions

$$\begin{aligned} D(\nu) &= \sum_{j=1}^{N_A} 2\beta m_j \int dt \langle \mathbf{v}_j(\tau)\mathbf{v}_j(\tau+t)\rangle_\tau e^{-i2\pi\nu t} \\ &= \sum_{j=1}^{N_A} D_j(\nu). \end{aligned} \tag{30}$$

This essentially means that the vibrational partition function can be written, without any further loss of generality, as

$$\begin{aligned} Q_{\text{vib}} &= \prod_j^{N_A}\prod_i q(\nu_i)^{D_j(\nu_i)} \\ &= \prod_j^{N_A} Q_j^{\text{vib}} \end{aligned} \tag{31}$$

because

$$\prod_j^{N_A} q(\nu_i)^{D_j(\nu_i)} = q(\nu_i)^{D(\nu_i)}. \tag{32}$$

Thus, the vibrational free energy can be written as a sum over atomic contributions

$$\begin{aligned} A_{\text{vib}} &= -\beta^{-1}\ln Q_{\text{vib}} \\ &= -\beta^{-1}\ln\left[\prod_j^{N_A} Q_j^{\text{vib}}\right] \\ &= -\beta^{-1}\sum_j^{N_A}\ln Q_j^{\text{vib}} \\ &= \sum_j^{N_A} A_j^{\text{vib}}. \end{aligned} \tag{33}$$

The above partitioning of the vibrational free energy is not restricted to atoms. It can without any further assumption be grouped into any meaningful collection of atoms such as residues or functional groups. This ansatz can be used as an aide to interpret and localize the changes occurring in the system, as shown in our previous work.[50] Please note that it uses the vibrational free energy only and not the total free energy. For the partitioning of the latter, approximate schemes exist[21] but have been discussed to lead to unreliable results.[51,52] Our only assumption along with the harmonic approximation is that the regions with the most prominent changes in the vibrational free energy are also those which contribute the most to the change in E and thus the total free energy change.

III. COMPUTATIONAL DETAILS

A. Classical and molecular dynamics simulations

The free energies of the numerical examples have been obtained from classical NVT simulations of a particle (of mass 1 u) in a one-dimensional harmonic (V_H), Morse (V_A), and double-well (V_D) potentials, respectively,

$$V_H(x) = \frac{1}{2}kx^2, \tag{34}$$

$$V_A(x) = D_E \times (1 - \exp\{-ax\})^2, \tag{35}$$

$$V_D(x) = \frac{1}{2}bx^2(x-0.5)^2. \tag{36}$$

The exact values for k, D_E, a, and b are listed in the supplementary material. To obtain an exact reference, we have integrated Eq. (5) numerically on a grid using $\approx 10^6$ points and a step width of 0.01 Bohr. For the harmonic oscillator, this procedure leads to an error below 10^{-4} kJ/mol.

The free energies of the molecular systems have been obtained from Born-Oppenheimer molecular dynamics simulations at the HF-3c[39] level of theory using the FermiONs++ program package[53–55] with DFTD3 v3.1[56,57] and gCP v2.02.[58] The center of mass translation and the overall rotation of the system have been removed at every step of the simulation.

All simulations use the Velocity Verlet[59,60] propagator and the random rescaling thermostat by Bussi, Donadio, and Parrinello,[61] keeping the average temperature at 298.15 K. A different thermostat, like a Langevin-thermostat, would in general have been better suited to sample our small systems.[47] However, the random changes of the nuclear forces would severely impact the velocity autocorrelation function and render our analysis impossible. Initial velocities have been drawn from a Maxwell-Boltzmann distribution at 298.15 K. Energies, velocities, and coordinates were written to files every 1 fs. The numerical examples are simulated for 110 ps (10 ps equilibration time, 0.1 fs time step) or, in some cases, 1010 ps (10 ps equilibration time, 0.1 fs time step). The simulation times of the molecular systems are 310 ps (10 ps equilibration time, 0.1 fs time step) in the case of ammonia and 202 ps (2 ps equilibration time, 0.2 fs time step) in the case of serine and cysteine. For every λ window, we have calculated five independent trajectories and an equidistant $\Delta\lambda$ of 0.1 has been applied. To show the convergence behavior of DSI in the ammonia example, we have additionally conducted 10 independent trajectories of 910 ps (10 ps equilibration time, 0.1 fs time step) for NH_3 and NH_3^+.

B. *Ab initio* alchemical transformations

Alchemical transformations are normally used in a molecular mechanics (MM) context, where transforming one system (0) into another system (1) is equal to gradually turning on (or off) contributions to the potential energy.[2] Here, we want to use this concept with *ab initio* calculations, which do not allow for such a fragmentation of the energy. To circumvent this problem, we use an ansatz developed by Reddy *et al.*[62] We perform two energy and forces calculations (for systems 0 and 1) at every step of the simulation and continue the trajectory along a weighted force $\mathbf{F}_\lambda$,

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

$$\mathbf{F}_\lambda = (1-\lambda)\mathbf{F}_0 + \lambda\mathbf{F}_1, \tag{37}$$

where $\mathbf{F}_0$ and $\mathbf{F}_1$ denote the determined forces of system 0 and system 1, respectively. Consequently, we use the weighted mass, temperature, center of mass velocity, inertia tensor, and total angular momenta in our thermostat and when removing the overall translation and rotation. In this work, we follow the single-topology ansatz.[63] In the case of the serine-cysteine transformation, this means that both systems (0 and 1) share the same structure, with the oxygen in system 0 being replaced by a sulfur in system 1. The use of a dual-topology ansatz (the OH-group in system 0 and the SH-group in system 1 have different structures, while the rest of the molecule is shared) is also possible.[62,64,65] However, when it is applied to simulations without explicit solvent molecules, an MM region, or geometrical constraints, the dual-topology ansatz leads to unstable trajectories. The reason for this is that, when $\lambda \approx 0$ or $\lambda \approx 1$, the OH– or SH–group is not "seen" by the shared part of the molecule, leading to unphysical geometries (large C–O or C–S bonds) and convergence problems of the self-consistent field algorithm. Constraints or surroundings will prevent this.

C. Free energy calculations

The density of states function (D) was calculated from the sampled velocities using Eq. (22) and subsequently rescaled so that Eq. (23) yields the $3N_A - 6$ vibrational degrees of freedom for the complete system. To allow for an easy comparison especially between chemically identical atoms, single atom spectra were rescaled so that Eq. (23) yields three. ΔA^{DSI} was calculated following the integration in Eq. (24) or (29) and for the molecular examples adding E. E is determined as the potential energy at the minimum geometry, which was for the intermediate systems ($0 < \lambda < 1$) obtained by performing a geometry optimization with the weighted forces [see Eq. (37)] until $\mathbf{F}_\lambda \approx \mathbf{0}$. ΔA^{BAR} is determined from the sampled potential energies by solving Eq. (11). In the case of the serine-cysteine transformation, the free energy change due to the mass change of the atom (oxygen to sulfur) was corrected by an analytically derived constant for each window. For the derivation, see the Appendix. ΔA^{NMA} and ΔA^{QHA} were obtained using Eqs. (15) and (16). The frequencies (ν_i) for the NMA were determined using the numerically calculated Hessian at the minimum energy geometry [Eq. (14)], while ν_i for the QHA were calculated as presented in Eq. (17).

The vibrational parts $\Delta A_{\mathrm{vib}}^{\mathrm{DSI}}$ and $\Delta A_{\mathrm{vib}}^{\mathrm{BAR}}$ are calculated as

$$\Delta A_{\mathrm{vib}}^{\mathrm{DSI}} = \Delta A^{\mathrm{DSI}} - \Delta E, \tag{38}$$

$$\Delta A_{\mathrm{vib}}^{\mathrm{BAR}} = \Delta A^{\mathrm{BAR}} - \Delta E. \tag{39}$$

As we have simulated several replicas for each λ window, we conduct a statistical analysis calculating the average free energy difference ($\langle \Delta A \rangle$) as

$$\langle \Delta A \rangle = \frac{1}{N_{\mathrm{E}} N_{\mathrm{P}}} \sum_i^{N_{\mathrm{E}}} \sum_j^{N_{\mathrm{P}}} \Delta A_{ij}, \tag{40}$$

where ΔA_{ij} is the free energy difference between the replicas i and j of the educt and product, respectively, and N_{E} and N_{P} are their total numbers. We, additionally, calculate the standard deviation of the different ΔA_{ij}. We do not list the inherent statistical error calculated by Shirts and Pande[45] since there is no analog for DSI.

IV. RESULTS AND DISCUSSION

A. Numerical examples

In order to prove that DSI yields the same results as other free energy methods and to investigate the effect of the shape and curvature of the potential on its accuracy, we carried out classical simulations in an one-dimensional harmonic [Eq. (34)], a Morse [Eq. (36)], and a double-well [Eq. (36)] potential and calculated three free energy changes for each potential (for details, see Sec. III). The free energy changes consist of changes in the curvature of the potential caused by variation of parameters k, a, D_{E}, and b, resembling changes in molecular angles, bonds, and dihedral angles, respectively (the exact values are listed in the supplementary material). The resulting $\langle \Delta A \rangle$s calculated using DSI, BAR, QHA, and NMA as well as the exact results are shown in Tables I and II. For the simulation-based methods, we also provide the standard deviation of ΔA from multiple trajectories. $\langle \Delta A \rangle$s of the individual λ-windows, potential plots, and density of states plots can be found in the

TABLE I. Calculated free energy changes (average and standard deviation of ΔA_{vib} in kJ/mol) of the harmonic and anharmonic potential (three transformations each) using NMA, QHA, BAR, and DSI. The exact result obtained from numerical integration is given as a reference. The wavenumber (in cm^{-1}) refers to the curvature of the potential at $x = 0$.

Potential	Wavenumbers	NMA	QHA	BAR	DSI	Exact
Harmonic	1000 → 2000	1.718	1.758 ± 0.041	1.709 ± 0.013	1.720 ± 0.002	1.718
	500 → 1000	1.718	1.780 ± 0.068	1.755 ± 0.013	1.732 ± 0.003	1.718
	100 → 500	3.990	4.488 ± 0.056	4.614 ± 0.042	4.166 ± 0.015	3.990
	100 → 500[a]		3.996 ± 0.048	3.858 ± 0.024	4.007 ± 0.002	
Anharmonic	1000 → 2000	1.718	1.796 ± 0.041	1.735 ± 0.015	1.757 ± 0.002	1.766
	500 → 1000	1.718	1.875 ± 0.076	1.787 ± 0.015	1.766 ± 0.004	1.764
	100 → 500	3.990	4.635 ± 0.043	4.806 ± 0.032	4.260 ± 0.011	4.033
	100 → 500[a]		4.104 ± 0.070	3.885 ± 0.030	4.047 ± 0.002	

[a]Trajectories with longer simulation times.

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

TABLE II. Calculated free energy changes (average and standard deviation of ΔA_{vib} in kJ/mol) of the double well potential (three transformations) using NMA, QHA, BAR, and DSI. The exact result obtained from numerical integration is given as a reference. The wavenumber (in cm^{-1}) refers to the curvature of the potential at $x = 0$. Double well potentials can be seen as "worst-case" examples for the frequency-based methods.

Potential	Wavenumbers	NMA	QHA	BAR	DSI	Exact
Double well	2000 → 2500	0.553	0.592 ± 0.062	0.581 ± 0.004	0.598 ± 0.005	0.608
	1500 → 2000	0.713	4.243 ± 0.076	0.810 ± 0.008	0.910 ± 0.014	0.825
	1000 → 1500	1.005	−0.159 ± 0.147	1.269 ± 0.008	1.682 ± 0.036	1.068
	1000 → 1500[a]		−0.151 ± 0.042	1.224 ± 0.008	1.611 ± 0.029	

[a]Trajectories with longer simulation times.

supplementary material. Please note that ΔE is for all numerical cases zero so that $\Delta A = \Delta A_{vib}$.

Comparing the methods that use sampled data along classical simulations (QHA, BAR, DSI) in the harmonic and anharmonic cases (see Table I), DSI performs best, showing smaller errors and standard deviations. In the latter case, it even outperforms NMA, which is only exact, when harmonic potentials are investigated. The error of all three simulation-based methods increases with decreasing curvature (wavenumbers) of the investigated potentials. The standard deviations of BAR and DSI also increase, while it remains constant in the case of QHA. The first reason for the larger errors (and standard deviations) is the sampling error as slower vibrations require longer simulation times to be sampled accurately. The second reason is the choice of the thermostat. As already discussed in Sec. III A, we are bound to velocity rescaling thermostats when applying DSI. These thermostats, however, introduce errors in the free energy calculations, especially for systems with only a few degrees of freedom and for slow modes. Consequently, we observe a decrease in the errors when the relaxation time of the thermostat is increased (resulting in a weaker thermostat). Longer simulation times tackle both problems discussed above and, therefore, improve the errors (and standard deviations of BAR and DSI) for the low-frequency harmonic and anharmonic cases significantly (see footnote a in Table I), which are then in good agreement with numerical test potentials found elsewhere.[45,66]

For the frequency-based methods, the double well potential is a "worst-case" example as it cannot be described exactly within the harmonic approximation. It features two types of movements: the movement within one well and the slower movement over the barrier (the inversion). With decreasing curvature (wavenumber), the barrier height shrinks, increasing the probability of the inversion. This set up is not problematic for BAR as it relies on energy averages and distributions and uses intermediate systems ($0 < \lambda < 1$) to enhance the sampling efficiency. This explains the small error and standard deviation in Table II, which is comparable to the values of the other examples.

For NMA, QHA, and DSI, it serves as a good showcase to illustrate the conceptual differences between these methods and to show how well the actual potential can be approximated. In Fig. 1(a), we have plotted the double well potential (1000 cm^{-1}) and the corresponding harmonic potentials, which are used by NMA and QHA to approximate the potential and to calculate the free energy. Figure 1(b) presents the same for DSI. The difference is that not a single but a series of harmonic potentials (illustrated by the red and blue areas) weighted by the density of states functions [also plotted in Fig. 1(b)] are assumed to describe the system.

The NMA harmonic oscillator, derived from a finite difference calculation around x = 0, mimics the fast vibration, while the QHA harmonic oscillator, derived from the distribution of x,

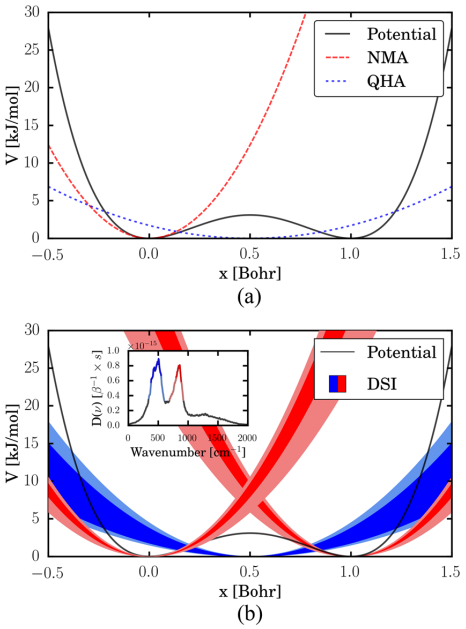


FIG. 1. (a) Double well potential (1000 cm^{-1}) and corresponding harmonic oscillators of NMA and QHA from which the free energy is calculated. (b) Double well potential (1000 cm^{-1}) and the series of weighted harmonic oscillators used in DSI to obtain the free energy. The frequencies have been extracted from the density of states function (see the subplot). Please note that the density of states function ($D(\nu)$) is given in $\beta^{-1} \times s$; we have omitted the factor β in Eq. (22).

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

is dominated by the slow vibration. In this example, the first is a good approximation (most likely due to error compensation), while the latter leads to erratic results (Table II). The density of states function shows that both vibrations are considered in DSI as two peaks with the same intensity appear in the spectrum. In this case, both vibrations have the same probability due to the low barrier height of around 2.5 kJ/mol ≈ RT. When larger barrier heights are applied, the intensity of the fast vibration becomes significantly larger than the intensity of the slow vibration (see the supplementary material). Note that the parameter b changes not only the barrier height but also the curvature of the potential at the same time.

In this example, longer simulation times lead to better DSI results, but the difference to the exact result is still significantly larger than in BAR. This indicates that even an arbitrary large number of harmonic oscillators are incapable of describing the system correctly. However, the description of the system in DSI is physically more correct than in the case of NMA and QHA, and the resulting free energy estimation is significantly better than in the case of QHA.

B. Ionization energy of ammonia

As a first molecular example, we have chosen the ionization of ammonia ($NH_3 \rightarrow NH_3^+$). (Alchemical) *ab initio* molecular dynamics simulations have been performed at the HF-3c level of theory (see Sec. III for details). The results for the overall reaction and for the individual λ-windows are listed in Table III.

Table III proves that the BAR result can be improved by taking into account intermediate λ-windows as the cumulative result (using all intermediate windows) differs from the direct result (using only the two end points) while featuring a 10 times smaller spread. This is not the case for DSI. Since all contributions of the intermediate λ-simulations for E and D cancel out, the results are identical. Only the standard deviation is larger in the cumulative case due to the noise between the intermediate states. The convergence of the direct DSI result and the cumulative BAR result are shown in Fig. 2.

TABLE III. Calculated vibrational free energy changes (average and standard deviation of ΔA_{vib} in kJ/mol) of the ionization of ammonia (direct from the first to the last λ-value and cumulative over all windows) and the intermediate λ-windows using BAR, DSI, and NMA.

λ-window	ΔA_{vib}^{DSI}	ΔA_{vib}^{BAR}	ΔA_{vib}^{NMA}
0.0 → 0.1	−0.26 ± 1.05	−1.08 ± 0.03	...
0.1 → 0.2	−0.45 ± 0.82	−0.94 ± 0.03	...
0.2 → 0.3	−0.31 ± 0.65	−0.44 ± 0.04	...
0.3 → 0.4	−0.96 ± 0.59	0.22 ± 0.04	...
0.4 → 0.5	0.90 ± 0.79	0.35 ± 0.06	...
0.5 → 0.6	−0.48 ± 1.06	0.08 ± 0.05	...
0.6 → 0.7	−0.24 ± 0.86	−0.03 ± 0.03	...
0.7 → 0.8	−0.39 ± 0.55	−0.11 ± 0.02	...
0.8 → 0.9	0.36 ± 0.42	−0.17 ± 0.03	...
0.9 → 1.0	−0.41 ± 0.30	−0.22 ± 0.03	...
Cumulative	−2.24 ± 2.70	−2.36 ± 0.12	...
Direct	−2.24 ± 0.77	−2.50 ± 1.56	−1.97
Quantum corrected	−13.76 ± 5.06	...	−13.26

The standard deviation on both curves decreases with increasing simulation time. After 250 ps simulation time (per trajectory), both methods yield the same result within one standard deviation of the BAR curve, and at ≈300 ps, the mean results are nearly identical (see also Table III). Even longer simulations (up to 900 ps) do not substantially affect the average DSI result, while the standard deviation is reduced (see Fig. S7 of the supplementary material). However, even when 2 × 10 trajectories of 900 ps are used in the DSI calculation, the standard deviation is still approximately two times larger than the one observed in BAR featuring an (almost) equivalent amount of data points. If one takes into account that no alchemical simulations (one energy and force calculation per step instead of two) are required for DSI, we could say that (for this example) the standard deviations of DSI and BAR behave similarly with respect to the computation time, while the average free energy change seems to converge faster in the case of DSI. At this point, we also want to mention that there are two factors which can decrease the accuracy of DSI: Too short simulations and too long intervals between the sampling of the nuclear velocities are applied (see Figs. S7 and S8 of the supplementary material).

The results of DSI and BAR for the intermediate λ-windows differ usually by about one standard deviation, except for 0.3 → 0.4. The histograms of the improper dihedral of ammonia (Fig. S4 of the supplementary material) reveal that for these cases the system is similar to the double well system we have discussed in Sec. IV A, which explains the larger error. In the other windows, the barrier is either too high for a frequent inversion of the molecule or vanishes entirely. The figure also shows that for $\lambda = 0.0$ the simulation has not spent equal amounts of time in the two minima of ammonia, which should bias the BAR results. In general, we observe a relatively high standard deviation for the DSI free energies of the intermediate λ-windows. The reason for this could be that mixed potential energy surfaces tend to be more anharmonic or even nonharmonic (e.g., the λ-window 0.3 → 0.4), showing larger errors and slower convergence.

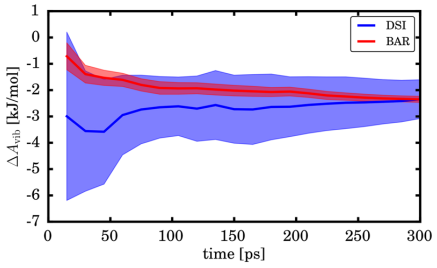


FIG. 2. Convergence of the total free energy change ($\langle \Delta A \rangle$, solid line) of the ionization of ammonia and the standard deviation (lighter area) with respect to the length of the used trajectories using BAR and DSI. BAR contains information from 11 × 5 trajectories (five replicas for all 11 λ-windows), whereas the DSI result is only based on 2 × 5 trajectories (five replicas for NH_3 and NH_3^+, respectively).

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

Additionally, the DSI method offers two features that are not accessible in energy-based methods. One can easily calculate the quantum corrected free energy change (see the last line in Table III), and one can map the change in $\langle \Delta A_{\text{vib}}^{\text{DSI}} \rangle$ to each atom or when dealing with larger problems, groups of atoms or molecules. In the case of the ionization of ammonia, the hydrogen atoms and the nitrogen atom gain vibrational free energy (0.59 kJ/mol and 0.46 kJ/mol, respectively) since the bonds in NH_3^+ are weaker than in NH_3. This is also reflected in the power spectrum (D, see Fig. 3), where nearly all modes of NH_3 are red-shifted in NH_3^+.

Figure 3 also shows the results of NMA and QHA as vertical dashed lines. As one can see, the frequencies estimated with NMA are in good agreement with D and are always positioned at the upper bound of the peaks in D. This is due to the fact that NMA does not consider any anharmonicity in the bond vibrations, which causes the slight decrease in the vibrational frequency and the vibrational free energy change (see Table III). QHA results clearly underestimate all frequencies and suggest unreasonably slow motions, especially for the inversion motion of NH_3.

C. Mutation from serine to cysteine

The second example consists of the mutation from serine to cysteine in vacuum. Mutations are a widely used tool in free energy calculations as they give access to, e.g., binding free energies. We conducted extensive (alchemical) *ab initio* molecular dynamics simulations of serine, cysteine, and intermediate structures (see Sec. III for details) and calculated the free energy for the overall reaction and for the individual λ-windows. The results are presented in Table IV.

TABLE IV. Calculated vibrational free energy changes (average and standard deviation of ΔA_{vib} in kJ/mol) of the mutation from serine to cysteine (direct from the first to the last λ-value and cumulative over all windows) and the intermediate λ-windows using BAR, DSI, and NMA.

λ-Window	$\Delta A_{\text{vib}}^{\text{DSI}}$	$\Delta A_{\text{vib}}^{\text{BAR}}$	$\Delta A_{\text{vib}}^{\text{NMA}}$
0.0 → 0.1	−0.79 ± 3.03	0.56 ± 0.40	...
0.1 → 0.2	0.06 ± 2.94	−0.06 ± 0.51	...
0.2 → 0.3	−2.10 ± 3.70	−0.74 ± 0.26	...
0.3 → 0.4	0.40 ± 4.03	−0.43 ± 0.31	...
0.4 → 0.5	−3.37 ± 3.80	−0.35 ± 0.32	...
0.5 → 0.6	0.59 ± 3.47	−0.27 ± 0.10	...
0.6 → 0.7	−1.66 ± 3.57	−0.48 ± 0.10	...
0.7 → 0.8	−1.26 ± 4.38	−0.35 ± 0.10	...
0.8 → 0.9	0.66 ± 4.31	−0.35 ± 0.05	...
0.9 → 1.0	−1.47 ± 3.16	−0.31 ± 0.05	...
Cumulative	−8.94 ± 12.46	−2.76 ± 0.85	...
Direct	−8.94 ± 2.76	290.93 ± 177.98	−3.69
Quantum corrected	−23.55 ± 7.24	...	−15.76

The one-step application of BAR in Table IV shows the wrong sign and is about two orders of magnitude too large. Significantly better results for similar one-step mutations have been reported for MM simulations.[67] However, the underlying data consisted of four 168 ns trajectories, which contain nearly 1000 times more conformations than our *ab initio* simulations.

The results of direct DSI are, in comparison with direct BAR, significantly better. The final results of DSI and BAR are within two

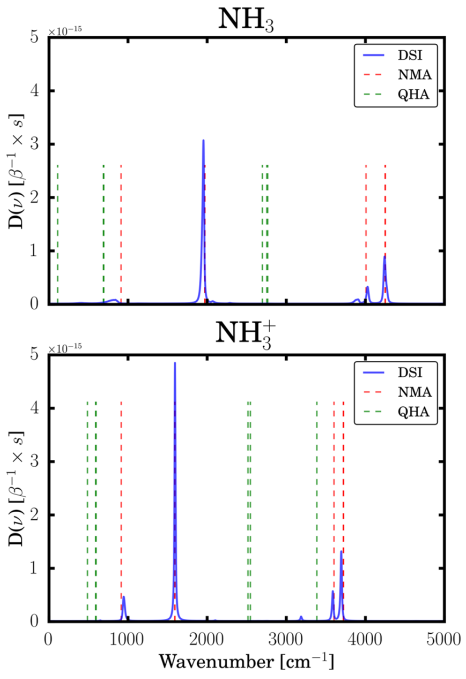


FIG. 3. Velocity density of states spectrum of NH_3 and NH_3^+. The frequencies obtained from NMA and QHA are shown in red and green, respectively. Please note that $D(\nu)$ is given in $\beta^{-1} \times s$; we have omitted the factor β in Eq. (22).

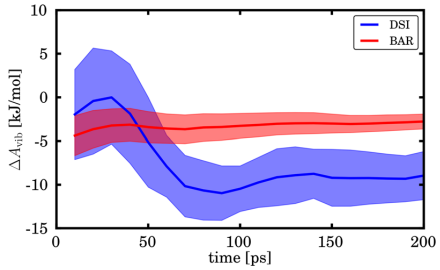


FIG. 4. Convergence of the total free energy change ($\langle \Delta A \rangle$, solid line) of the serine-cysteine transformation and its standard deviation (lighter area) with respect to the length of the used trajectories using BAR and DSI. BAR contains information from 11 × 5 trajectories (five replicas for all 11 λ-windows), whereas the DSI result is only based on 2 × 5 trajectories (five replicas for serine and cysteine, respectively).

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

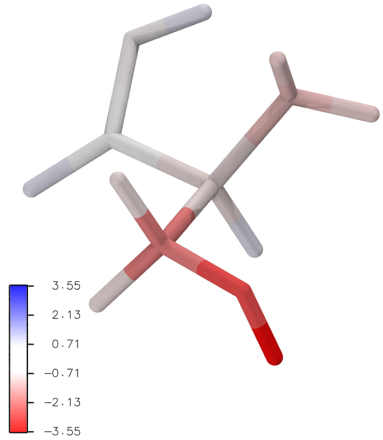


FIG. 5. Vibrational free energy change ($\langle \Delta A_{\text{vib}}^{\text{DSI}} \rangle$ in kJ/mol) of each atom during the mutation from serine to cysteine. The major changes occur exactly at the atoms that are connected or close to the O/S-mutation.

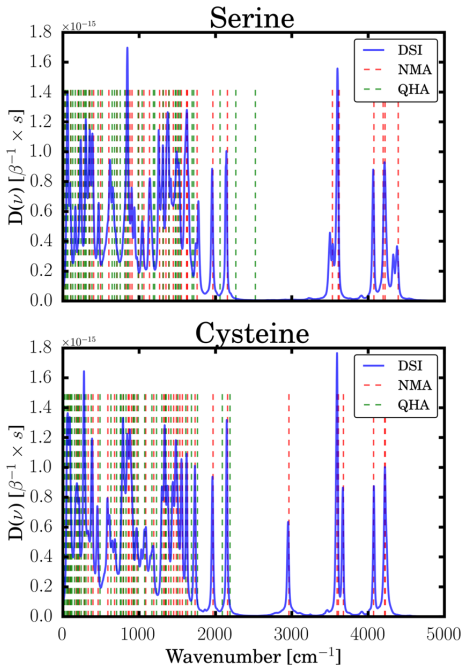


FIG. 6. Velocity density of states spectrum of serine and cysteine. The frequencies obtained from NMA and QHA are shown in red and green, respectively. Please note that $D(\nu)$ is given in $\beta^{-1} \times s$; we have omitted the factor β in Eq. (22).

standard deviations of one another and both are also more than one standard deviation off from the NMA results which lies between the two results. Nevertheless, the standard deviations of DSI are one order of magnitude larger than those of BAR. This and the convergence plot of the free energies (see Fig. 4) indicate that further sampling is required, mainly due to the larger amount of low-frequency modes (see Fig. 6) in this example, which is beyond the focus of the present manuscript.

Despite the numerical noise, the trend given by DSI is correct, and localizing the vibrational free energy (Fig. 5) offers an explanation where and why the free energy changes during the mutation. It clearly shows that main contributors are the C_β-atom as well as the connected hydroxyl- or thiol-group. Slight contributions stem also from the C_α-atom and the amino-group as the strength of the intermittently formed hydrogen bond between hydroxy/thiol-group and amino-group changes. This also has a small effect on the carboxyl-group. These results agree well with chemical intuition.

Figure 6 shows the power spectra of serine and cysteine as well as the NMA and QHA results. The position of the NMA frequencies and the main peaks in the power spectrum agree well. Both show a large amount of low-frequency modes, so-called "breathing modes." Nevertheless, NMA neglects the anharmonicity of the vibrations and the different minima of the system, leading to a different free energy (see Table IV). QHA significantly overestimates the existence of these modes and fails to find the high-frequency bond vibrations.

V. CONCLUSION

In this work, we have tested and compared the density of states method (DSI) to the more established free energy methods such as Normal Mode Analysis (NMA), Quasiharmonic Analysis (QHA), and Bennett's Acceptance Ratio method (BAR), calculating several numerical and two chemical examples. We show that DSI works similar to NMA and QHA, but features the ability to correctly include anharmonicites, as the partition function is approximated by an arbitrary large number of harmonic oscillators weighted by the density of states function.

DSI delivers the same result as BAR for the numerical examples and the ionization of ammonia. Regarding the mutation from serine to cysteine, DSI correctly reflects the trend of the free energy, but features larger standard deviations, mainly due to the large number of low-frequency modes in the systems. This indicates that long simulation times will be required for larger systems. Additional downsides of the method regarding free energy calculations are as follows: (1) larger memory requirements ($3 \times N_A$ velocities have to be stored in short intervals instead of one energy at arbitrary long intervals), (2) Monte Carlo simulations and enhanced sampling methods cannot be combined with DSI, and (3) ΔE has to be determined, which will become tedious for large systems with many degrees of freedom.

There are, however, also important advantages of the method, when comparing to BAR:

1. For DSI, only the end points (no intermediates) are required. This gives access to free energies of nearly all molecular

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

transformations even at the *ab initio* level of theory, circumventing endpoint-catastrophes and alchemical transformations. Additionally, this can lead to a reduction of computation time when DSI is applied to small molecular systems.
2. Quantum-corrected vibrational free energies are directly accessible.
3. A straightforward pattern to determine atomwise or residue-wise contributions to the vibrational free energies exists.

We, therefore, think that DSI can be a good alternative to standard free energy methods, especially when expensive *ab initio* methods are applied to transformations of small to medium-sized molecules. Furthermore, its ability to localize free energy changes at atoms or residues is a valuable tool to gain insights into the underlying process(es), which can always be combined with energy-based methods such as BAR.

SUPPLEMENTARY MATERIAL

The supplementary material comprises details on our numerical examples where we give details on the used potentials and show the corresponding density of states plots as well as ΔA's for the intermediate λ-windows. For the ionization of ammonia, we present dihedral distributions and ΔA convergence studies of the intermediate λ-windows. Additionally, we show the mapping of $\Delta A_{\mathrm{vib}}^{\mathrm{DSI}}$ on the individual atoms and the convergence of $A_{\mathrm{vib}}^{\mathrm{DSI}}$ with respect to simulation time and the sampling frequency. For the serine-cysteine transformation, we present the distributions of the C-O/S bond lengths for all λ-windows.

ACKNOWLEDGMENTS

Financial support was provided by the SFB 1309 "Chemical Biology of Epigenetic Modifications" (DFG), SFB 749 "Dynamics and Intermediates of Molecular Transformations" (DFG), and the DFG cluster of excellence (EXC 114) "Center for Integrative Protein Science Munich" (CIPSM). C.O. acknowledges further support as Max-Planck-Fellow at the MPI-FKF Stuttgart.

APPENDIX: INFLUENCE OF THE MASS CHANGE IN THE FREE ENERGY OF THE TRANSFORMATION FROM SERINE TO CYSTEINE

Changing the mass of a particle has an impact on the free energy as the kinetic energy distribution of the particle changes. The canonical partition function of a system consisting of N_A distinguishable particles has the form

$$Q = \frac{1}{h^{3N_A}} \int \mathrm{d}x^{3N_A} \int \mathrm{d}p^{3N_A} \exp\{-\beta H(x^{3N_A}, p^{3N_A})\}. \tag{A1}$$

The Hamiltonian (H) is usually split into the potential (U) and kinetic energy (T) which are functions of the generalized coordinates ($\mathbf{x}$) and generalized impulses ($\mathbf{p}$), respectively. Hence, the above integral can be split into the product of kinetic and potential energy contributions

$$\begin{aligned} Q &= \frac{1}{h^{3N_A}} \int \mathrm{d}x^{3N_A} \exp\{-\beta U(x^{3N_A})\} \int \mathrm{d}p^{3N_A} \exp\{-\beta T(p^{3N_A})\} \\ &= \frac{1}{h^{3N_A}} \int \mathrm{d}x^{3N_A} \exp\{-\beta U(x^{3N_A})\} \int_{-\infty}^{\infty} \mathrm{d}p^{3N_A} \exp\{-\beta \sum_i^{3N_A} \frac{p_i^2}{2m_i}\}. \end{aligned} \tag{A2}$$

The integration over the kinetic part can be carried out analytically and yields

$$Q = \frac{1}{h^{3N_A}} \prod_i^{3N_A} \sqrt{\frac{2\pi m_i}{\beta}} \int \mathrm{d}x^{3N_A} \exp\{-\beta U(x^{3N_A})\}. \tag{A3}$$

If we consider now the free energy difference between two systems, where not only the potential energy function changes but also the mass of one particle, we can write

$$\begin{aligned} \Delta A_{0\to 1} &= -\beta^{-1} \ln\left[\frac{Q_1}{Q_0}\right] \\ &= -\beta^{-1} \ln\left[\frac{\prod_i^{3N_A} \sqrt{\frac{2\pi m_1^i}{\beta}} \int \mathrm{d}x^{3N_A} \exp\{-\beta U_1(x^{3N_A})\}}{\prod_i^{3N_A} \sqrt{\frac{2\pi m_0^i}{\beta}} \int \mathrm{d}x^{3N_A} \exp\{-\beta U_0(x^{3N_A})\}}\right] \\ &= -\beta^{-1} \ln\left[\sqrt{\frac{m_1}{m_0}}^3 \langle \exp\{-\beta \Delta U\}\rangle_0\right]. \end{aligned} \tag{A4}$$

In our case, m_1 is m_O, the atomic mass of oxygen, and m_1 is m_S, the atomic mass of sulfur

$$\Delta A_{0\to 1} = -\frac{3}{2}\beta^{-1} \ln \frac{m_S}{m_O} - \beta^{-1} \ln\langle \exp\{-\beta \Delta U\}\rangle_0. \tag{A5}$$

Thus, the results of BAR have to be corrected by

$$\Delta A_{0\to 1}^{\mathrm{mass}} = -\frac{3}{2}\beta^{-1} \ln \frac{m_S}{m_O} = -2.58\ \mathrm{kJ/mol}. \tag{A6}$$

For each individual λ-window, the correction reads

$$\Delta A_{\lambda\to\lambda+\Delta\lambda}^{\mathrm{mass}} = -\frac{3}{2}\beta^{-1} \ln\left[1 + \Delta\lambda \frac{m_S - m_O}{(1-\lambda)m_O + \lambda m_S}\right]. \tag{A7}$$

REFERENCES

[1] *Free Energy Calculations*, edited by C. Chipot and A. Pohorille (Springer-Verlag, Berlin, Heidelberg, 2007).
[2] P. Kollman, Chem. Rev. **93**, 2395 (1993).
[3] C. D. Christ, A. E. Mark, and W. F. van Gunsteren, J. Comput. Chem. **31**, 1569 (2010).
[4] M. A. Olsson and U. Ryde, J. Chem. Theory Comput. **13**, 2245 (2017).
[5] U. Ryde and P. Söderhjelm, Chem. Rev. **116**, 5520 (2016).
[6] O. K. Dudko, G. Hummer, and A. Szabo, Phys. Rev. Lett. **96**, 108101 (2006).
[7] E. Neria, S. Fischer, and M. Karplus, J. Chem. Phys. **105**, 1902 (1996).
[8] D. A. McQuarrie and S. D. Simon, *Molecular Thermodynamics* (University Science Books, Sausalito, CA, 1999).
[9] E. Fermi, J. Pasta, and S. Ulam, Los Alamos Report No. LA-1940, 1955.
[10] B. J. Alder and T. E. Wainwright, J. Chem. Phys. **31**, 459 (1959).
[11] A. Rahman, Phys. Rev. A **136**, A405 (1964).
[12] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).

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

[13]W. K. Hastings, Biometrika **57**, 97 (1970).
[14]B. Brooks and M. Karplus, Proc. Natl. Acad. Sci. U. S. A. **80**, 6571 (1983).
[15]B. Tidor and M. Karplus, J. Mol. Biol. **238**, 405 (1994).
[16]M. Karplus and J. N. Kushick, Macromolecules **14**, 325 (1981).
[17]R. M. Levy, A. R. Srinivasan, and W. K. Olson, Biopolymers **23**, 1099 (1984).
[18]R. W. Zwanzig, J. Chem. Phys. **22**, 1420 (1954).
[19]J. G. Kirkwood, J. Chem. Phys. **3**, 300 (1935).
[20]C. H. Bennett, J. Comput. Phys. **22**, 245 (1976).
[21]J. Gao, K. Kuczera, B. Tidor, and M. Karplus, Science **244**, 1069 (1989).
[22]M. Karplus and G. A. Petsko, Nature **347**, 631 (1990).
[23]J. M. Rickman and R. LeSar, Annu. Rev. Mater. Res. **32**, 195 (2002).
[24]P. A. Bash, U. C. Singh, R. Langridge, and P. A. Kollman, Science **236**, 564 (1987).
[25]P. A. Bash, U. C. Singh, F. K. Brown, R. Langridge, and P. A. Kollman, Science **235**, 574 (1987).
[26]R. Rathore, M. Sumakanth, M. S. Reddy, P. Reddanna, A. A. Rao, M. D. Erion, and M. Reddy, Curr. Pharm. Des. **19**, 4674 (2013).
[27]C. Chipot, Wiley Interdiscip. Rev.: Comput. Mol. Sci. **4**, 71 (2014).
[28]D. L. Mobley and M. K. Gilson, Annu. Rev. Biophys. **46**, 531 (2017).
[29]A. P. Bhati, S. Wan, D. W. Wright, and P. V. Coveney, J. Chem. Theory Comput. **13**, 210 (2017).
[30]M. Cecchini, J. Chem. Theory Comput. **11**, 4011 (2015).
[31]P. H. Berens, D. H. J. Mackay, G. M. White, and K. R. Wilson, J. Chem. Phys. **79**, 2375 (1983).
[32]S.-T. Lin, M. Blanco, and W. A. Goddard, J. Chem. Phys. **119**, 11792 (2003).
[33]S.-T. Lin, P. K. Maiti, and W. A. Goddard, J. Phys. Chem. B **114**, 8191 (2010).
[34]R. A. X. Persson, V. Pattni, A. Singh, S. M. Last, and M. Heyden, J. Chem. Theory Comput. **13**, 4467 (2017).
[35]S. Belsare, V. Pattni, M. Heyden, and T. Head-Gordon, J. Phys. Chem. B **122**, 5300 (2018).
[36]M. Heyden, Wiley Interdiscip. Rev.: Comput. Mol. Sci. **9**, e1390 (2019).
[37]D. Marx and J. Hutter, in *Modern Methods and Algorithms of Quantum Chemistry – Proceedings*, 2nd ed., NIC Series, Vol. 3, edited by J. Grotendorst (NIC-Directors, Juelich, 2000), pp. 329–477.
[38]B. Kirchner, P. J. di Dio, and J. Hutter, Top. Curr. Chem. **307**, 109 (2012).
[39]R. Sure and S. Grimme, J. Comput. Chem. **34**, 1672 (2013).
[40]D. M. Zuckerman and T. B. Woolf, Phys. Rev. Lett. **89**, 180602 (2002).
[41]N. Lu and D. A. Kofke, J. Chem. Phys. **114**, 7303 (2001).
[42]W. L. Jorgensen and C. Ravimohan, J. Chem. Phys. **83**, 3050 (1985).
[43]M. R. Shirts, E. Bair, G. Hooker, and V. S. Pande, Phys. Rev. Lett. **91**, 140601 (2003).
[44]S. Bruckner and S. Boresch, J. Comput. Chem. **32**, 1303 (2011).
[45]M. R. Shirts and V. S. Pande, J. Chem. Phys. **122**, 144107 (2005).
[46]A. de Ruiter, S. Boresch, and C. Oostenbrink, J. Comput. Chem. **34**, 1024 (2013).
[47]G. König and B. R. Brooks, J. Comput.-Aided Mol. Des. **26**, 543 (2012).
[48]G. König, B. R. Brooks, W. Thiel, and D. M. York, Mol. Simul. **44**, 1062 (2018).
[49]F. M. Ytreberg, R. H. Swendsen, and D. M. Zuckerman, J. Chem. Phys. **125**, 184114 (2006).
[50]J. C. B. Dietschreit, L. D. M. Peters, J. Kussmann, and C. Ochsenfeld, J. Phys. Chem. A **123**, 2163 (2019).
[51]P. E. Smith and W. F. van Gunsteren, J. Phys. Chem. **98**, 13735 (1994).
[52]A. E. Mark and W. F. van Gunsteren, J. Mol. Biol. **240**, 167 (1994).
[53]J. Kussmann and C. Ochsenfeld, J. Chem. Phys. **138**, 134114 (2013).
[54]J. Kussmann and C. Ochsenfeld, J. Chem. Theory Comput. **11**, 918 (2015).
[55]J. Kussmann and C. Ochsenfeld, J. Chem. Theory Comput. **13**, 3153 (2017).
[56]S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, J. Chem. Phys. **132**, 154104 (2010).
[57]S. Grimme, S. Ehrlich, and L. Goerigk, J. Comput. Chem. **32**, 1456 (2011).
[58]H. Kruse and S. Grimme, J. Chem. Phys. **136**, 154101 (2012).
[59]L. Verlet, Phys. Rev. **159**, 98 (1967).
[60]W. C. Swope, H. C. Andersen, P. H. Berens, and K. R. Wilson, J. Chem. Phys. **76**, 637 (1982).
[61]G. Bussi, D. Donadio, and M. Parrinello, J. Chem. Phys. **126**, 014101 (2007).
[62]M. R. Reddy, U. C. Singh, and M. D. Erion, J. Am. Chem. Soc. **126**, 6224 (2004).
[63]S. Boresch and M. Karplus, J. Phys. Chem. A **103**, 103 (1999).
[64]H. Hu and W. Yang, J. Chem. Phys. **123**, 041102 (2005).
[65]G. König, P. S. Hudson, S. Boresch, and H. L. Woodcock, J. Chem. Theory Comput. **10**, 1406 (2014).
[66]G. König, B. T. Miller, S. Boresch, X. Wu, and B. R. Brooks, J. Chem. Theory Comput. **8**, 3650 (2012).
[67]G. König, S. Bruckner, and S. Boresch, J. Comput. Chem. **30**, 1712 (2009).

Calculating Free Energies from the Vibrational Density of States Function: Validation and Critical Assessment - Supporting Information

Laurens D. M. Peters,[1,2,a)] Johannes C. B. Dietschreit,[1,2,a)] Jörg Kussmann,[1,2] and Christian Ochsenfeld[1,2,b)]

[1)] *Chair of Theoretical Chemistry, Department of Chemistry, University of Munich (LMU), Butenandtstr. 7, D-81377 München, Germany*

[2)] *Center for Integrated Protein Science (CIPSM) at the Department of Chemistry, University of Munich (LMU), Butenandtstr. 5–13, D-81377 München, Germany*

(Dated: 12 April 2019)

[a)] These two authors contributed equally to this work

[b)] Electronic mail: christian.ochsenfeld@uni-muenchen.de

I. NUMERICAL EXAMPLES

A. Harmonic Potential

TABLE S1. Used values (in atomic units) for k in the simulations of the initial (index 0) and final (index 1) state.

Wavenumbers	k_0	k_1
$100 \rightarrow 500$	3.78e-04	9.46e-03
$500 \rightarrow 1000$	9.46e-03	3.78e-02
$1000 \rightarrow 2000$	3.78e-02	1.51e-01

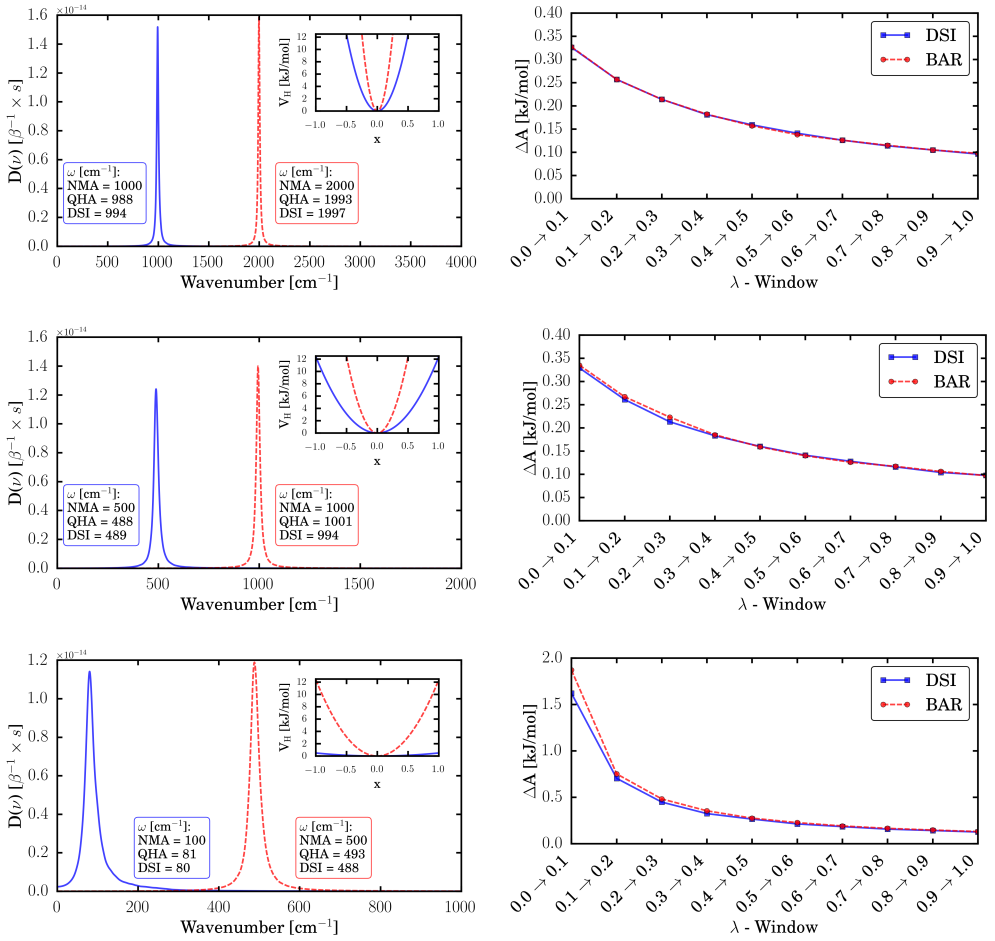


FIG. S1. (left) Harmonic potentials and density of states plots of the simulations. In addition, their maxima (DSI) and the frequencies calculated with NMA and QHA are listed. (right) $\langle \Delta A \rangle$s of the individual λ-windows of the corresponding simulations. Please note that the density of states function ($D(\nu)$) is given in $\beta^{-1} \times s$; We have omitted the factor β in eq. (22).

B. Anharmonic Potential

TABLE S2. Used values (in atomic units) for D_E and a in the simulations of the initial (index 0) and final (index 1) state.

Wavenumbers	D_{E0}	D_{E1}	a_0	a_1
$100 \rightarrow 500$	1.55e-02	3.10e-02	1.11e-01	3.91e-01
$500 \rightarrow 1000$	1.55e-02	3.10e-02	5.53e-01	7.82e-01
$1000 \rightarrow 2000$	1.55e-02	3.10e-02	1.11e+00	1.56e+00

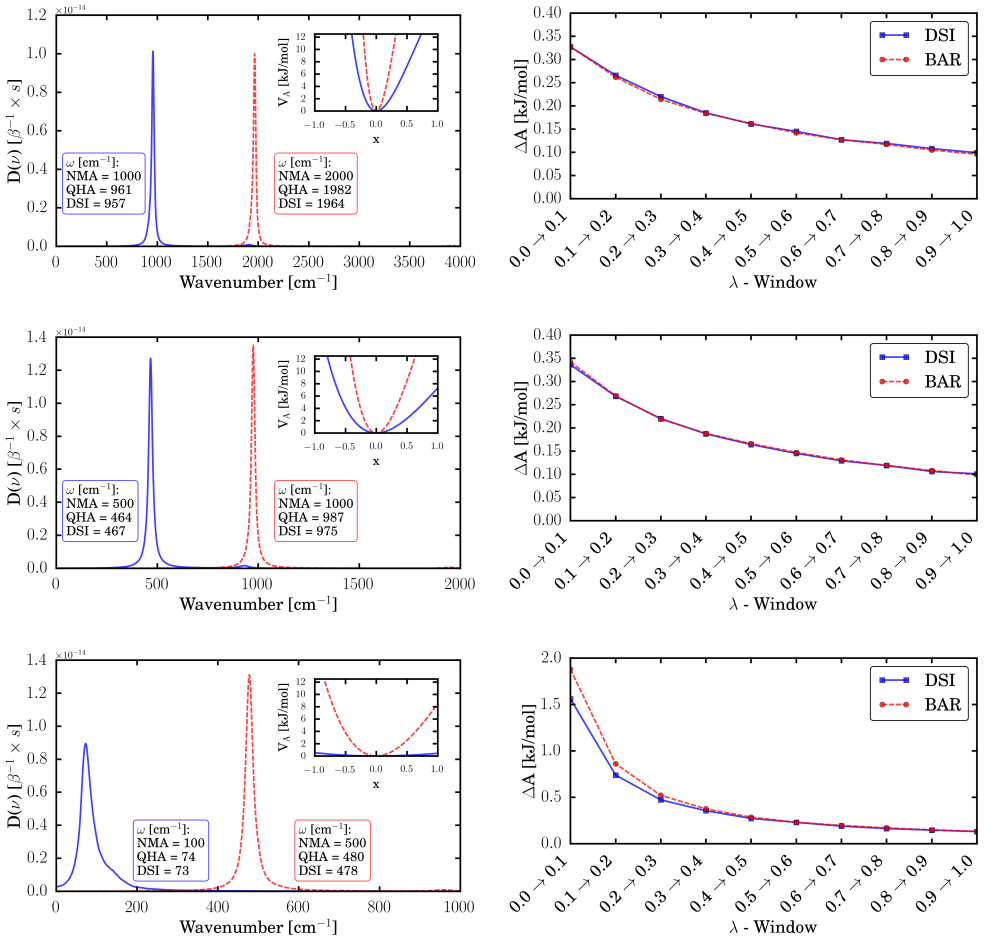


FIG. S2. (left) Anharmonic potentials and density of states plots of the simulations. In addition, their maxima (DSI) and the frequencies calculated with NMA and QHA are listed. (right) $\langle \Delta A \rangle$s of the individual λ-windows of the corresponding simulations. Please note that the density of states function ($D(\nu)$) is given in $\beta^{-1} \times s$; We have omitted the factor β in eq. (22).

C. Double Well Potential

TABLE S3. Used values (in atomic units) for b in the simulations of the initial (index 0) and final (index 1) state.

Wavenumbers	b_0	b_1
$100 \rightarrow 1500$	3.78e-02	8.51e-02
$1500 \rightarrow 2000$	8.51e-02	1.51e-01
$2000 \rightarrow 2500$	1.51e-01	2.37e-01

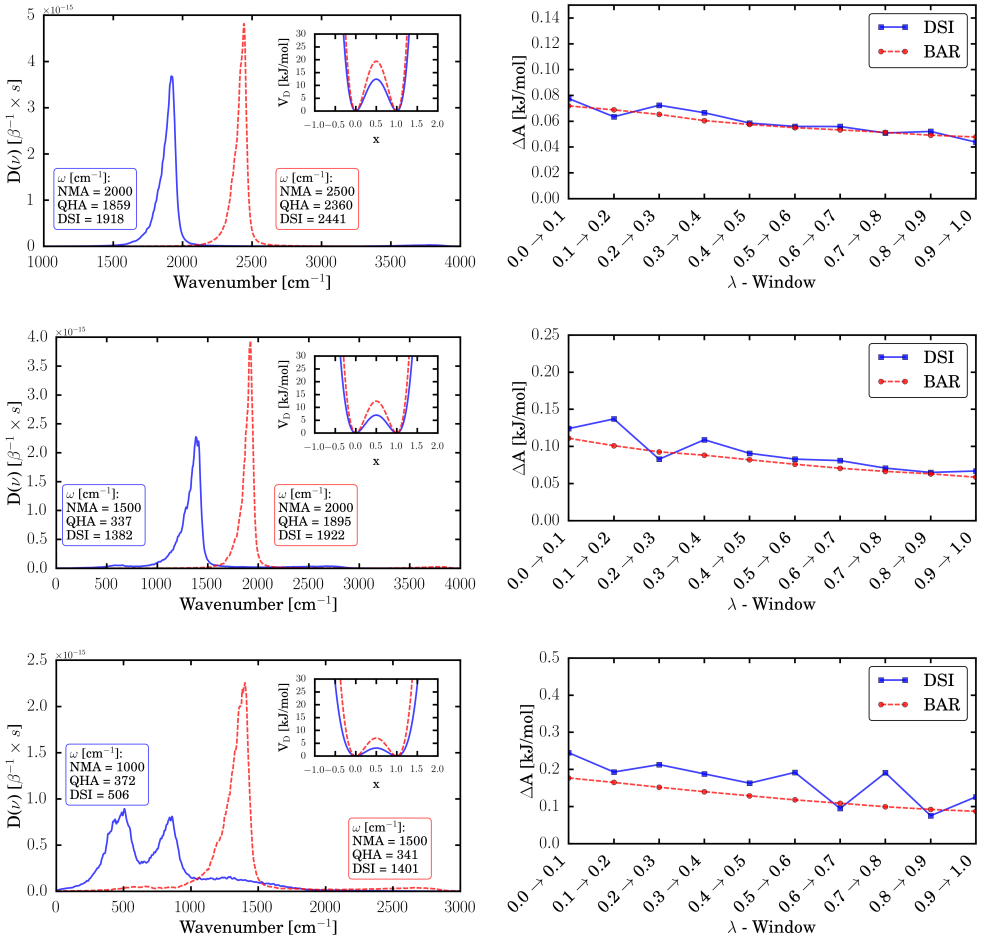


FIG. S3. (left) Double well potentials and density of states plots of the simulations. In addition, their maxima (DSI) and the frequencies calculated with NMA and QHA are listed. (right) $\langle \Delta A \rangle$s of the individual λ-windows of the corresponding simulations. Please note that the density of states function ($D(\nu)$) is given in $\beta^{-1} \times s$; We have omitted the factor β in eq. (22).

II. IONIZATION ENERGY OF AMMONIA

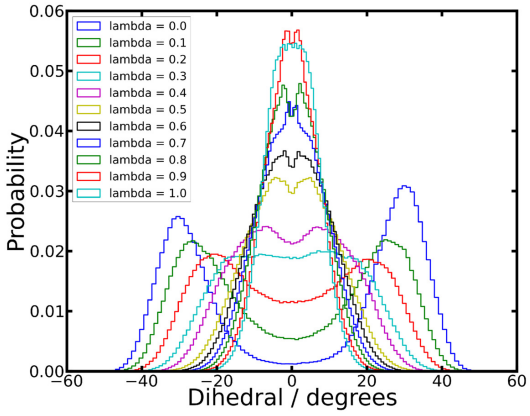


FIG. S4. Normalized distribution of the improper dihedral of ammonia for each λ-value over 100 bins. The histograms include the data of all five simulations per λ-value.

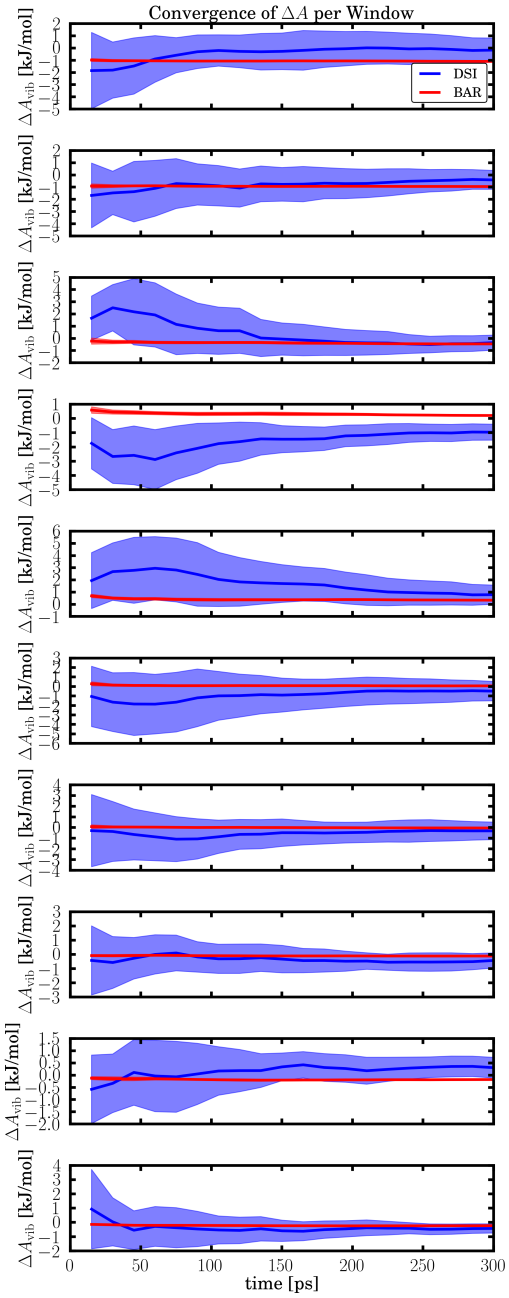


FIG. S5. Convergence of the mean vibrational free energy change ($\langle \Delta A \rangle$, solid line) and the standard deviation of different trajectories (lighter area) using BAR and DSI for each λ-window. The top panel shows the changes for $0.0 \rightarrow 0.1$, the second panel from the top pertains to $0.1 \rightarrow 0.2$, and so forth.

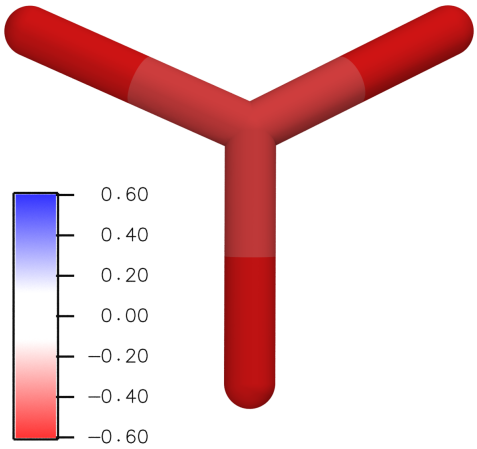


FIG. S6. Vibrational free energy change ($\langle \Delta A_{\mathrm{vib}}^{\mathrm{DSI}} \rangle$ in kJ/mol) of each atom during the ionization of ammonia. The hydrogen atoms loose 0.59 kJ/mol and the nitrogen atom 0.46 kJ/mol.

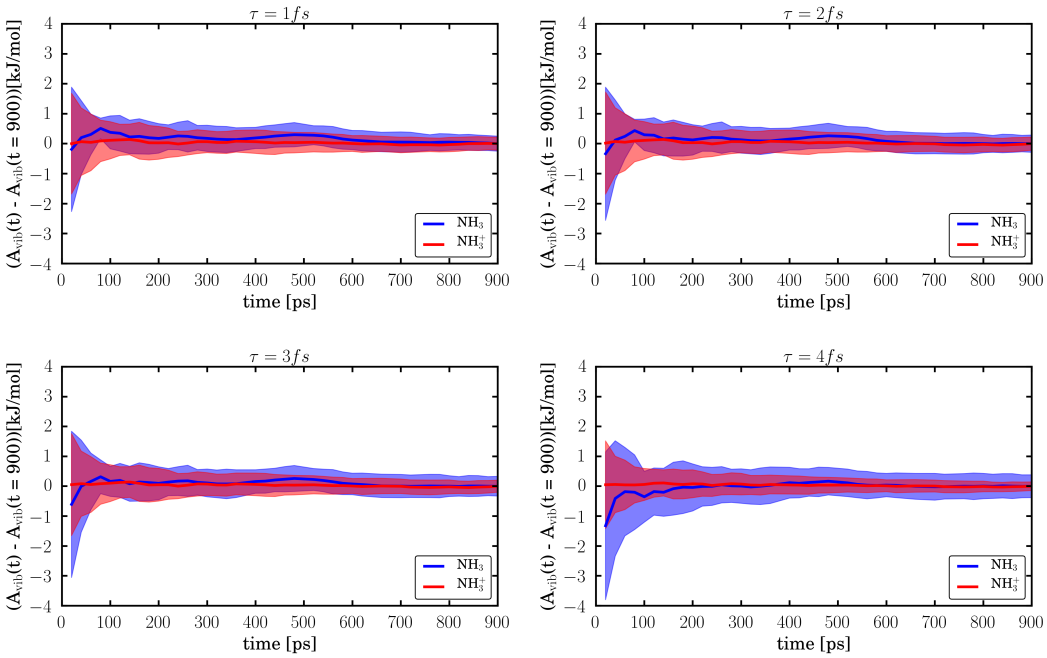


FIG. S7. Convergence of the mean vibrational free energy ($\langle A_{\mathrm{vib}} \rangle$, solid line) and the standard deviation of ten different trajectories (lighter area) of NH_3 and NH_3^+ using DSI and different intervals (τ) between the sampling of the nuclear velocities. The mean value at t = 900 ps is set to zero. The mean vibrational free energy converges after a sampling of $\approx$ 200 ps. Its standard deviation decreases constantly with increasing simulation time. The values of NH_3 show a slower convergence and a larger standard deviation, due to the higher amount of low-frequency modes.

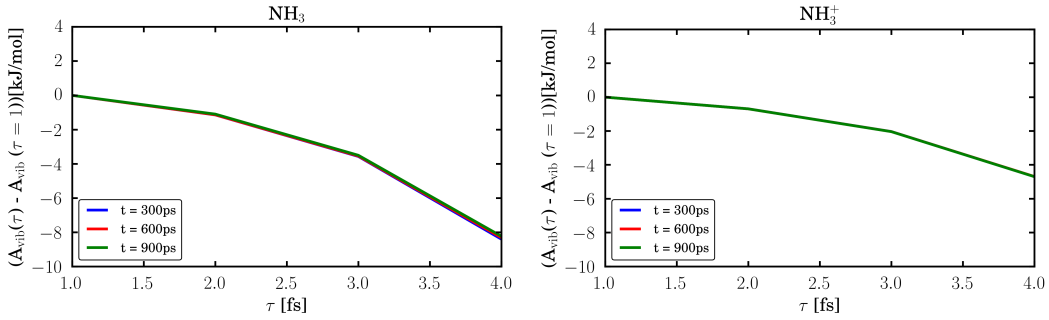


FIG. S8. Change of the mean vibrational free energies with the length of the interval between the sampling of the nuclear velocities (τ) using DSI. Ten independent trajectories of NH_3 and NH_3^+ and three different simulation times (t) are considered. The mean values at $\tau = 1\ fs$ are set to zero. The sampling rate has a larger impact on the result than the simulation length. An increasing τ decreases the intensity of the N-H bond-stretching modes, leading to changes in the vibrational free energy. The reason for this is that the sampling of these high-frequency modes becomes worse. Applying $\tau > 4\ fs$ the bond vibrations do not appear in the density of states spectrum. The effect of τ on NH_3 is larger, as the frequency of its bond vibrations is higher.

III. MUTATION FROM SERINE TO CYSTEINE

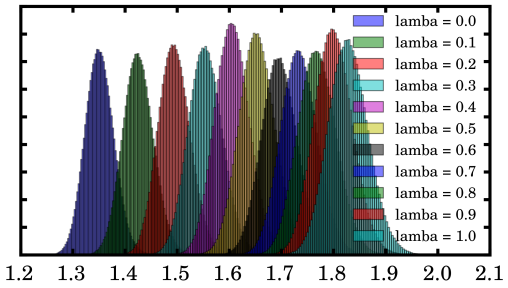


FIG. S9. Histograms of the C-O/S bond length for each λ-window. The bond length histograms are closely connected to the distribution of ΔU in each window.

3.2 Publication II: Identifying Free Energy Hot-Spots in Molecular Transformations

Johannes C. B. Dietschreit, Laurens D. M. Peters,
Jörg Kussmann, and Christian Ochsenfeld
"Identifying Free Energy Hot-Spots in Molecular Transformations"
J. Phys. Chem. A **2019**, *123*, 2163-2170

Abstract: The free energy is one of the central quantities in material and natural sciences. While being well-established, e.g., in drug design or catalyst optimization, computational methods lack a straightforward way to gain deeper insights into the calculated free energy, and thus the underlying chemical or physical processes. Here, we present a generally applicable, spectrum-based ansatz that tackles this shortcoming by identifying contributions from specific atoms or groups to the vibrational free energy. We illustrate this in studies of the bromodomain-inhibitor binding and the anomeric effect in glucose providing quantitative evidence in line with chemical intuition in both cases. For the latter example we also report an experimental infrared spectrum and find excellent agreement with our simulated spectra.

THE JOURNAL OF PHYSICAL CHEMISTRY A
Article
Cite This: J. Phys. Chem. A 2019, 123, 2163−2170
pubs.acs.org/JPCA

Identifying Free Energy Hot-Spots in Molecular Transformations

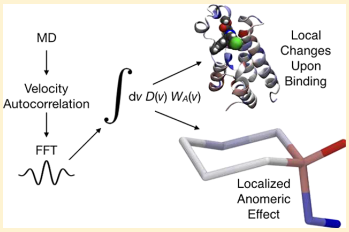
Johannes C. B. Dietschreit,[†,‡,§] Laurens D. M. Peters,[†,‡,§] Jörg Kussmann,[†,‡] and Christian Ochsenfeld*[,†,‡]

[†]Chair of Theoretical Chemistry, Department of Chemistry, University of Munich (LMU), Butenandtstr. 7, D-81377 München, Germany

[‡]Center for Integrated Protein Science (CIPSM) at the Department of Chemistry, University of Munich (LMU), Butenandtstr. 5−13, D-81377 München, Germany

Supporting Information

ABSTRACT: The free energy is one of the central quantities in material and natural sciences. While being well-established, e.g., in drug design or catalyst optimization, computational methods lack a straightforward way to gain deeper insights into the calculated free energy, and thus the underlying chemical or physical processes. Here, we present a generally applicable, spectrum-based ansatz that tackles this shortcoming by identifying contributions from specific atoms or groups to the vibrational free energy. We illustrate this in studies of the bromodomain-inhibitor binding and the anomeric effect in glucose providing quantitative evidence in line with chemical intuition in both cases. For the latter example we also report an experimental infrared spectrum and find excellent agreement with our simulated spectra.



INTRODUCTION

The free energy is the driving force behind every chemical reaction. It determines, for example, the rate of an enzymatic reaction or the scope of products formed during a catalytic reaction. The prediction of free energies is, therefore, a key challenge in modern quantum chemistry.[1,2] For small, unimolecular systems this is usually done via a frequency analysis of the molecule using quantum mechanics (QM) calculations. It is assumed that, in the vicinity of the minimum energy geometry, all vibrations can be described as independent harmonic oscillators (harmonic approximation). For larger or multimolecular systems, this approach is not feasible, as the harmonic approximation is not valid anymore (due to the increasing number of anharmonic modes) and the potential energy surface features an enormous number of local minima, which have to be considered. In these cases, one focuses on free energy differences, which can be computed without having to know the absolute energy of both states, and determines the free energy from sampled energies along Monte Carlo[3,4] or molecular dynamics (MD)[5−7] simulations applying, e.g., exponential averaging theory[8] or Bennett's acceptance ratio method.[9]

While the mentioned methods have been used extensively in different fields,[10−13] the interpretation of their results is in most cases not straightforward. The reason for this is that it is not possible to separate the total free energy change into contributions from different atoms or residues[14] and, therefore, to understand the underlying effects (e.g., bond weakening, sterical clashes, new noncovalent interactions) causing the free energy to change. Applying the conventional energy-based methods,[8,9] some approximate fragmentation is possible for simple force fields;[12,15] however, this is not possible when nonadditive force fields (like the emerging polarizable force fields[16−18]), QM calculations, or combined quantum mechanics/molecular mechanics (QM/MM) are used.

In this work, we use and present a method that calculates the vibrational part of the free energy from the vibrational density of states function, which itself was the topic of experimental[19] and theoretical studies.[20,21] This approach has originally been introduced by Berens et al.[22] to estimate quantum corrections to thermodynamic properties. It has been used occasionally to compute absolute entropies,[23] solvation effects such as entropy,[24] or helped identifying different water species around a protein in solution[25] by employing the additional assumptions of the two-phase model.[26] The calculation of free energy changes in discrete volume units (so-called "voxels") by Heyden is also based on this approach.[27]

The applicability of the method of Berens et al.[22] to free energy calculations has been determined in a different study of ours,[28] where a more detailed derivation, validation, and analysis of the method can be found. Here, we will focus entirely on its capability of calculating atom- or residue-wise contributions to the vibrational free energy and how these free energy hot-spots can help to understand and interpret free energy changes during molecular transformations. We start with a brief summary of the density of states integration method (DSI)[22,28] in Section 2. There, we will also discuss shortcomings of the method and how they affect the

Received: December 21, 2018
Revised: February 14, 2019
Published: March 1, 2019

DOI: 10.1021/acs.jpca.8b12309
J. Phys. Chem. A 2019, 123, 2163−2170

The Journal of Physical Chemistry A Article

applicability and the interpretation of the results of our approach. In Section 3 we list computational (and experimental) details. In Section 4 we apply our method to two prototypical, illustrative examples: (1) The binding of an inhibitor to a bromodomain-containing protein and (2) the visualization of the anomeric effect in glucose. An outlook is given in Section 5.

THEORY

Density of States Integration Method. We extract the free energy from the velocities ($\mathbf{v}_j$) of each atom j during a molecular dynamics simulation. This is done by calculating the density of states function ($D(\nu)$) as the Fourier transform of the velocity autocorrelation function

$$D(\nu) = 2\beta \sum_{j=1}^{N} m_j \int dt \langle \mathbf{v}_j(\tau)\mathbf{v}_j(t+\tau)\rangle_\tau e^{-i2\pi\nu t} \tag{1}$$

β is equal to $1/(k_B T)$ with k_B being the Boltzmann constant and T the absolute temperature. m_j is the mass of atom j, and N is the total number of nuclei. When neglecting contributions from translation and rotation, the free energy (A) can be calculated from a weighted integral over the frequency (ν)[22,28]

$$A = E + A_{vib} = E + \beta^{-1} \int_0^\infty d\nu D(\nu) \ln[\beta h\nu] \tag{2}$$

with E being the potential energy at the global minimum energy geometry and h the Planck constant. The vibrational part of the free energy (A_{vib}) is thus calculated as a sum of an arbitrarily large number of harmonic oscillators weighted by $D(\nu)$.

Equations 1 and 2 indicate that A_{vib} can be split into contributions from the individual nuclei or residues, as $D(\nu)$ is calculated as a sum over all atoms. In order to obtain atom- or residue-resolved free energies, we recast eq 1 to

$$\begin{aligned} D(\nu) &= 2\beta \sum_{j=1}^{N} m_j \int dt \langle \mathbf{v}_j(\tau)\mathbf{v}_j(t+\tau)\rangle_\tau e^{-i2\pi\nu t} \\ &= \sum_{j=1}^{N} 2\beta m_j \int dt \langle \mathbf{v}_j(\tau)\mathbf{v}_j(t+\tau)\rangle_\tau e^{-i2\pi\nu t} \\ &= \sum_{j=1}^{N} D_j(\nu) \\ &= \sum_{i}^{N_{regions}} \sum_{j}^{\{N\}_i} D_j(\nu) \end{aligned} \tag{3}$$

$N_{regions}$ is the number of regions in which we split the total system and $\{N\}_i$ is the set of atoms that belong to the region i. The regions can be chosen completely freely ranging from the entire system over residues to individual atoms. This helps us rewrite eq 2 to

$$A = E + \sum_{i}^{N_{regions}} A_{vib}(i) \tag{4}$$

$A_{vib}(i)$ is the vibrational free energy localized in region i. If we consider free energy changes

$$\Delta A = \Delta E + \sum_{i}^{N_{regions}} \Delta A_{vib}(i) \tag{5}$$

$\Delta A_{vib}(i)$ indicates a change in the potential energy surface in region i. Please note that the fragmentation of ΔA_{vib} directly evolves from its calculation from $D(\nu)$ and requires no additional approximations or assumptions.

Interpreting Free Energy Hot-Spots. Two shortcomings have to be considered when interpreting $\Delta A_{vib}(i)$ for different atoms or residues i. The first one is that we are investigating only the vibrational part of the free energy ($\Delta A_{vib}(i)$ instead of $\Delta A(i)$) without $\Delta E(i)$ and contributions from rotations and translations. The latter are neglected, as the overall rotation and translation of the system are removed, because keeping them can lead to unwanted artifacts. Therefore, the sum over all hot-spots yields ΔA_{vib} and not the full free energy difference (ΔA). A comparison to other free energy methods is, thus, only meaningful when adding ΔE, which has been done for small systems in ref 28 and is not possible for larger systems (as the exact determination of ΔE becomes impossible). However, A_{vib} identifies regions where the potential energy surface (and thus A) is changing and is, therefore, an excellent tool to find and quantify free energy hot-spots, as shown below.

The second shortcoming is that the Density of States Integration (DSI) uses the harmonic approximation. It was shown in ref 28 that DSI can (in contrast to other free energy methods based on vibrational frequencies) describe the anharmonic behavior of vibrations, as it considers the system as a linear combination of a (nearly) infinite number of harmonic oscillators. Free energy changes arising from vibrations involving movements over local maxima in the potential energy surface and very slow modes (such as rotations of entire protein domains) can only be described qualitatively but not quantitatively, leading to errors in the calculation of A_{vib}.[28] However, these errors partly cancel out, because we investigate free energy differences. Moreover, slow modes are normally delocalized over large parts of the systems and thus do not substantially affect the free energy hot-spots and their interpretation, as we focus mainly on local changes.

METHODS

Classical Mechanical Simulations. The crystal structures of the apo-bromodomain (PDB 5O38)[29] and the inhibitor-domain complex (PDB 5O3B)[29] were used as starting structures. All molecules that were not protein, inhibitor, or water were removed. Antechamber, part of the AmberTools 16,[30] was used to parametrize the inhibitor. The force field ff14SB[31] was used for the simulations. The proteins were solvated in a rectangular box with 10 Å of TIP3P[32] water, and neutralized with 2 chlorine ions. The simulation engine NAMD[33] was used. The energy of the system was minimized: For the first 10,000 steps only the water molecules and for the next 10,000 steps the full system. The system was heated over 30 ps to 300 K. In the following it was equilibrated for 200 ps, and then a production run of 1 ns was carried out. The time step was 0.5 fs, as no constraints such as SHAKE[34] or RATTLE[35] were imposed on the system during the production runs. Nonbonded interactions were evaluated at every step. Periodic electrostatic interactions were computed with the particle mesh Ewald summation method, with a sixth order interpolation. We used a cutoff radius of 12 Å and a switching

DOI: 10.1021/acs.jpca.8b12309
J. Phys. Chem. A 2019, 123, 2163−2170

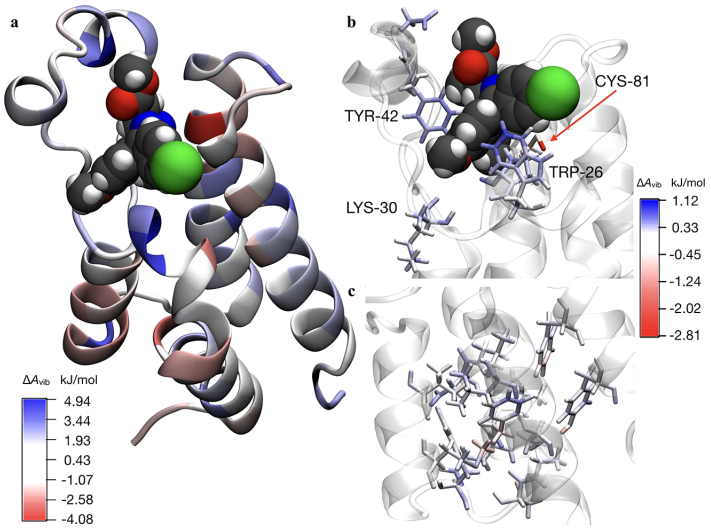


Figure 1. (a) Changes of the vibrational free energy (ΔA_{vib}) within the bromodomain upon binding the inhibitor. The residues are colored according to the changes in A_{vib} going from the apo- to the complexed-form per residue. Blue residues indicate a gain ($\Delta A_{vib} > 0$, less movement), and red residues indicate a loss in vibrational free energy ($\Delta A_{vib} < 0$, more freedom). The inhibitor is shown with van-der-Waals-spheres colored according to atom types. (b) Interaction between the residues and the inhibitor. TRP-26 (van-der-Waals interaction), TYR-42 (hydrogen bond), and CYS-81 (two hydrogen bonds) are highlighted. (c) Interactions within the helical part of the domain. Note that the color scales differ between part a and parts b and c of the figure.

function that smoothly switches off interaction between 10 and 12 Å. A Verlet nearest neighbor list with a radius of 13.5 Å was used. The temperature was controlled with the Berendsen rescaling algorithm.[36] Translation and rotation of the protein were removed from the velocities after the simulation. *MDAnalysis*[37,38] was used to extract and process the velocities. The convergence of the free energy difference is shown for residue TRP-26 in Figure S6.

Quantum Mechanical Simulations. The quantum chemistry package *FermiONs++*[39−41] developed in our group was used for the *ab initio* Born−Oppenheimer molecular dynamics simulations. We used the HF-3c[42] method that includes dispersion (DFTD3 v3.1)[43,44] and counterpoise corrections (gCP v2.02).[45] The Velocity Verlet algorithm[46,47] and a stochastic rescaling thermostat[48] were applied. The structures of ***α*-Glu**, ***β*-Glu**, ***α*-HCT**, and ***β*-HCT** were minimized at the same level of theory before the calculations. The initial atom velocities were drawn at random from a Maxwell−Boltzmann distribution at 298.15 K. The time step is 0.5 fs, and a ninth order extended Lagrangian scheme[49] was used to improve the SCF convergence. The system was equilibrated for 5 ps. The production runs were 200 ps long, and 20 independent simulations (different starting velocity vectors and pseudorandom numbers in the thermostat) were conducted for each molecule. Translation and rotation of the molecule were removed at every step of the simulation. As starting points, we used two different minima, both obtained by energy minimizations. Additional conformers or a subsequent weighting of the single simulations were not required as the thermal energy of the molecule and the simulation time were sufficient to explore the conformational space. The sampling was monitored by the convergence of the mean free energy difference; see Figure S5.

Free Energy Calculations. Vibrational free energies (A_{vib}) are calculated from the sampled nuclear velocities applying eqs 1 and 2. All atomic spectra were rescaled such that every atom receives the same fraction of the total amount of degrees of freedom.

Infrared Spectra. The experimental spectrum of D(+)-glucose 1-hydrate (ITW Reagents, > 99%) has been measured in this work as an average of 16 scans with 1 cm^{-1} resolution using a Thermo Fischer Nicolet 6700 FT-IR apparatus.

RESULTS AND DISCUSSION

Inhibitor Bound to the Bromodomain—An MM Application. As a first demonstration of the presented approach, we investigate the change of A_{vib} during the binding of a bromodomain-containing protein to an inhibitor. Proteins of the bromo- and extra-terminal domain (**BET**) family are involved in the recognition of acetylated lysine residues and play an important role in epigenetic communication.[50] Very recently, potent mutant-selective inhibitors for **BET** have been developed,[29] which are meant as a tool for future *in vivo* studies. Upon binding to the inhibitor, the potential energy surface of **BET** is modified leading to conformational changes in the protein which one would generally assume causes the binding site to become tighter. Here, we want to stress that we are focusing on calculating the *changes* of A_{vib} upon binding and do not attempt to compute the binding free energy, for which energy-based methods such as the Bennett's Acceptance Ration method[9] are more suitable. We expect though that the atoms highlighted by our method are those which are the main contributors to the free energy of binding.

We used the cocrystal structure of **9-ME-1** and **BET** as well as the apo-crystal structure (PDB 5O3C and 5O38[29]) as a starting point to investigate the effect of inhibitor binding to the bromodomain motif. We conducted two independent

DOI: 10.1021/acs.jpca.8b12309
J. Phys. Chem. A 2019, 123, 2163−2170

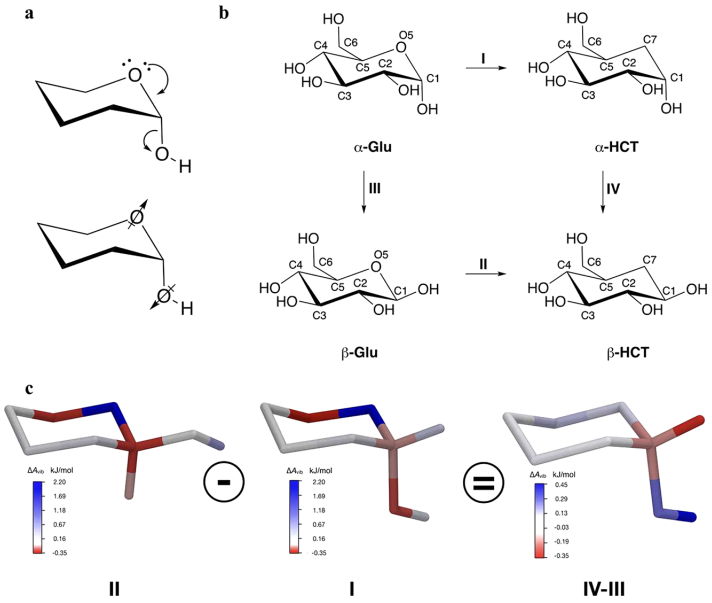


Figure 2. (a) (top) Hyperconjugation and (bottom) dipole interaction discussed as origins of the anomeric effect in glucose.[52] (b) Structures, abbreviations, and atom labels of the investigated molecules. The possible transformations **I** to **IV** are aranged in a thermodynamic cycle. (c) Change in the vibrational free energy per atom from ***β*-Glu** → ***β*-HCT** (transformation **II**), ***α*-Glu** → ***α*-HCT** (transformation **I**), and their difference (**II-I** = **IV-III**). The latter is equivalent to the change of the free energy during the appearance of the anomeric effect reflecting the bond strengthening of the O5−C1 bond and the bond weakening of the C1−O1 and C5−O5 bond.

classical molecular dynamics simulations in a water box and computed the difference in vibrational free energy (ΔA_{vib}) per residue and per atom, where the residue-wise vibrational free energy is just the sum over the corresponding atoms. The changes in the vibrational free energy are shown in Figure 1. Blue indicates a stiffening of the residue ($\Delta A_{vib} > 0$, shift toward vibrations of higher frequency) upon the binding of the inhibitor, whereas red ($\Delta A_{vib} < 0$, shift toward vibrations of lower frequency) means that the residue can move more freely.

The overall change in ΔA_{vib} is positive, already indicating a stiffening of the motif due to the interaction with the inhibitor. Many of those stiffer residues can be found in the binding pocket (see Figure 1a+b), where their side chains show distinct interactions with the inhibitor. Examples are TRP-26, which interacts with the inhibitor via van-der-Waals interactions, and TYR-42, which forms a hydrogen bond. Interactions between the inhibitor and the backbone are also occurring. The peptide group of LYS-30, for example, communicates with **9-ME-1** via a water molecule. The only residue in the binding pocket, which becomes more flexible, is CYS-81; its SH-group switches between two hydrogen bond acceptors (one at **BET** and one at the inhibitor). Both hydrogen bonds are not formed in the apoprotein, where CYS-81 is constantly bound to a residue which is inaccessible in the presence of the inhibitor. The analysis, however, also shows that A_{vib} does not only change at the binding site, as parts of the α-helices are colored as well (see Figure 1c). Reasons for these significantly smaller contributions are subtle changes in the arrangement of the helices.

In order to discuss if the values of the free energy hot-spots ($\Delta A_{vib}(i)$) can be interpreted quantitatively (not only qualitatively), one has to consider the individual vibrations contributing to the free energy hot-spots. The changes of the free energy in the binding pocket and at the helices are dominated by local changes of stretching or bending vibrations, arising from the interactions with the inhibitor or the resulting changes in the arrangement of the helices. These (rather) high-frequency modes should be described quantitatively correct with our method, as already discussed in Section 2 and ref 28. CYS-81 might be an exception as the switching between two hydrogen bond acceptors features a low-frequency mode over a local maximum. It is, however, an excellent example how local effects change ΔA_{vib} dramatically and how this is detected by our method. Of course, we cannot exclude that our method neglects possible low-frequency modes, which affect large parts of the system. Yet, we were not able to find such vibrations in RMSD-plots.

In previous experimental and theoretical studies,[19,51] a general mode softening (increase of the density of states function for very small wave numbers) was reported for similar protein-inhibitor systems. These new low-frequency modes were identified as linear combinations of the rotational or translational modes of the inhibitor and modes of the protein. In our present study, this effect is not visible. The main reason for this is that all our simulations were carried out including explicit solvent molecules, while previous studies involved gas-phase simulations and measurements of dried samples. As a consequence, we describe the more realistic system and the replacement of solvent by the inhibitor in the binding pocket and not the mere binding event. Since the solvent molecules located in the binding pocket also couple to the protein, no mode softening during its replacement by the inhibitor is observed.

DOI: 10.1021/acs.jpca.8b12309
J. Phys. Chem. A 2019, 123, 2163−2170

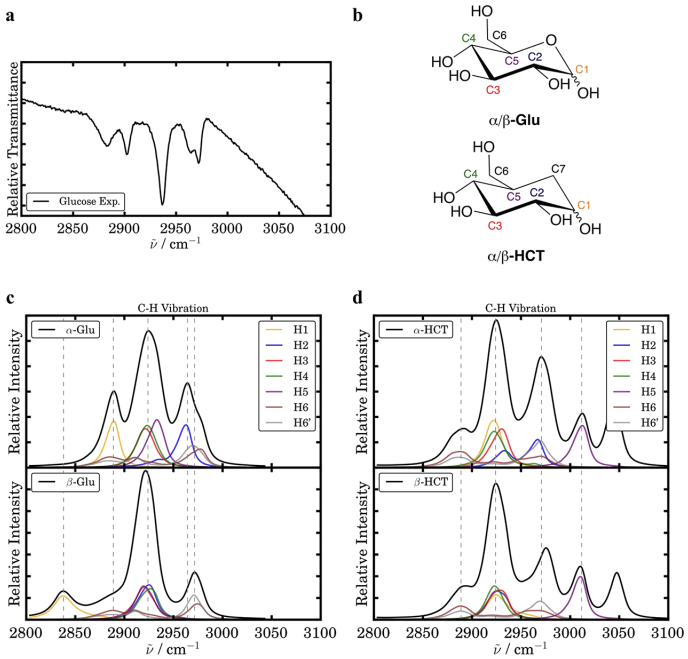


Figure 3. (a) Excerpt of the experimental IR-spectrum of crystalline glucose (monohydrate and mixture of the α- and β-anomer) showing the C−H stretching vibrations for comparison to the simulated spectra below. (b) Labels of carbon atoms in ***α*-Glu**, ***β*-Glu**, ***α*-HCT**, and ***β*-HCT**. (c) Calculated spectra ($D(\tilde{\nu})$) of (top) ***α*-Glu** and (bottom) ***β*-Glu** showing the C−H stretching vibrations of the entire molecule (black) and the contributions from the different C−H bonds (color). The splitting of molecular peaks enables a detailed inspection of the surroundings of the individual atoms. For comparison to the experiment, the frequencies of the simulated spectra have been scaled by a factor of 0.82 (similar to the reported 0.81 in our previous work[55]). (d) Same analysis for (top) ***α*-HCT** and (bottom) ***β*-HCT**. The peaks of H7 and H7′ (around 3050 cm^{-1}) are not shown as they cannot be compared to glucose.

In summary, our approach has revealed the formation of a tight inhibitor−protein complex, which also affects parts of the helices of the protein. All interactions reported here were found simply through the application of the presented method, and no complex analysis of bond distances or dihedral angle distributions and, thus, no *a priori* knowledge about the binding process was necessary to identify any of them. In addition, the power spectra ($D(\nu)$) of the residues or atoms were used to interpret the free energy hot-spots.

Anomeric Effect—A QM Application. In the first example we focused on intermolecular interactions, but our method can also visualize changes of covalent bonds. Therefore, we use, as a second example, *ab initio* molecular dynamics instead of force field calculations to investigate the anomeric effect. The anomeric effect appears in heterocycles based on cyclohexane and leads to a stabilization of the axial position of heteroatomic substituents adjacent to the heteroatom within the six-membered ring.[52] We investigate this effect by looking at one famous representative, namely glucose. Here, one encounters an unexpected stabilization of α-glucose (***α*-Glu**) with respect to β-glucose (***β*-Glu**). The origin of the anomeric effect has been under discussion for a long time including experimental[53] as well as theoretical[54] contributions (see Figure 2a). The two debated causes are hyperconjugation and dipole interactions, both stereoelectronic effects. In this work, we restrict ourselves to the visualization of the effect and highlight the involved atoms, as our method does not allow for a distinction between the two models.

The sole comparison of ***α*-Glu** and ***β*-Glu** would not only incorporate the anomeric but also other effects, for example, changes of the hydrogen bonds or the 1,3-diaxial interactions. To isolate and visualize the anomeric effect, we have simulated ***α*-Glu**, ***β*-Glu**, and their two analogues with the ring oxygen being replaced by a CH_2 moiety (***α*-HCT** and ***β*-HCT**, the abbreviation **HCT** stands for the IUPAC name 5-**H**ydroxy-methyl-**c**yclohexane-(1,2,3,4)-**t**etrole) at the HF-3c[42] level of theory. Please note that the anomeric effect is only present in ***α*-Glu**. It can, therefore, be investigated by analyzing the difference between the transformations ***α*-Glu** → α-**HCT** (**I**) and ***β*-Glu** → β-**HCT** (**II**) or between the transformations ***α*-Glu** → β-**Glu** (**III**) and ***α*-HCT** → β-**HCT** (**IV**), as the other effects cancel out. For structures and atomic labels see Figure 2b.

In order to visualize the anomeric effect, we show the vibrational free energy differences (ΔA_{vib}) of the transformations **I** and **II** for selected atoms in Figure 2c (for the corresponding vibrational spectra see Figure S1). Their difference, which can be interpreted as the appearance of the anomeric effect, is also shown. In both transformations (**I** and **II**), the centers near the mutation site C7/O5 and (in the case of **I**) O1 contribute to ΔA_{vib}, whereas C2−4 are not affected by the mutation. Their difference (**II** − **I**) reveals that ΔA_{vib} of the anomeric effect is mainly localized at C1, O1, and their hydrogen atoms as well as at C5 and O5. It is in good

DOI: 10.1021/acs.jpca.8b12309
J. Phys. Chem. A 2019, 123, 2163−2170

agreement with the bond strengthening of the O5−C1 bond (C1 is red) and the bond weakening of the C1−O1 bond (O1 is blue), when the anomeric effect appears. Additionally, the C5−O5 bond is slightly weakened (C5 is blue). The comparably small contribution at O5 is the result of two counteractive effects, the simultaneous strengthening and weakening of its bonds to C1 and C5, respectively. All effects are also visible in the distributions of bond lengths during the simulations (see Figure S2) and can be interpreted quantitatively, as they feature solely changes in (rather) high-frequency modes.

The C−H stretching vibrations of the systems (Figure 3) can also be used to prove the anomeric effect, as they are very good sensors for changes in the surrounding chemical environment. The superposition spectra of **α-Glu** and **β-Glu** (black lines in Figure 3c) are in good agreement with the experimental infrared spectrum measured in the present work (Figure 3a) featuring both six peaks. Comparing **α-Glu** and **β-Glu** (Figure 3c), we can identify two red-shifts (C1−H1 and C5−H5), which do not appear in the **HCT**-spectra (Figure 3d). They can, therefore, be assigned to the anomeric effect corroborating the previous result that the anomeric effect affects the vibrations of C1, C5, and O1 as well as the connected hydrogen atoms and not C2−C4 and C6.

Again, our method has discovered all atoms involved in the anomeric effect, verifying the common picture of this complex stereoelectronic effect (see Figure 2a) without any prior knowledge or assumption. A detailed inspection of the spectra ($D(\nu)$) offers even more insights in the vibrational behavior of the investigated molecules, as shown for this specific case.

CONCLUSION

Overall, the use of vibrational spectra calculated from nuclear velocities, can lead to new and valuable insights into molecular transformations. As the two examples have shown, we are able to localize and, therefore, explain free energy changes. The calculation is straightforward and does not require any *a priori* knowledge about the system before the actual evaluation. We have also shown that it is applicable to any level of theory for the molecular dynamics simulations, ranging from force-field to full quantum-mechanical calculations. Our results are in absolute agreement with chemical intuition for which our method provides a solid and generally valid fundament. Although the central quantity of the approach is the vibrational free energy (ΔA_{vib}) and not the total free energy (ΔA), our method allows for a quantification of effects, especially when rather high-frequency vibrations are involved and when one compares different ΔA_{vib}'s of, e.g., different inhibitors or different substituents. We suggest this ansatz to be used, e.g, in drug or catalyst design, in addition to the calculation of energies and the investigation of structural parameters for gaining the complete picture of the problem at hand.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jpca.8b12309.

The Supporting Information includes details about our image processing, atomic VDOS spectra of all heavy atoms in glucose and HCT, distributions of important bond lengths and dihedral angles in glucose and HCT, comparison of experimental and simulated IR-spectra, a selected region from the VDOS spectrum featuring the anomeric effect, convergence plots for glucose and HCT, and a convergence plot for a selected residue of the bromodomain (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: christian.ochsenfeld@uni-muenchen.de.

ORCID

Christian Ochsenfeld: 0000-0002-4189-6558

Author Contributions

§J.C.B.D. and L.D.M.P. contributed equally to this work

Notes

The authors declare no competing financial interest.

An interactive tutorial with analysis scripts is available at http://www.cup.lmu.de/pc/ochsenfeld/download/.

ACKNOWLEDGMENTS

The authors acknowledge Sophia Schwarz and Professor Oliver Trapp (LMU Munich) for their help in measuring the experimental infrared spectrum, and Professor Fritz Schaefer (University of Georgia, Athens, USA) for useful comments on our manuscript. Financial support was provided by the SFB 749 "Dynamics and Intermediates of Molecular Transformations" (DFG), the SFB 1309 "Chemical Biology of Epigenetic Modifications" (DFG), and the DFG cluster of excellence (EXC 114) "Center for Integrative Protein Science Munich" (CIPSM). C.O. acknowledges further support as Max-Planck-Fellow at the MPI-FKF Stuttgart.

REFERENCES

(1) Chipot, C., Pohorille, A., Eds. *Free Energy Calculations*; Springer-Verlag: Berlin Heidelberg, 2007.

(2) Hansen, N.; Van Gunsteren, W. F. Practical aspects of free-energy calculations: A review. *J. Chem. Theory Comput.* **2014**, *10*, 2632−2647.

(3) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087−1092.

(4) Hastings, W. K. Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika* **1970**, *57*, 97−109.

(5) Fermi, E.; Pasta, J.; Ulam, S. Studies of nonlinear problems. *Los Alamos Rep. LA-1940*; **1955**, DOI: 10.2172/4376203.

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

(7) Rahman, A. Correlations in the motion of atoms in liquid argon. *Phys. Rev.* **1964**, *136*, 405−411.

(8) Zwanzig, R. W. High-temperature equation of state by a perturbation method. I. Nonpolar gases. *J. Chem. Phys.* **1954**, *22*, 1420−1426.

(9) Bennett, C. H. Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.* **1976**, *22*, 245−268.

(10) Rickman, J. M.; LeSar, R. Free-Energy calculations in materials research. *Annu. Rev. Mater. Res.* **2002**, *32*, 195−217.

(11) Karplus, M.; Petsko, G. A. Molecular dynamics simulations in biology. *Nature* **1990**, *347*, 631−639.

(12) Gao, J.; Kuczera, K.; Tidor, B.; Karplus, M. Hidden thermodynamics of mutant proteins: a molecular dynamics analysis. *Science* **1989**, *244*, 1069−1072.

(13) Bash, P. A.; Singh, U. C.; Langridge, R.; Kollman, P. A. Free energy calculations by computer simulation. *Science* **1987**, *236*, 564−568.

(14) Smith, P. E.; van Gunsteren, W. When are free energy components meaningful? *J. Phys. Chem.* **1994**, *98*, 13735−13740.

DOI: 10.1021/acs.jpca.8b12309
J. Phys. Chem. A 2019, 123, 2163−2170

(15) Irwin, B. W. J.; Huggins, D. J. Estimating atomic contributions to hydration and binding using free energy perturbation. *J. Chem. Theory Comput.* **2018**, *14*, 3218−3227.
(16) Warshel, A.; Kato, M.; Pisliakov, A. V. Polarizable Force Fields: History, Test Cases, and Prospects. *J. Chem. Theory Comput.* **2007**, *3*, 2034−2045.
(17) Lipparini, F.; Barone, V. Polarizable Force Fields and Polarizable Continuum Model: A Fluctuating Charges/PCM Approach. 1. Theory and Implementation. *J. Chem. Theory Comput.* **2011**, *7*, 3711−3724.
(18) Baker, C. M. Polarizable force fields for molecular dynamics simulations of biomolecules. *WIREs Comput. Mol. Sci.* **2015**, *5*, 241−254.
(19) Balog, E.; Becker, T.; Oettl, M.; Lechner, R.; Daniel, R.; Finney, J.; Smith, J. C. Direct determination of vibrational density of states change on ligand binding to a protein. *Phys. Rev. Lett.* **2004**, *93*, 028103.
(20) Zhang, C.; Guidoni, L.; Kühne, T. D. Competing factors on the frequency separation between the OH stretching modes in water. *J. Mol. Liq.* **2015**, *205*, 42−45.
(21) Martinez, M.; Gaigeot, M.; Borgis, D.; Vuilleumier, R. Extracting effective normal modes from equilibrium dynamics at finite temperature. *J. Chem. Phys.* **2006**, *125*, 144106.
(22) Berens, P. H.; Mackay, D. H. J.; White, G. M.; Wilson, K. R. Thermodynamics and quantum corrections from molecular dynamics for liquid water. *J. Chem. Phys.* **1983**, *79*, 2375−2389.
(23) Lin, S.-T.; Maiti, P. K.; Goddard, W. A., III Two-Phase thermodynamic model for efficient and accurate absolute entropy of water from molecular dynamics simulations. *J. Phys. Chem. B* **2010**, *114*, 8191−8198.
(24) Persson, R. A.; Pattni, V.; Singh, A.; Kast, S. M.; Heyden, M. Signatures of solvation thermodynamics in spectra of intermolecular vibrations. *J. Chem. Theory Comput.* **2017**, *13*, 4467−4481.
(25) Pattni, V.; Vasilevskaya, T.; Thiel, W.; Heyden, M. Distinct protein hydration water species defined by spatially resolved spectra of intermolecular vibrations. *J. Phys. Chem. B* **2017**, *121*, 7431−7442.
(26) Lin, S. T.; Blanco, M.; Goddard, W. A., III The two-phase model for calculating thermodynamic properties of liquids from molecular dynamics: Validation for the phase diagram of Lennard-Jones fluids. *J. Chem. Phys.* **2003**, *119*, 11792−11805.
(27) Heyden, M. Disassembling solvation free energies into local contributions - Toward a microscopic understanding of solvation processes. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2019**, No. e1390.
(28) Peters, L. D. M.; Dietschreit, J. C. B.; Kussmann, J.; Ochsenfeld, C. Calculating free energies from the vibrational density of states function: Validation and critical assessment. *J. Chem. Phys.* **2018**, submitted.
(29) Runcie, A. C.; Zengerle, M.; Chan, K.-H.; Testa, A.; van Beurden, L.; Baud, M. G. J.; Epemolu, O.; Ellis, L. C. J.; Read, K. D.; Coulthard, V.; et al. Optimization of a 'bump-and-hole' approach to allele-selective BET bromodomain inhibition. *Chem. Sci.* **2018**, *9*, 2452−2468.
(30) Case, D.; Cerutti, D.; Cheatham, T.; Darden, T.; Duke, R.; Giese, T.; Gohlke, H.; Goetz, A.; Greene, D.; Homeyer, N.; et al. *AMBER 2017*; 2017.
(31) 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.
(32) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926−935.
(33) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781−1802.
(34) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. 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.
(35) Andersen, H. C. Rattle: A "velocity" version of the shake algorithm for molecular dynamics calculations. *J. Comput. Phys.* **1983**, *52*, 24−34.
(36) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684−3690.
(37) Gowers, R. J.; Linke, M.; Barnoud, J.; Reddy, T. J. E.; Melo, M. N.; Seyler, S. L.; Dotson, D. L.; Domanski, J.; Buchoux, S.; Kenney, I. M.; et al. In *Proceedings of the 15th Python in Science Conference*; Benthall, S., Rostrup, S., Eds.; 2016; *Chapter MDAnalysis: A Python package for the rapid analysis of molecular dynamics simulations*, pp 102−109.
(38) Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.* **2011**, *32*, 2319−2327.
(39) Kussmann, J.; Ochsenfeld, C. Pre-selective screening for matrix elements in linear-scaling exact exchange calculations. *J. Chem. Phys.* **2013**, *138*, 134114.
(40) 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.
(41) Kussmann, J.; Ochsenfeld, C. Hybrid CPU/GPU Integral Engine for Strong-Scaling Ab Initio Methods. *J. Chem. Theory Comput.* **2017**, *13*, 3153−3159.
(42) Sure, R.; Grimme, S. Corrected small basis set Hartree-Fock method for large systems. *J. Comput. Chem.* **2013**, *34*, 1672−1685.
(43) 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*, 154104.
(44) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **2011**, *32*, 1456−1465.
(45) Kruse, H.; Grimme, S. A geometrical correction for the inter- and intra-molecular basis set superposition error in Hartree-Fock and density functional theory calculations for large systems. *J. Chem. Phys.* **2012**, *136*, 154101.
(46) Verlet, L. Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Phys. Rev.* **1967**, *159*, 98−103.
(47) Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *J. Chem. Phys.* **1982**, *76*, 637−649.
(48) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.
(49) Niklasson, A. M. N.; Steneteg, P.; Odell, A.; Bock, N.; Challacombe, M.; Tymczak, C. J.; Holmström, E.; Zheng, G.; Weber, V. Extended Lagrangian Born-Oppenheimer molecular dynamics with dissipation. *J. Chem. Phys.* **2009**, *130*, 214109.
(50) Filippakopoulos, P.; Picaud, S.; Mangos, M.; Keates, T.; Lambert, J.-P.; Barsyte-Lovejoy, D.; Felletar, I.; Volkmer, R.; Müller, S.; Pawson, T.; et al. Histone recognition and large-scale structural analysis of the human bromodomain family. *Cell* **2012**, *149*, 214−231.
(51) Moritsugu, K.; Njunda, B. M.; Smith, J. C. Theory and Normal-Mode analysis of change in protein vibrational dynamics on ligand binding. *J. Phys. Chem. B* **2010**, *114*, 1479−1485.
(52) Filloux, C. M. The problem of origins and origins of the problem: Influence of language on studies concerning the anomeric effect. *Angew. Chem., Int. Ed.* **2015**, *54*, 8880−8894.
(53) Cocinero, E. J.; Carcabal, P.; Vaden, T. D.; Simons, J. P.; Davis, B. G. Sensing the anomeric effect in a solvent-free environment. *Nature* **2011**, *469*, 76−80.
(54) Mo, Y. Computational evidence that hyperconjugative interactions are not responsible for the anomeric effect. *Nat. Chem.* **2010**, *2*, 666−671.

DOI: 10.1021/acs.jpca.8b12309
J. Phys. Chem. A 2019, 123, 2163−2170

(55) Peters, L. D. M.; Kussmann, J.; Ochsenfeld, C. Efficient and accurate Born-Oppenheimer molecular dynamics for large molecular systems. *J. Chem. Theory Comput.* **2017**, *13*, 5479−5485.

DOI: 10.1021/acs.jpca.8b12309
J. Phys. Chem. A 2019, 123, 2163−2170

Supporting Information: Identifying Free Energy Hot-Spots in Molecular Transformations

Johannes C. B. Dietschreit,[1,2] Laurens D. M. Peters,[1,2]
Jörg Kussmann[1,2], Christian Ochsenfeld[1,2]

[1]Center for Integrated Protein Science (CIPSM) at the Department of Chemistry, University of Munich (LMU), Butenandtstr. 5–13, D-81377 München, Germany
[2]Chair of Theoretical Chemistry, Department of Chemistry, University of Munich (LMU), Butenandtstr. 7, D-81377 München, Germany

Contents

1 Image Processing

All images containing molecular geometries which are coloured according to changes in vibrational free energy were produced using *VMD*.[1] All plots showing spectra and distributions were produced using the python-package *matplotlib*.[2] The chemical structures were drawn with *ChemDraw*.

2 Data and Materials Availability

All inputs and trajectories are available upon request. PDB files (with the free energy colouring), analysis scripts, and an interactive tutorial are available at `http://www.cup.lmu.de/pc/ochsenfeld/download/`. *NAMD* is freely available for non-commercial users, while *FermiONs++* is not yet available.

3 Figures

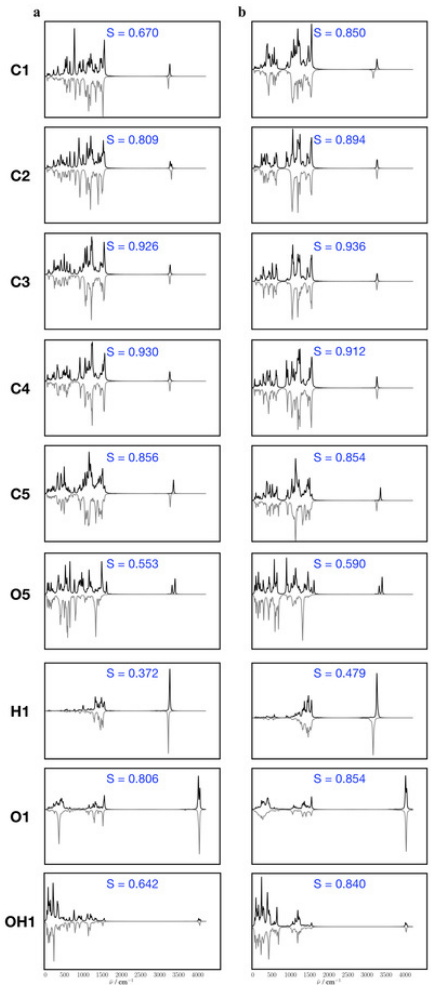


Figure S 1: **a**+**b**, Vibrational spectra per atom for (**a**) α-**HCT** (black) and α-**Glu** (inverted grey) and (**b**) β-**HCT** (black) and β-**Glu** (inverted grey). The overlap (S) between the two spectra is calculated as $S = \int_0^\infty I_1(\tilde{\nu})I_2(\tilde{\nu})\mathrm{d}\tilde{\nu} / \sqrt{\int_0^\infty I_1^2(\tilde{\nu})\mathrm{d}\tilde{\nu} \int_0^\infty I_2^2(\tilde{\nu})\mathrm{d}\tilde{\nu}}$, with I_1 and I_2 being the intensity of the two spectra at a wavenumber ($\tilde{\nu}$)). The overlap between α-**Glu** and α-**HCT** is generally smaller than the overlap between β-**Glu** and β-**HCT**. The wave number ($\tilde{\nu}$) increases from left to right.

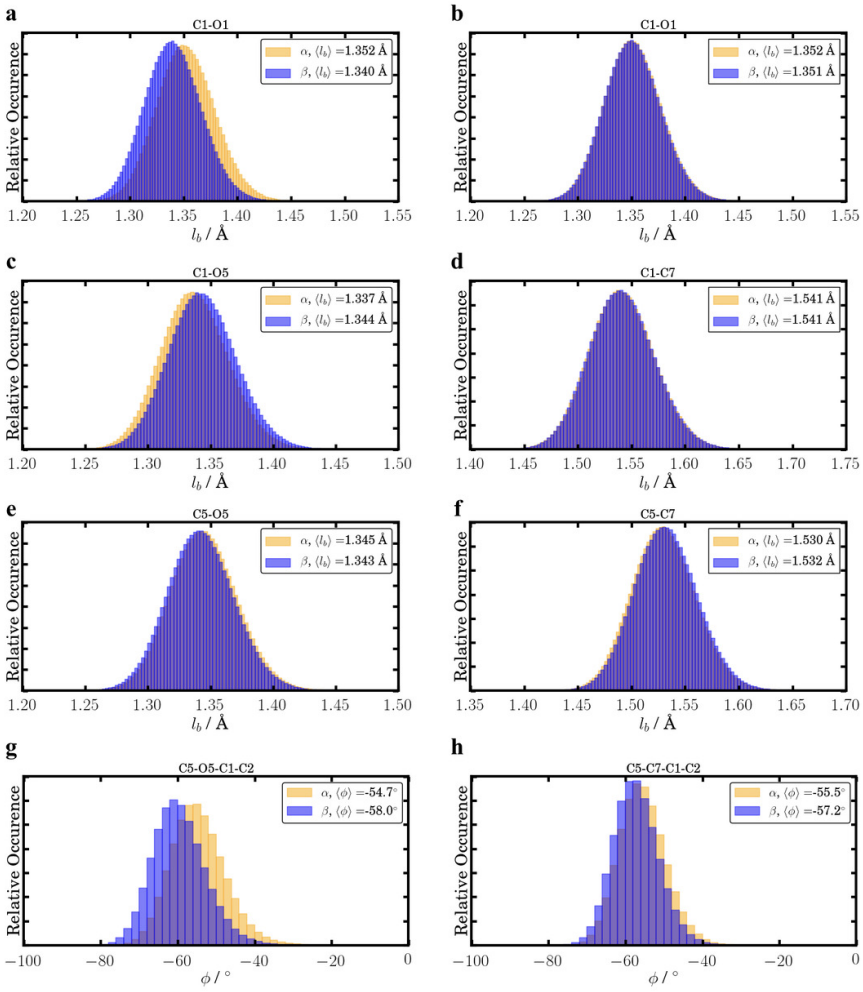


Figure S 2: Distribution of bond lengths and angles of the simulations of (left column) α-**Glu** and β-**Glu** and (right column) α-**HCT** and β-**HCT**, that serve as an indicator for the anomeric effect. α-**Glu** exhibits, in comparison to β-**Glu**, a shorter C1-O5 bond, a longer C1-O1 and C5-O5 bond, and a smaller C5-O5-C1-C2 dihedral angle. Similar observations cannot be made, when comparing the simulations of α-**HCT** and β-**HCT**.

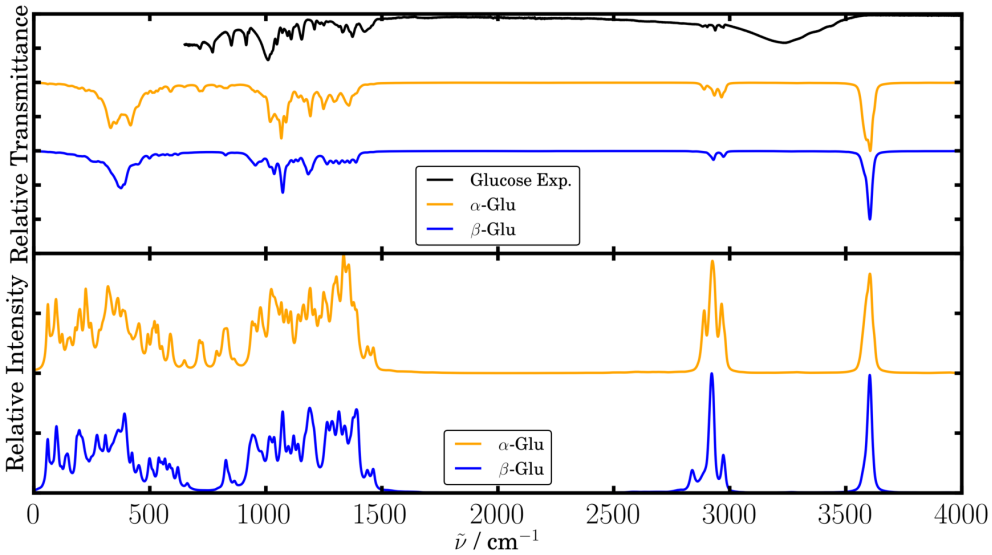


Figure S 3: (Top) Experimental IR spectrum of crystalline glucose-monohydrate (black) and simulated IR spectra (calculated as presented in ref. [3]) of α-**Glu** (orange) and β-**Glu** (blue). (Bottom) Computed vibrational power spectra of α-**Glu** (orange) and β-**Glu** (blue). The power spectrum has different intensities and also shows not IR-active vibrations exhibiting a small or no change in the dipole moment. For comparison to the experiment, the frequencies of the simulated spectra have been scaled by a factor of 0.82 (similar to the reported 0.81 in ref. [3]).

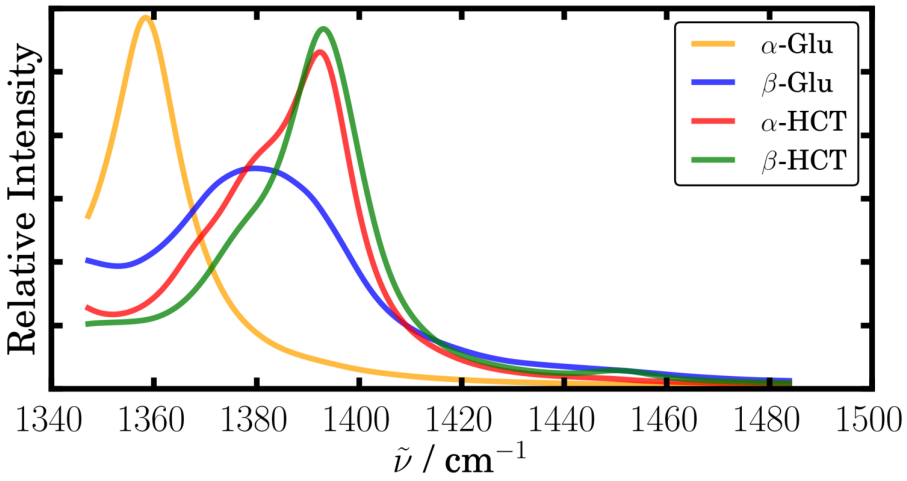


Figure S 4: Other vibrations than the C-H stretching bond are affected by the anomeric effect, e.g., the C-H deformation mode, which is clearly shifted in the case of α-**Glu**. For comparison to the experiment, the frequencies of the simulated spectra have been scaled by a factor of 0.82 (similar to the reported 0.81 in ref. [3]).

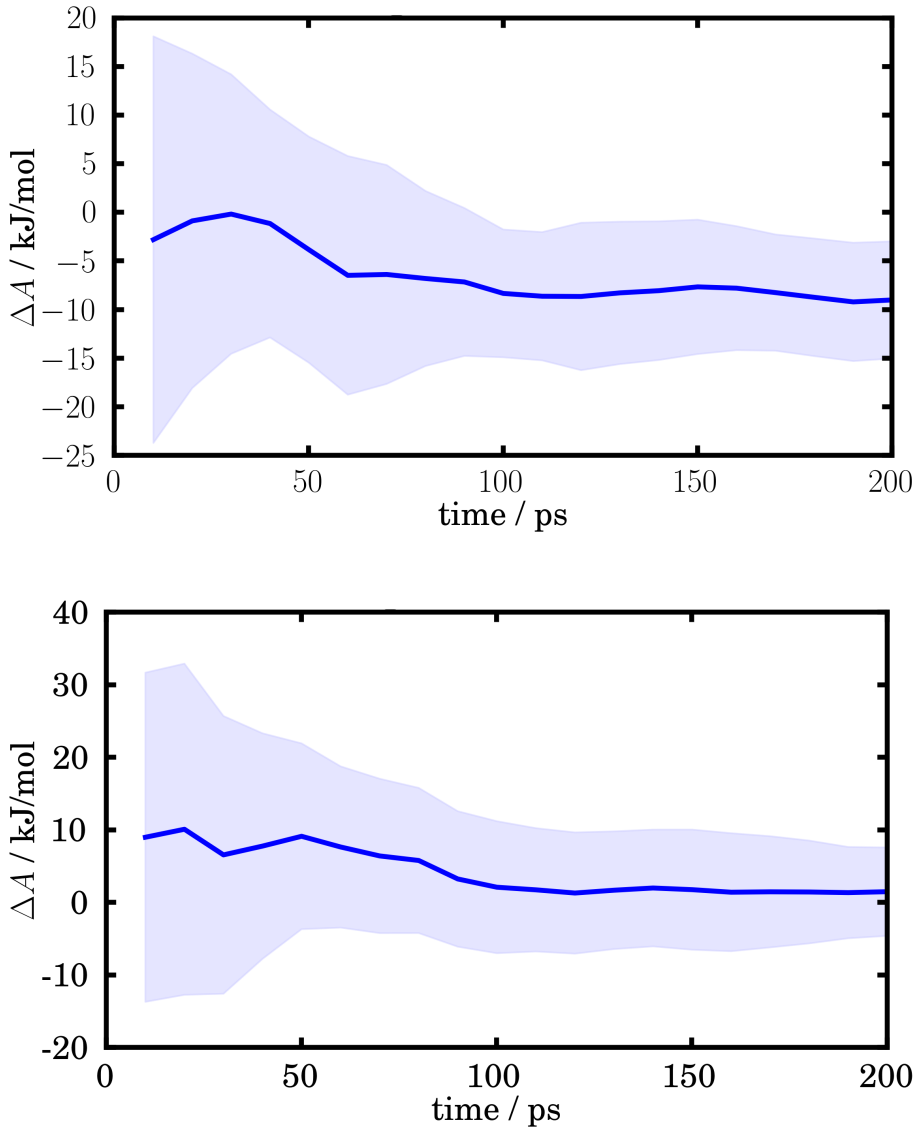


Figure S 5: Convergence of ΔA_{vib} for $\alpha \rightarrow \beta$ glucose (top) and HCT (bottom).

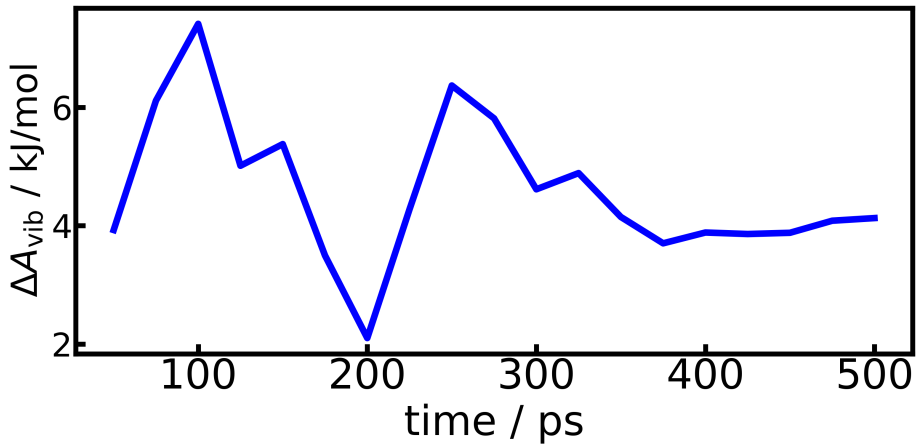


Figure S 6: Convergence of ΔA_{vib} for the Bromodomain shown for residue TRP-26, as it showed a large change in the free energy.

References

[1] W. Humphrey, A. Dalke, K. Schulten, *Journal of Molecular Graphics* **1996**, *14*, 33–38.

[2] J. D. Hunter, *Computing In Science & Engineering* **2007**, *9*, 90–95.

[3] L. D. M. Peters, J. Kussmann, C. Ochsenfeld, *J. Chem. Theory Comput.* **2017**, *13*, 5479–5485.

3.3 Publication III: Finding Reactive Configurations: A Machine Learning Approach for Estimating Energy Barriers Applied to Sirtuin 5

Beatriz von der Esch, **Johannes C. B. Dietschreit**, Laurens D. M. Peters, and Christian Ochsenfeld
"Finding Reactive Configurations: A Machine Learning Approach for Estimating Energy Barriers Applied to Sirtuin 5"
J. Chem. Theory Comput. **2019**, *15*, 6660-6667

Abstract: Sirtuin 5 is a class III histone deacetylase that, unlike its classification, mainly catalyzes desuccinylation and demanoylation reactions. It is an interesting drug target that we use here to test new ideas for calculating reaction pathways of large molecular systems such as enzymes. A major issue with most schemes (e.g., adiabatic mapping) is that the resulting activation barrier height heavily depends on the chosen educt conformation. This makes the selection of the initial structure decisive for the success of the characterization. Here, we apply machine learning to a large number of molecular dynamics frames and potential energy barriers obtained by quantum mechanics/molecular mechanics calculations in order to identify (1) suitable start-conformations for reaction path calculations and (2) structural features relevant for the first step of the desuccinylation reaction catalyzed by Sirtuin 5. The latter generally aids the understanding of reaction mechanisms and important interactions in active centers. Using our novel approach, we found eleven key features that govern the reactivity. We were able to estimate reaction barriers with a mean absolute error of 3.6 kcal/mol and identified reactive configurations.

JCTC Journal of Chemical Theory and Computation
Article
Cite This: J. Chem. Theory Comput. 2019, 15, 6660−6667
pubs.acs.org/JCTC

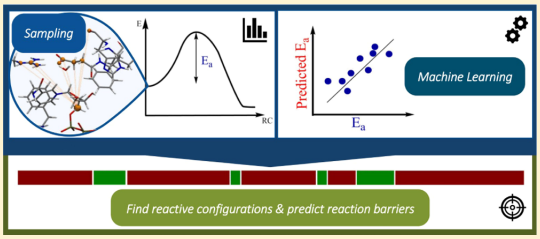
Finding Reactive Configurations: A Machine Learning Approach for Estimating Energy Barriers Applied to Sirtuin 5

Beatriz von der Esch,[†] Johannes C. B. Dietschreit,[†] Laurens D. M. Peters, and Christian Ochsenfeld*

Chair of Theoretical Chemistry, Department of Chemistry, University of Munich (LMU), Butenandtstr. 7, D-81377 München, Germany

Supporting Information

ABSTRACT: Sirtuin 5 is a class III histone deacetylase that, unlike its classification, mainly catalyzes desuccinylation and demanoylation reactions. It is an interesting drug target that we use here to test new ideas for calculating reaction pathways of large molecular systems such as enzymes. A major issue with most schemes (e.g., adiabatic mapping) is that the resulting activation barrier height heavily depends on the chosen educt conformation. This makes the selection of the initial structure decisive for the success of the characterization. Here, we apply machine learning to a large number of molecular dynamics frames and potential energy barriers obtained by quantum mechanics/molecular mechanics calculations in order to identify (1) suitable start-conformations for reaction path calculations and (2) structural features relevant for the first step of the desuccinylation reaction catalyzed by Sirtuin 5. The latter generally aids the understanding of reaction mechanisms and important interactions in active centers. Using our novel approach, we found eleven key features that govern the reactivity. We were able to estimate reaction barriers with a mean absolute error of 3.6 kcal/mol and identified reactive configurations.



INTRODUCTION

Computationally obtained reaction barriers are an excellent link to experiment. They allow us to verify or propose new reaction mechanisms, gain insights into kinetics, or compare reactivities. However, the calculation of reliable activation energies is a demanding task, especially for large molecular systems, for example, enzymes.

There exists a large collection of static and dynamic methods to model chemical reactions (e.g., adiabatic mapping,[1] nudged elastic band,[2,3] string methods,[4,5] transition path sampling,[6] umbrella sampling,[7] metadynamics,[8] and many more). Regardless of the method, there are two major challenges: (1) the choice of the theoretical description, and (2) the sampling bottleneck that leads to a ubiquitous dependency on the chosen start conformation. One of the main tools of choice for studying enzymatic systems is the combination of quantum mechanics and molecular mechanics (QM/MM) (see e.g., refs 9−11). The description of the QM part varies between semi-empirical (e.g., AM1 or SCC-DFTB)[12−15] and ab initio methods (e.g., HF or DFT).[16−18] Besides the level of theory chosen for the QM region, the extent of the QM sphere[19−21] and the treatment of the boundary region play an important role.[9−11] With increasing computational power and novel efficient methods, we are able to increase our attention to detail (e.g., solvent effects), and apply higher level theoretical methods. However, the second issue of selecting an initial configuration is nearly as important as the accuracy of the description of the electronic structure. In order to circumvent the need to search for suitable starting structures from a vast number of frames,[22,23] extensive sampling would be needed. Unfortunately, it becomes more and more demanding to sample the phase space with increasing system size and accuracy of the Hamiltonian. Therefore, for extended system such as enzymes, exploring the entire phase space remains prohibitively expensive at the QM/MM level. Start configurations can be taken from an MM-MD simulation. Alternatively, it has been suggested to start from the crystal structure, avoiding the selection problem entirely (see e.g., ref 24). However, the X-ray structures, which often differ from structures in solution, are not guaranteed to be reactive.[11] Even if they are suited for the initial step of a reaction, problems might arise for subsequent reaction steps.

Thus, it is paramount to develop a straight forward approach for pinpointing reactive configurations visited during the MM-MD, which are located at the beginning of reaction paths. The work of Lodola et al.[25] supports the importance of exploring the influence of conformational changes. They show the power of statistical tools, for example, principal component analysis, to identify conformational changes dominating enzymatic reactivity.[25] In a recent study, Bonk et al.[26] tried to link geometry and reactivity using machine learning during extensive transition interface sampling which enabled them to find reactive trajectories more often.

Received: August 30, 2019
Published: November 25, 2019

DOI: 10.1021/acs.jctc.9b00876
J. Chem. Theory Comput. 2019, 15, 6660−6667

Here, we apply QM/MM adiabatic mapping to a large selection of MM-MD frames to obtain an estimate of the reaction barrier starting from these snapshots. Adiabatic mapping is a straight forward approach to calculate the potential energy profile of a reaction, where a predefined reaction coordinate is gradually changed while the remaining system is relaxed. It should be noted that adiabatic mapping is not suitable for modelling reactions involving large structural rearrangements or changes in solvation.[11] We relate the initial structures taken from the MD trajectory and the calculated transition barriers using simple machine learning in order to understand which conformational changes influence the reactivity, and build a predictive model for activation energies. The model is subsequently applied to all MD frames in order to identify reactive regions within the trajectory. This set up is intended to help identify suitable start frames and therefore alleviate the need of extensive sampling, which is a true limitation at the QM/MM level.

As a model reaction, we investigate the first step of the desuccinylation reaction catalyzed by Sirt5, which belongs to the class of histone deacetylases.[27,28] Despite what its enzyme class name suggests, Sirt5 mainly catalyzes the desuccinylation or demanoylation of lysines and not a deacetylation.[29] This desuccinylation is thought to be a three step reaction which is initiated by a nucleophilic attack of the substrate on the NAD^+ cofactor that leads to the dissociation of nicotinamide.[27,30−32]

METHODS

Data Acquisition. *Structure Preparation.* The crystal structure (PDB: 3RIY[33]) consists of a dimer of Sirt5 in the complex with a histone tail peptide containing a succinylated lysine (SLL) as well as NAD^+. We selected the first monomer in the file (chain A for Sirt5 and chain D for the peptide) as well as the respective NAD^+. Hydrogen atoms were added using the program tleap from the program suite Amber16.[34] The protonation state of titratable residues was set according to PropKa.[35,36] The zinc finger in Sirt5 was parametrized using the ZAFF (Zinc Amber Force Field) parameters.[37] For the residue SLL, GAFF (Generalized Amber Force Field) parameters[38] were assigned using the Antechamber code, which determined the atomic partial charges from an AM1[12] calculation with bond-charge corrections (AM1−BCC).[39] The parameters for NAD^+ were taken from the AMBER parameter database.[40,41] All other parameters were taken from AmberFF14.[42] Finally, the system was solvated by placing it in a TIP3P[43] water box with a distance of 17 Å in all three dimensions at a density of 0.832 g/cm^3. The system was neutralized with one chloride ion.

MM-MD Simulation. Two minimizations (10 000 steps) were carried out to prepare the solvated system. During the first minimization, the protein was constrained and only the water molecules were optimized. In the second step, the entire system was subjected to the minimization. The system was heated to 300 K by increasing the temperature by 1 K every 100 fs. Afterward, the system was equilibrated for 100 000 time steps. During heating and equilibration, the temperature was controlled with simple velocity rescaling. The following production run was performed in the *NPT* ensemble for 200 ns. The pressure was kept at one atmosphere and the temperature at 300 K with the Langevin Piston barostat and Langevin thermostat implemented in NAMD.[44] The time step for equilibration and production was set to 2 fs. Nonbonded interactions were evaluated explicitly within 10 Å and smoothly switched off at 12 Å. A Verlet nearest neighbor list[45] with a radius of 13.5 Å was used to speed up the computations. Periodic boundary conditions were used in all three directions. Electrostatic interactions were evaluated with the particle mesh Ewald method[46] and an interpolation of the sixth order. The MD simulations were carried out with the NAMD[44] program package.

QM/MM Calculations. We selected frames (every 0.5 ns) from the production run as starting points for QM/MM calculations. All the frames were minimized twice at the MM level for 10 000 steps, again minimizing first only the solvent and then the full system. Subsequently, the frames were subjected to a QM/MM optimization. The QM region always included the residues Arg105, Phe70, Phe223, His158, part of NAD^+, and the succinyl-lysine residue, as well as all water molecules within 4 Å of the C1′ atom of the ribose in NAD^+, which are in total 139−151 atoms, depending on the number of water molecules in the active site (Supporting Information, SFigure 1 shows the QM region). The QM region was described at the HF-3c[47] level of theory and the MM region as specified in the section "Structure Preparation". The two subsystems were coupled via electrostatic embedding. The QM/MM calculations were performed within the ChemShell[48] code, with the QM part treated by the program package FermiONs++.[49,50]

We performed a small benchmark comparing HF-3c with higher level DFT methods to show that it is well suited for our endeavor. HF-3c consistently overestimates the reaction barrier. Trends in higher and lower barriers are reflected properly compared to DFT (see the Supporting Information for more details).

The optimized structures were used as starting points for adiabatic mapping pathways. The reaction coordinate was defined as the difference between the C1′−O bond and the C1′−N bond. While the C1′−O distance was reduced, the C1′−N bond was elongated. In each step, the bond difference was changed by 0.2 Å and fixed, while the remaining system was minimized.

Machine Learning. *Data Preprocessing.* We are interested in the relation between the educt configuration and the reaction barrier. Therefore, a representation of the geometry is needed that is suited to describe structural changes. There are several representations which are well established for chemical investigations such as Bag of Bonds,[51] *XYZ*-coordinates, Coulomb-matrices,[52,53] or SMILES.[54] Each of these representations is appropriate for different problems. Even though there is a number of established representations, we decided to simply select the distances between all nonhydrogen atoms within the QM region to describe the geometry in the active site. This representation allows for a preliminary correlation analysis which reduces the number of features in our system (see next section). Additionally, no further calculation of, for example, atomic charges is needed (which are heavily influenced by the employed QM method). The collection of interatomic distances is invariant to translation and rotation, and therefore, avoids any problems that might otherwise occur. Additionally, the number of water molecules within the QM region was considered as an additional feature. All in all, this added up to 2629 features.

Dimensionality Reduction. Because the outcome of a machine learning fit is dependent on expressive features and can be impaired by redundant or even insignificant variables, the features were purged. The dimensions were reduced by a

DOI: 10.1021/acs.jctc.9b00876
J. Chem. Theory Comput. 2019, 15, 6660−6667

simple correlation analysis. All features with absolute correlations <0.375 to the reaction barrier were omitted. This value was chosen quite low to ensure that most of the variations were captured. Further, the remaining features were checked for strong absolute correlations >0.9 among each other. If a pair of features were strongly correlated, one of them was omitted. This resulted in a subset containing only 15 features out of the original 2629.

Model Selection, Refinement, and Application. There exists a vast number of machine learning algorithms to choose from. Because we want to predict activation energies, we are trying to solve a typical regression problem from the mathematical point of view. There are different types of regression models, simple linear regression (least squares), polynomial regression, support vector regression, decision tree regression, to name a few.[55,56] The predictive power of the different machine learning models depends strongly on the structure and size of the data, and the relation between the target and feature variables. All machine learning scripts were performed in python with a combination of pandas[57] and scikit-learn.[58] We tested different supervised learning regression techniques, the results can be found in section "Machine Learning Model Comparison" of the Supporting Information. After testing we chose a sparse regression model, the elastic net regressor.[58,59]

Elastic net regression includes variable selection and regularization, which leads to a greater predictive power and enhances the interpretability of the results. Methods including regularization are especially suited for problems where little data is available. They suppress overfitting by introducing a cost function.[59] Based on all 150 samples, an elastic net model was built. The hyperparameter α, which controls the strength of the bias, and the l1_ratio (the ratio between the l1- and l2-type cost functions) were determined using fivefold cross validation. To evaluate the performance of the model the mean-absolute-error (MAE), RMSE, and R^2 value were calculated using threefold cross validation. To additionally visualize the skill of the machine learning model on new samples, the data set was randomly divided into a training and testing set (2:1), fitted to the training set and applied to the test set.

Lastly, the model was fitted to all the available data (all 150 frames), with the previously determined hyperparameters. The final model was then used to predict the reaction barrier for every MD frame (every 10 ps of the trajectory). For ten frames with low predicted reaction barriers, adiabatic mapping as described in the section "QM/MM Calculations" was carried out to show that the model helps to find reactive frames. The model generated here is not transferable, but the presented protocol can be employed for other extended systems.

RESULTS AND DISCUSSION

Reaction Barriers Obtained by Adiabatic Mapping. The combined QM/MM adiabatic mapping calculation of 250 reaction pathways starting from snapshots taken from an MM-MD simulation gave reaction barrier heights between 22 and 80 kcal/mol. Figure 1 shows how the calculated reaction barriers increase with an increasing number of water molecules.

As the MD simulation advances, the peptide and NAD^+ slightly unbind and more water molecules coordinate the carbonyl-oxygen involved in the first reaction step, and thus, its nucleophilicity decreases. After 75 ns, the adiabatic mapping approach was mostly incapable of describing the nicotinamide

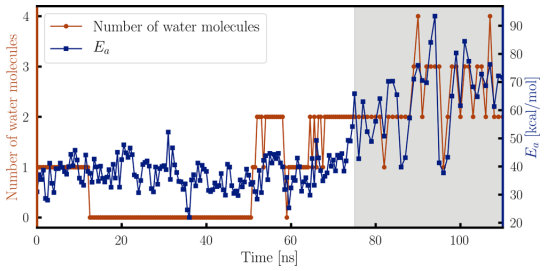


Figure 1. Number of water molecules within 4 Å of the C1′ atom of the ribose in NAD^+ is shown in red. The computed HF-3c activation energies are plotted in blue. The shaded area highlights the region that was not included in the subsequent machine learning steps.

cleavage, the desired products were no longer obtained. The incapability to model the reaction expresses itself in very high reaction barriers. Only the first 150 reaction pathways, starting from snapshots taken within the first 75 ns of the MD-trajectory, were included in the data-set for machine learning.

Figure 1 also shows that extended periods of the MD trajectory are especially nonreactive. This underlines that if only very few frames are picked or a very short MD simulation is used as basis for further calculations, one can miss reactive periods completely. The first 150 samples, each 0.5 ns apart along the MD-trajectory, yield energy barriers between 21 and 60 kcal/mol. The distribution of the barrier heights is shown in Figure 2. It highlights that educt configurations that lead to a

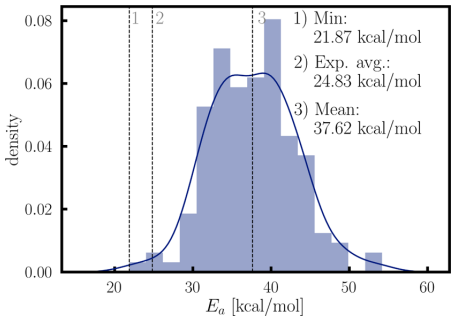


Figure 2. Distribution of the calculated energy barriers (adiabatic mapping with HF-3c). The blue line indicates a smooth distribution function fitted to the histogram. (1) Lowest barrier found, (2) exponentially averaged barrier, (3) mean barrier.

low energy transition are extremely rare. This emphasizes how difficult it is to find an appropriate start frame, that closely resembles the reactive enzyme complex and provides a reasonable energy barrier. The large variation of reaction profiles obtained with different initial configurations, and therefore the importance of suited starting points has been recognized early on (see e.g., refs 9, 25, 60−62).

As presumed by Ryde,[63] the energy barriers roughly form a Gaussian distribution. The arithmetic average is 37.62 kcal/mol, and the minimum activation barrier is 21.87 kcal/mol. The exponential average gives a good estimate for the barrier and is suitable for comparison with experiments, under the condition that the picked snapshots are well chosen;[64] here we obtain a value of 24.83 kcal/mol. The

DOI: 10.1021/acs.jctc.9b00876
J. Chem. Theory Comput. 2019, 15, 6660−6667

standard deviation, σ, within these 150 samples is 5.40 kcal/mol. Based on the conclusions by Ryde,[63] more than 10^6 samples would be needed to obtain an estimate of the activation barrier within chemical accuracy (within 1 kcal/mol) of the exponential average. In contrast, most studies have only used a few snapshots (about 3−10).[65−68] To avoid having to calculate millions of pathways, we propose a strategic, scalable approach. We suggest using machine learning based on selected distances to pinpoint reactive regions within the MM trajectory. This allows for strategic sampling of reaction pathways that contribute significantly to the exponential average, giving a more accurate estimate of the energy barrier using less samples. Alternatively, productive snapshots from the MD trajectory found by the machine learning model, can be used for further (more accurate) QM/MM studies.

Machine Learning Performance. Using Elastic Net regression with 15 input features and 150 samples, a model was built to predict reaction energies from geometrical features. With α set to 0.09 and l1_ratio equal to 0.5, a MAE of 3.58 kcal/mol and a root mean squared error (RMSD) 4.46 kcal/mol is obtained. The R^2 score, which describes the percentage of the response variable variation that is explained, is 0.28. In general an R^2 score of 0.28 may be regarded as very poor. However, with respect to the complexity of the system and the very limited number of training points (100) a score of 0.28 is surprisingly high. Additionally, only 15 features were needed to describe the problem to this level of accuracy, which is possibly influenced by a much greater collection of residues.

To visualize the performance of the machine learning model, the data set was randomly spilt into a test (50 samples) and a training set (100 samples). This 1:2 division is similar to the one made during one cycle of the threefold cross validation used to assess performance. Subsequently the model was fitted on the training set and applied to the test set. The predictions for the test set versus the activation barriers calculated using adiabatic mapping are shown in Figure 3. The predictions of the regression model are in quite good agreement with the actual activation barriers. The model is least accurate for the extreme barriers, it overestimates low barriers and underestimates high barriers. These are the regions in the training distribution of reaction barriers with the least number of samples. In order to increase the predictive power of the model, without prior knowledge of the system, more training points are needed. We suggest to test if semi-empirical methods might be a solution to the sampling bottleneck.

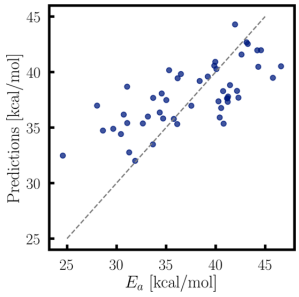


Figure 3. Scatter plot showing the performance of the Elastic Net Regressor on a test set (50 random points which were not used for learning).

Another way to increase the predictive power is to iteratively apply the model: calculate frames with predicted low barrier heights, add the results to the training data, and enhance its performance with every cycle. However, enhancing the performance of the model is only necessary if the goal is to predict a final energy barrier using this approach. That being said, the model is able to differentiate between less and more reactive frames. Therefore, this straight forward approach is sufficient to identify regions of interest within the MD trajectory, which is the intent of this work. Appropriate starting geometries identified by the built model can then be used for involved QM/MM studies, for example, aiming at calculating a free energy reaction profile.

Analysis of the Resulting Feature Subset. The group of features that remained after the two selection steps (explained in section "Dimensionality Reduction") is shown in the following three tables (Tables 1−3). For each of the features, the indices of the involved atoms (pdb file of the entire system is attached to the Supporting Information), the Pearson correlation coefficient to the activation energy, the elastic net coefficient, and an explanatory figure are given. The distances are grouped into 3 categories. The first category contains distances between the binding pocket and either the substrate or the cofactor (Table 1). The second group consists of intramolecular distances of SLL and NAD^+ (Table 2). The last group contains the intermolecular interactions between SLL and NAD^+ (Table 3).

The features given in Table 1 are distances from the substrate and the cofactor to the surrounding amino acids. Two distances are between atoms of PHE 70 to SLL (1) and NAD^+ (2), which are anticorrelated to the activation energy and contribute to the predictive model. The two features indicate that PHE 70 and the attached backbone must allow enough space for the nicotinamide leaving group to move out of the binding pocket. Hence, if the SLL−PHE 70 and the NAD^+−PHE 70 distances increase, the activation barriers become lower. Feature 3 and 4 show that the binding pocket has to be compact and the NAD^+ cofactor has to be located deep in the active center for the reaction to take place.

The second category includes intramolecular distances. It shows that small conformational changes within the reactants clearly influence the reactivity. Features 5 and 6 express the relative position of the nicotinamide to the ribose ring. As they are very similar, feature 6 was eliminated by the elastic net model due to its redundancy.

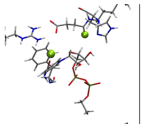
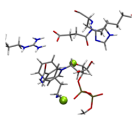
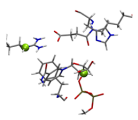
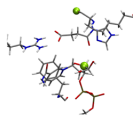
The alignment of the succinyl group plays a major role. Feature 7 has the highest absolute coefficient of all the features and therefore has the greatest impact on the predicted transition barrier. Feature 7 expresses the distance between the C4 atom and the terminal carboxyl group. This distance is anticorrelated to the activation barrier, and thus the barrier is lowest when the negatively charged carboxyl group is furthest away from the reactive centers.

The last group is the largest, it contains eight features which describe the relative positions of NAD^+ and SLL. Features 8, 9, and 10 are related to the previously explained feature 7. These distances are also a measure of the relative position of the carboxyl group, and therefore redundant, their coefficients are small or zero. The other five distances between NAD^+ and SLL show all positive correlation to the energy barrier. They indicate that the substrate and the cofactor have to be sufficiently aligned in order for the reaction to take place. Additionally, based on the large number of features containing

DOI: 10.1021/acs.jctc.9b00876
J. Chem. Theory Comput. 2019, 15, 6660−6667

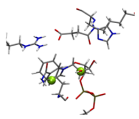
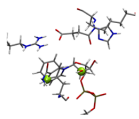
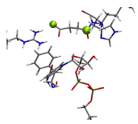
Journal of Chemical Theory and Computation Article

Table 1. Features 1−4 Used in the Elastic Net Model[a]

				
Number	1	2	3	4
Atom indices[b]	4192, 584	4281, 594	4279, 1165	4280, 2022
Corr. to E_a	-0.38	-0.39	0.39	0.38
Coefficient	-1.12	-0.78	0.82	1.19

[a]These four features describe the overall configuration of the active site. The atoms in between which the distance is measured is colored in green. [b]Atom indices as in the pdb-file (see Supporting Information).

Table 2. Features 5−7 Used in the Elastic Net Model for Describing Interactions Within SLL and NAD$^+$[a]

			
Number	5	6	7
Atom indices[b]	4294, 4284	4293, 4284	4194, 4191
Corr. to E_a	0.41	0.40	-0.45
Coefficient	0.66	0.0	-2.99

[a]The atoms in between which the distance is measured is colored in green. [b]Atom indices as in the pdb-file (see Supporting Information).

the ribose ring, we suspect that the pucker of the ring plays an important role.

Application of the Trained Model to the Entire MD Simulation. The final model, which was trained on all 150 samples, was applied to the entire MD-trajectory. The predicted barrier heights for the initial step of the dessuccinylation are shown in Figure 4. One can see the general agreement between predicted (blue) and calculated MD frames (orange). The changes in the reactivity are captured and reflected by the estimated barriers. It is interesting to note that there are periods in the MD trajectory which are either reactive or nonreactive, and others in which the reactivity oscillates very strongly.

The distribution of the predicted activation energies is shown in Figure 5 on the left. It is compared to the initial collection of barrier heights used for learning. The comparison

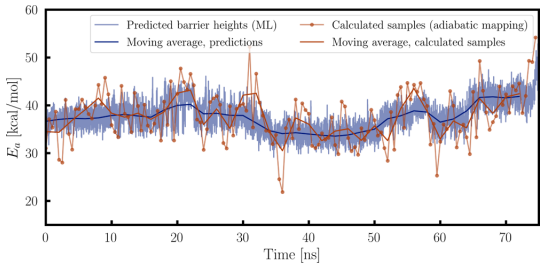
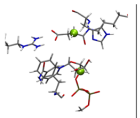
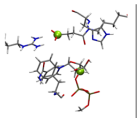
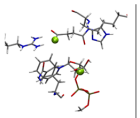
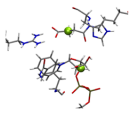
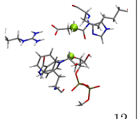
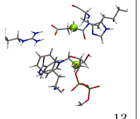
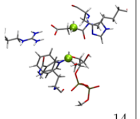
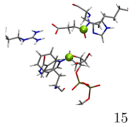


Figure 4. Section from the entire MD for which the activation energies were predicted with the previously built model. The red dots indicate the energy barriers calculated with adiabatic mapping.

shows that the distribution of the predictions is much narrower. This already suggests that the model will overestimate low energy transitions and underestimate high barriers.

In order to check the reliability of the model for predicting reactive regions within the MD trajectory, we selected 10 frames for which a low barrier was forecast and three additional snapshots to represent the frames with higher predicted activation energies. These three additional samples are the frames at 25, 50, and 75% of the distribution of predicted transition barriers. Starting from these snapshots, adiabatic mapping calculations were carried out. The results for the picked frames that were modeled are shown in Figure 5 on the right. The predicted (ML) values and the calculated results (adiabatic mapping with HF-3c) are compared. They are put

Table 3. Features 8−15 Used in the Elastic Net Model for Describing the Interactions between SLL and NAD$^+$[a]

				
Number	8	9	10	11
Atom indices[b]	4279, 4188	4279, 4190	4279, 4195	4279, 4189
Corr. to E_a	0.43	0.41	0.42	0.43
Coefficient	0.0	-0.16	0.0	1.81
				
Number	12	13	14	15
Atom indices[b]	4282, 4188	4284, 4188	4286, 4188	4286, 4191
Corr. to E_a	0.45	0.43	0.54	0.48
Coefficient	0.92	0.0	0.80	0.10

[a]The atoms in between which the distance is measured is colored in green. [b]Atom indices as in the pdb-file (see Supporting Information).

DOI: 10.1021/acs.jctc.9b00876
J. Chem. Theory Comput. 2019, 15, 6660−6667

Journal of Chemical Theory and Computation Article

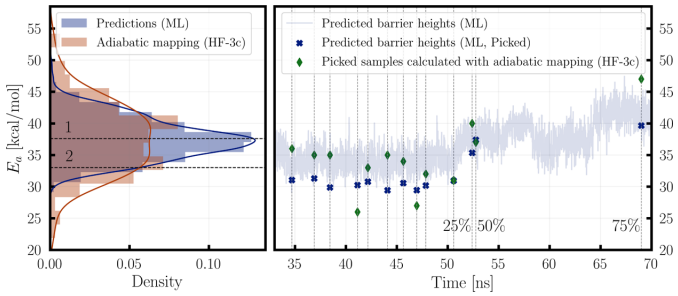


Figure 5. Left: Distribution of the predicted barrier heights (blue). For comparison the distribution of the initially calculated barriers (adiabatic mapping with HF-3c), used for learning, is given (orange). The arithmetic mean (1) and exponential average (2) of the predicted barriers are 37.60 and 33.02 kcal/mol, respectively. Right: Comparison of predicted (ML) and calculated (adiabatic mapping with HF-3c) reaction barriers for 10 frames with low energy transitions and three additional representative snapshots. The values shown here are listed in Table 4.

into context with the predicted barriers for all frames of the MM trajectory and the samples originally given to the model for training.

The predicted barrier heights and the calculated reaction barriers for the thirteen frames are listed in Table 4.

Table 4. Comparison of Predicted and Calculated Barrier Heights for Ten Frames with Low Estimated Reaction Barriers by the ML Model[a]

time [ns]	$E_a^{Pred.}$	$E_a^{Calc.}$ [kcal/mol]	ΔE
34.70	31	36	5
36.89	31	35	4
38.44	30	35	5
41.14	30	**26**	−4
42.16	31	33	2
44.08	30	35	5
45.64	31	34	3
46.96	29	**27**	−2
47.85	30	32	2
50.58	31	31	0
52.39[b]	35	40	5
52.76[c]	37	34	−3
69.00[d]	40	47	7

[a]Three additional values are given for frames from 25, 50, and 75% of the distribution of predicted transition barriers. Bold numbers indicate calculated barriers that are below 30 kcal/mol. In general, all calculated activation energies are close to the predicted values. The MAE for these 13 samples is 3.6 kcal/mol. [b]25%. [c]50%. [d]75%.

The comparison of the calculated and predicted activation energies shows that the designed model overestimates low energy transitions. The start geometries that lead to low transitions are few compared to the number of snapshots that are unsuitable starting points for QM/MM reaction path studies. From the original 150 samples only 9 had energy barriers below 30 kcal/mol. Using the machine learning model 2 out of the 10 frames, thought to be suited, lead to barriers lower than 30 kcal/mol. Therefore, the model allows us to identify relevant frames that will contribute significantly to the exponential average of the reaction barrier. For an accurate estimate of the exponential average, more data points used for training would be required. Improving the predictive model and subsequently calculating the exponential average from all predicted barriers could be an interesting approach to approximate the true activation barrier, which then can be compared to experiments. Overall, we are able to meet our goal to strategically find reactive regions within the MD-trajectory. Using the model, we are able to exclude the majority of frames without needing to calculate them specifically.

CONCLUSIONS

Using simple machine learning techniques, we are able to find reactive periods within the MD trajectory without prior knowledge of the structural factors that govern the reactivity of Sirtuin 5. The applied protocol enables us to identify the structural features that stabilize the transition state, and thus enhance the reactivity.

We found that the cofactor NAD^+ and the substrate SLL have to be located close together and be well aligned; therefore, the compactness of the binding pocket is a prerequisite. At the same time, there has to be sufficient room for nicotinamide, the leaving group, to exit the active site. Configurational changes within NAD^+ and SLL are also connected to the reactivity. The relative position of the nicotinamide to the ribose ring in NAD^+, the orientation of the terminal carboxyl group of SLL and its salt bridge to the neighboring ARG 105 are important structural features. Using measurements of these changes we were able to estimate activation energies with a MAE of 3.6 kcal/mol. For the initial step of the desuccinylation, we found transitions with barriers as low as 26 kcal/mol. We expect that the inclusion of dynamic effects through free energy simulations and even more accurate methods will yield a more reliable transition barrier than found in the scope of this work. These results also support the assumption that the dessucinylation investigated here has a reaction mechanism which is analogous to the deacetylation by Sirtuin 2, which has already been studied in greater detail.[30−32] The straightforward approach we applied here to estimate transition barriers is transferable to any extended system. It greatly simplifies the search for appropriate educt conformations, which significantly influences the outcome of most QM/MM-schemes to model enzymatic reaction mechanisms. The approach is scalable and can be easily customized to meet individual needs, by employing other descriptions for the MM or the QM part, adjusting the number of samples or adding further features.

DOI: 10.1021/acs.jctc.9b00876
J. Chem. Theory Comput. 2019, 15, 6660−6667

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.9b00876.

Visualization of the QM region, benchmark: HF-3c versus other functionals, and machine learning model comparison (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: christian.ochsenfeld@uni-muenchen.de.

ORCID

Johannes C. B. Dietschreit: 0000-0002-5840-0002

Christian Ochsenfeld: 0000-0002-4189-6558

Author Contributions

†B.v.d.E. and J.C.B.D. contributed equally to this work

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge financial support by the SFB 1309 "Chemical Biology of Epigenetic Modifications" (DFG) and the DFG cluster of excellence (EXC 114) "Center for Integrated Protein Science Munich" (CIPSM). C.O. acknowledges further support as Max-Planck-Fellow at the MPI-FKF Stuttgart.

■ REFERENCES

(1) 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.

(2) Mills, G.; Jónsson, H. Quantum and thermal effects in H_2 dissociative adsorption: Evaluation of free energy barriers in multidimensional quantum systems. *Phys. Rev. Lett.* **1994**, *72*, 1124.

(3) 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.

(4) Weinan, E.; Ren, W.; Vanden-Eijnden, E. String method for the study of rare events. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2002**, *66*, 052301.

(5) Behn, A.; Zimmerman, P. M.; Bell, A. T.; Head-Gordon, M. Efficient exploration of reaction paths via a freezing string method. *J. Chem. Phys.* **2011**, *135*, 224108.

(6) Dellago, C.; Bolhuis, P. G.; Csajka, F. S.; Chandler, D. Transition path sampling and the calculation of rate constants. *J. Chem. Phys.* **1998**, *108*, 1964−1977.

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

(8) Laio, A.; Parrinello, M.; Li, Y.; Zhang, R.; Du, L.; Zhang, Q.; Wang, W. Escaping Free-Energy Minima. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12562.

(9) Eurenius, K. P.; Chatfield, D. C.; Brooks, B. R.; Hodoscek, M. Enzyme Mechanisms with Hybrid Quantum and Molecular Mechanical Potentials. I. Theoretical Considerations. *Int. J. Quantum Chem.* **1996**, *60*, 89−1200.

(10) Senn, H. M.; Thiel, W. QM/MM Methods for Biomolecular Systems. *Angew. Chem., Int. Ed.* **2009**, *48*, 1198−1229.

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

(12) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.

(13) 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: Condens. Matter Mater. Phys.* **1998**, *58*, 7260−7268.

(14) Hur, S.; Bruice, T. C. The near attack conformation approach to the study of the chorismate to prephenate reaction. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 12015−12020.

(15) Guo, H.; Cui, Q.; Lipscomb, W. N.; Karplus, M. Understanding the Role of Active-Site Residues in Chorismate Mutase Catalysis from Molecular-Dynamics Simulations. *Angew. Chem., Int. Ed.* **2003**, *42*, 1508−1511.

(16) Blank, I. D.; Sadeghian, K.; Ochsenfeld, C. A Base-Independent Repair Mechanism for DNA Glycosylase−No Discrimination Within the Active Site. *Sci. Rep.* **2015**, *5*, 10369.

(17) Roßbach, S.; Ochsenfeld, C. Quantum-Chemical Study of the Discrimination against dNTP in the Nucleotide Addition Reaction in the Active Site of RNA Polymerase II. *J. Chem. Theory Comput.* **2017**, *13*, 1699−1705.

(18) Kreppel, A.; Blank, I. D.; Ochsenfeld, C. Base-Independent DNA Base-Excision Repair of 8-Oxoguanine. *J. Am. Chem. Soc.* **2018**, *140*, 4522−4526.

(19) Roßbach, S.; Ochsenfeld, C. Influence of Coupling and Embedding Schemes on QM Size Convergence in QM/MM Approaches for the Example of a Proton Transfer in DNA. *J. Chem. Theory Comput.* **2017**, *13*, 1102−1107.

(20) Sumowski, C. V.; Ochsenfeld, C. A Convergence Study of QM/MM Isomerization Energies with the Selected Size of the QM Region for Peptidic Systems. *J. Phys. Chem. A* **2009**, *113*, 11734−11741.

(21) Kulik, H. J.; Zhang, J.; Klinman, J. P.; Martínez, T. J. How Large Should the QM Region Be in QM/MM Calculations? The Case of Catechol O-Methyltransferase. *J. Phys. Chem. B* **2016**, *120*, 11381−11394.

(22) Sadiq, S. K.; Coveney, P. V. Computing the Role of Near Attack Conformations in an Enzyme-Catalyzed Nucleophilic Bimolecular Reaction. *J. Chem. Theory Comput.* **2015**, *11*, 316−324.

(23) Santos-Martins, D.; Calixto, A. R.; Fernandes, P. A.; Ramos, M. J. A Buried Water Molecule Influences Reactivity in α-Amylase on a Subnanosecond Time Scale. *ACS Catal.* **2018**, *8*, 4055−4063.

(24) Neves, R. P. P.; Fernandes, P. A.; Ramos, M. J. Mechanistic insights on the reduction of glutathione disulfide by protein disulfide isomerase. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, E4724−E4733.

(25) Lodola, A.; Sirirak, J.; Fey, N.; Rivara, S.; Mor, M.; Mulholland, A. J. Structural Fluctuations in Enzyme-Catalyzed Reactions: Determinants of Reactivity in Fatty Acid Amide Hydrolase from Multivariate Statistical Analysis of Quantum Mechanics/Molecular Mechanics Paths. *J. Chem. Theory Comput.* **2010**, *6*, 2948−2960.

(26) Bonk, B. M.; Weis, J. W.; Tidor, B. Machine Learning Identifies Chemical Characteristics That Promote Enzyme Catalysis. *J. Am. Chem. Soc.* **2019**, *141*, 4108−4118.

(27) Schemies, J.; Uciechowska, U.; Sippl, W.; Jung, M. NAD^+-dependent histone deacetylases (sirtuins) as novel therapeutic targets. *Med. Res. Rev.* **2009**, *30*, 861−889.

(28) Parihar, P.; Solanki, I.; Mansuri, M. L.; Parihar, M. S. Mitochondrial sirtuins: Emerging roles in metabolic regulations, energy homeostasis and diseases. *Exp. Gerontol.* **2015**, *61*, 130−141.

(29) Nakagawa, T.; Guarente, L. Sirtuins at a glance. *J. Cell Sci.* **2011**, *124*, 833−838.

(30) Liang, Z.; Shi, T.; Ouyang, S.; Li, H.; Yu, K.; Zhu, W.; Luo, C.; Jiang, H. Investigation of the Catalytic Mechanism of Sir2 Enzyme with QM/MM Approach: SN1 vs SN2? *J. Phys. Chem. B* **2010**, *114*, 11927−11933.

(31) Hawse, W. F.; Hoff, K. G.; Fatkins, D. G.; Daines, A.; Zubkova, O. V.; Schramm, V. L.; Zheng, W.; Wolberger, C. Structural Insights into Intermediate Steps in the Sir2 Deacetylation Reaction. *Structure* **2008**, *16*, 1368−1377.

DOI: 10.1021/acs.jctc.9b00876
J. Chem. Theory Comput. 2019, 15, 6660−6667

(32) Wang, Y.; Fung, Y. M. E.; Zhang, W.; He, B.; Chung, M. W. H.; Jin, J.; Hu, J.; Lin, H.; Hao, Q. Deacylation Mechanism by SIRT2 Revealed in the 1-SH-2-O-Myristoyl Intermediate Structure. *Cell Chem. Biol.* **2017**, *24*, 339−345.
(33) Du, J.; Zhou, Y.; Su, X.; Yu, J. J.; Khan, S.; Jiang, H.; Kim, J.; Woo, J.; Kim, J. H.; Choi, B. H.; et al. Sirt5 is a NAD-dependent protein lysine demalonylase and desuccinylase. *Science* **2011**, *334*, 806−809.
(34) Case, D.; Betz, R.; Cerutti, D.; Cheatham, T. E.; Darden, T.; Duke, R.; Giese, T.; Gohlke, H.; Goetz, A.; Homeyer, N.; et al. *AMBER*, 2016.
(35) Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *J. Chem. Theory Comput.* **2011**, *7*, 525−537.
(36) Søndergaard, C. R.; Olsson, M. H. M.; Rostkowski, M.; Jensen, J. H. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values. *J. Chem. Theory Comput.* **2011**, *7*, 2284−2295.
(37) Peters, M. B.; Yang, Y.; Wang, B.; Füsti-Molnár, L.; Weaver, M. N.; Merz, K. M., Jr. Structural Survey of Zinc-Containing Proteins and Development of the Zinc AMBER Force Field (ZAFF). *J. Chem. Theory Comput.* **2010**, *6*, 2935−2947.
(38) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general Amber force field. *J. Comput. Chem.* **2004**, *25*, 1157−1174.
(39) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *J. Comput. Chem.* **2000**, *21*, 132−146.
(40) Pavelites, J. J.; Gao, J.; Bash, P. A.; MacKerell, A. D. A molecular mechanics force field for NAD^+ NADH, and the pyrophosphate Groups of nucleotides. *J. Comput. Chem.* **1997**, *18*, 221−239.
(41) Walker, R. C.; De Souza, M. M.; Mercer, I. P.; Gould, I. R.; Klug, D. R. Large and fast relaxations inside a protein: Calculation and measurement of reorganization energies in alcohol dehydrogenase. *J. Phys. Chem. B* **2002**, *106*, 11658−11665.
(42) 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.
(43) 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.
(44) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781−1802.
(45) Verlet, L. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.* **1967**, *159*, 98−103.
(46) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N·log (N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089−10092.
(47) Sure, R.; Grimme, S. Corrected small basis set Hartree-Fock method for large systems. *J. Comput. Chem.* **2013**, *34*, 1672−1685.
(48) Sherwood, P.; De Vries, A. H.; Guest, M. F.; Schreckenbach, G.; Catlow, C. R. A.; French, S. A.; Sokol, A. A.; Bromley, S. T.; Thiel, W.; Turner, A. J.; et al. QUASI: A general purpose implementation of the QM/MM approach and its application to problems in catalysis. *J. Mol. Struct.: THEOCHEM* **2003**, *632*, 1−28.
(49) Kussmann, J.; Ochsenfeld, C. Pre-selective screening for matrix elements in linear-scaling exact exchange calculations. *J. Chem. Phys.* **2013**, *138*, 134114.
(50) 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.
(51) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; Von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *J. Phys. Chem. Lett.* **2015**, *6*, 2326−2331.
(52) Rupp, M.; Tkatchenko, A.; Müller, K. R.; Von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.
(53) Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; Von Lilienfeld, O. A.; Tkatchenko, A.; Müller, K.-R. Assessment and validation of machine learning methods for predicting molecular atomization energies. *J. Chem. Theory Comput.* **2013**, *9*, 3404−3419.
(54) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(55) Rupp, M. Machine learning for quantum mechanics in a nutshell. *Int. J. Quantum Chem.* **2015**, *115*, 1058−1073.
(56) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547−555.
(57) McKinney, W. Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, 2010; pp 51−56.
(58) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(59) Hastie, T.; Zou, H. Regularization and Variable Selection via the Elastic Net. *J. R. Stat. Soc. Ser. B* **2005**, *67*, 301−320.
(60) Scheiner, S. Comparison of proton transfers in heterodimers and homodimers of NH_3 and OH_2. *J. Chem. Phys.* **1982**, *77*, 4039−4050.
(61) Scheiner, S. Proton Transfers in Hydrogen-Bonded Systems. Cationic Oligomers of Water. *J. Am. Chem. Soc.* **1981**, *103*, 315−320.
(62) Scheiner, S.; Harding, L. B. Proton Transfers in Hydrogen-Bonded Systems. 2. Electron Correlation Effects in $(N_2H_7)^+$. *J. Am. Chem. Soc.* **1981**, *103*, 2169−2173.
(63) Ryde, U. How Many Conformations Need to Be Sampled to Obtain Converged QM/MM Energies? the Curse of Exponential Averaging. *J. Chem. Theory Comput.* **2017**, *13*, 5745−5752.
(64) Cooper, A. M.; Kästner, J. Averaging Techniques for Reaction Barriers in QM/MM Simulations. *ChemPhysChem* **2014**, *15*, 3264−3269.
(65) Lonsdale, R.; Houghton, K. T.; Żurek, J.; Bathelt, C. M.; Foloppe, N.; de Groot, M. J.; Harvey, J. N.; Mulholland, A. J. Quantum Mechanics/Molecular Mechanics Modeling of Regioselectivity of Drug Metabolism in Cytochrome P450 2C9. *J. Am. Chem. Soc.* **2013**, *135*, 8001−8015.
(66) Sokkar, P.; Boulanger, E.; Thiel, W.; Sanchez-Garcia, E. Hybrid Quantum Mechanics/Molecular Mechanics/Coarse Grained Modeling: A Triple-Resolution Approach for Biomolecular Systems. *J. Chem. Theory Comput.* **2015**, *11*, 1809−1818.
(67) Lonsdale, R.; Reetz, M. T. Reduction of α,β-Unsaturated Ketones by Old Yellow Enzymes: Mechanistic Insights from Quantum Mechanics/Molecular Mechanics Calculations. *J. Am. Chem. Soc.* **2015**, *137*, 14733−14742.
(68) Li, Y.; Zhang, R.; Du, L.; Zhang, Q.; Wang, W. Insight into the Catalytic Mechanism of Meta-Cleavage Product Hydrolase BphD: A Quantum Mechanics/Molecular Mechanics Study. *RSC Adv.* **2015**, *5*, 66591−66597.

DOI: 10.1021/acs.jctc.9b00876
J. Chem. Theory Comput. 2019, 15, 6660−6667

Supporting Information: "Finding reactive configurations: A machine learning approach for estimating energy barriers applied to Sirutin 5"

Beatriz von der Esch,[†,‡] Johannes C. B. Dietschreit,[†,‡] Laurens D. M. Peters,[†] and Christian Ochsenfeld[*,†]

†*Chair of Theoretical Chemistry, Department of Chemistry, University of Munich (LMU), Butenandtstr. 7, D-81377 München, Germany*

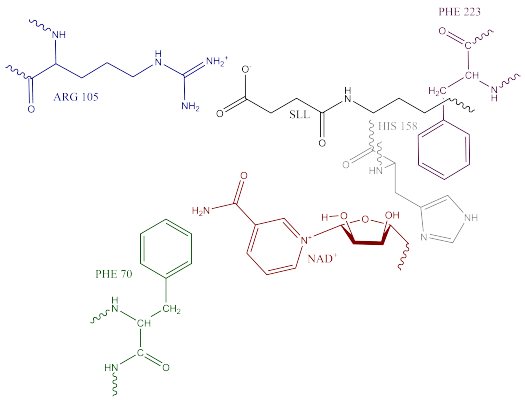
‡*These authors contributed equally to this work*

E-mail: christian.ochsenfeld@uni-muenchen.de

Visualisation

All images of molecular geometries were generated using *VMD*.[1] All plots were produced using the python-packages *matplotlib*[2] and *seaborn*. The chemical structures were drawn with *ChemDraw*.

QM-region

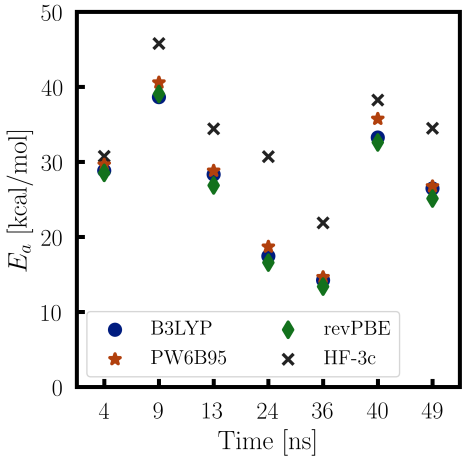


SFigure 1: Visualisation of the QM-region. The substrate SLL is shown in black and the co-factor NAD^+ in red. Additionally, four amino-acids and zero to four water molecules were included.

Besides the two reactants, SLL and the co-factor NAD+, four additional amino-acids and zero to four water molecule were included. Therefore, the number of atoms included in the QM-region varied from 139 to 152. The residues contained in the QM-region are shown in Figure 1. The substrate SLL is shown in black, NAD^+ in red, HIS 158 in grey, ARG 105 in blue, PHE 223 in magenta and PHE 70 in green. The residues were chosen based on proximity to the reactive centers.

Benchmark: HF-3c vs other functionals

To assess the accuracy of the HF-3c/minix for the description of the QM region, seven frames that covered a 25 kcal/mol range were compared to results obtained by higher theory methods. For those frames single point calculations were carried out for the educt and the transition state at the B3LYP-D3/def2-tzvp, revPBE-D3/def2-tzvp, and PW6B95-D3/def2-tzvp level of theory.[3–11] The functionals were selected because of their general use (B3LYP or revPBE) or because they were especially created for kinetic barriers (PW6B95). The activation barriers were calculated from the single point energies. The QM/MM partitioning and all interactions were treated as described in section "QM/MM Simulations".



SFigure 2: Comparison of predicted barrier heights based on the QM/MM adiabatic mapping paths generated with HF-3c. In all cases HF-3c is an upper limit to the barrier height, and thus it consistently overestimates the activation energy. The values on the x-axis show when the frames were picked from the MD-trajectory.

Figure 2 clearly shows that HF-3c is always proportional to the energy barriers estimated with the other methods and consistently overestimates the barrier height. This consistency allows us to use HF-3c/minix to distinguish frames higher and lower barriers, as we do not aim to use it in order to estimate a value comparable to experiment.

Machine Learning Model Comparison

Listed in the Table 1 are the results for the tested regression models. The numerical hyperparameters were determined using 5-fold cross validation. The MAE, RMSE, and R2 value were calculated using 3-fold cross validation.

STable 1: Summary of the tested machine learning models. The mean absolute error (MAE), the root-mean-squared-error (RMSE) and the R2 value for each model are listed. Besides these measures of performance the chosen hyperparameters are given.

Model	Hyperparameters	MAE [kcal/mol]	RMSE [kcal/mol]	R2
Linear Regression		4.28	5.41	-0.06
Descision Tree Regression	max depth=9	5.08	6.91	-0.54
Ridge Regression	$\alpha = 20$	3.57	4.46	0.28
Lasso Regression	$\alpha = 0.1$	3.71	4.59	0.23
Kernel Ridge Regression	$\alpha = 20$, kernel='linear'	3.55	4.44	0.28
Elastic Net Regression	$\alpha = 0.06$, l1 ratio=0.5	3.59	4.46	0.28

References

(1) Humphrey, W.; Dalke, A.; Schulten, K. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics* **1996**, *14*, 33–38.

(2) Hunter, J. D. Matplotlib: A 2D graphics environment. *Computing In Science & Engineering* **2007**, *9*, 90–95.

(3) Vosko, S.; Wilk, L.; Nusair, M. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Can. J. Phys.* **1980**, *58*, 1200–1211.

(4) 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*, 785–789.

(5) Becke, A. D. Densityfunctional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(6) Stephens, P.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.* **1994**, *98*, 11623–11627.

(7) Zhang, Y.; Yang, W. Comment on Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1998**, *80*, 890.

(8) Zhao, Y.; Truhlar, D. G. The Design of Density Functionals that are Broadly Accurate for Thermochemistry, Thermochemical Kinetics, and Nonbonded Interactions. *J. Phys. Chem. A* **2005**, *109*, 5656–5667.

(9) 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*, 3297–305.

(10) 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*.

(11) Sure, R.; Grimme, S. Corrected small basis set Hartree-Fock method for large systems. *J. Comput. Chem.* **2013**, *34*, 1672–1685.

3.4 Manuscript IV: QM/MM Free Energy Investigation of the Initial Step of the Desuccinylation Reaction Catalyzed by Sirtuin 5 Points Towards a Conserved Mechanism among Sirtuins

Johannes C. B. Dietschreit, Beatriz von der Esch, and Christian Ochsenfeld
"QM/MM Free Energy Investigation of the Initial Step of the Desuccinylation Reaction Catalyzed by Sirtuin 5 Points Towards a Conserved Mechanism among Sirtuins"
2020, *in preparation*

Abstract: Human Sirtuin 5 (Sirt5) catalyzes the NAD^+-dependent desuccinylation and demalonylation of lysine residues. While computational studies have so far mainly focused on the deacetylation reaction catalyzed by related sirtuins, the desuccinylation reaction mechanism remains computationally uncharacterized. As succession to our previous study (von der Esch et al., JCTC 2019), where we analyzed the first reaction step of the desuccinylation by means of QM/MM adiabatic mapping and machine learning, we use sophisticated Umbrella Sampling to compute the free energy reaction profile of the initial reaction step. The computational investigation leads to the conclusion that the NAD^+ transfer, the first step of the deacylation reaction, is highly conserved among all sirtuins and proceeds via an S_N2-type reaction mechanism in Sirt5. Further, difficulties when estimating free energy barriers via exponential averaging and limitations of free energy surface reweighting are discussed in detail.

The following manuscript has not yet been submitted to a peer reviewed journal.

QM/MM Free Energy Investigation of the Initial Step of the Desuccinylation Reaction Catalyzed by Sirtuin 5 Points Towards a Conserved Mechanism among Sirtuins

Johannes C. B. Dietschreit,[1,2,a)] Beatriz v. d. Esch,[1,2,a)] and Christian Ochsenfeld[1,2,b)]
[1)] *Chair of Theoretical Chemistry, Department of Chemistry, University of Munich (LMU), Butenandtstr. 7, D-81377 München, Germany*
[2)] *Center for Integrated Protein Science (CIPSM) at the Department of Chemistry, University of Munich (LMU), Butenandtstr. 5–13, D-81377 München, Germany*

(Dated: 1 December 2020)

Human Sirtuin 5 (Sirt5) catalyzes the NAD^+-dependent desuccinylation and demalonylation of lysine residues. While computational studies have so far mainly focused on the deacetylation reaction catalyzed by related sirtuins, the desuccinylation reaction mechanism remains computationally uncharacterized. As succession to our previous study (von der Esch et al., JCTC 2019), where we analyzed the first reaction step of the desuccinylation by means of QM/MM adiabatic mapping and machine learning, we use sophisticated Umbrella Sampling to compute the free energy reaction profile of the initial reaction step. The computational investigation leads to the conclusion that the NAD^+ transfer, the first step of the deacylation reaction, is highly conserved among all sirtuins and proceeds via an S_N2-type reaction mechanism in Sirt5. Further, difficulties when estimating free energy barriers via exponential averaging and limitations of free energy surface reweighting are discussed in detail.

Keywords: Sirt5, Reaction Pathway, Free Energy, QM/MM, Umbrella Sampling

I. INTRODUCTION

Post-translational modifications (PTMs) describe the chemical alteration of proteins after their expression. They greatly increase the variety of a cell's proteome by expanding the chemical space of the 20 canonical amino acids and play an important role in, for example, protein activity, cell signalling, or transcription.[1] A frequently modified residue is lysine. Best known is the interplay of lysine acetylation[2,3] and methylation[4,5] fixing its charge state to either neutral or positively charged, especially in histone tails.

Acetylation is one of the possible modifications subsumed under the group of ϵ-N-acylation of lysine. In humans there are 18 lysine deacylases (KDACs). They can be divided into four classes. Classes I, II, and IV are Zn^{2+}-dependent enzymes; their active site contains a catalytically active zinc ion. Class III KDACs, known as sirtuins, also contain Zn^{2+}, but they are NAD^+-dependent. The catalytic center is located next to an NAD^+-binding Rossmann-fold subdomain, whereas the zinc binding motif is spatially separated and ensures the structural integrity of the enzymes.[6] Sirtuins are the mammalian homologs of the silent information regulator 2 (Sir2), a highly conserved family of proteins found in archea and eucariots.[7,8] There are seven different sirtuin isoforms in mammals that cover a wide range of lysine deacylations. They not only catalyze lysine deactelyation, but also for example desuccinylation and demyristoylation.[9,10] In line with their wide range of catalytic activity, sirtuins can be found in several different cell compartments such as the nucleus or the mitochondria[11], where they are involved in various biological processes.[12,13]

This paper will focus on the catalytic activity of Sirt5, which shows no detectable deacetylation but rather demalonylation and desuccinylation activity.[14] It is located in the mitochondria and its main target is the carbamoyl phosphate synthetase 1 (CPS1).[15] Its active site consists of a hyrdophobic pocket with a positively charged arginine (Arg105) at the end. Together with Tyr102, those two residues position the negatively charged end of the dicarboxylic acid modification for removal. Sirt5 transfers succinyl (and malonyl) to its cosubstrate by cleaving the ribosyl bond in NAD^+ and thereby generating nicotinamide, a natural sirtuin inhibitor,[16,17] and a mixture of 2'- and 3'-O-succinyl-ADP-ribose.[14]

A mutagenesis study of the His116 in the active site, modification of NAD^+, and sirtuin crystal structures strongly suggest that no residue in the catalytic pocket takes actively part in the first step of the reaction,[18,19] namely the cleavage of the glycosidic bond between ribose and nicotinamide, and the addition of the substrate's amide carbonyl oxygen to ribose, forming an imminium adduct (henceforth called intermediate). Said intermediate was captured by using thioamide substrate analogs.[20,21] The NAD^+ exchange reaction can either proceed via an S_N1-like step-wise or an S_N2-like concerted mechanism. So far computational studies have only focused on the initial step of the deacetylation reaction in the bacterial sirtuin analogue Sir2Tm[22] and the yeast homolog yHst2[23]. Both concluded that the first step is very likely concerted.

In our previous publication we have analyzed the first

[a)] These two authors contributed equally to this work
[b)] Electronic mail: christian.ochsenfeld@uni-muenchen.de

reaction step catalyzed by Sirt5 by means of quantum mechanics / molecular mechanics (QM/MM).[24] We calculated minimal energy paths for the first reaction step by means of adiabatic mapping.[25] Adiabatic mapping calculations minimise the system while constraining a collective variable to a specific set of values (in our case the difference between the breaking glycosidic bond and the forming bond between the amide carbonyl oxygen and C1' of ribose). For these paths we used 150 different reactant configurations which were extracted from a MM-molecular dynamics (MD) simulation of the Sirt5-substrate complex solved in water. The study connected the configuration of the active site with the calculated activation energy by means of machine learning. We were able to identify interactions of the substrate (a succinylated peptide) and residues within the active site that could increase or decrease the activation barrier. Due to the complexity of the high-dimensional potential energy surface, the procedure drags the system from reactant to intermediate by visiting many local minima. This leads to a large scattering of the activation barrier as the minimised reactant geometries also correspond to many different local minima.

The effective, free energy activation barrier can best be estimated as exponential average from many of those minimal free energy barriers.[26] However, Ryde[27] has cautioned that one needs quite a large number of minimal energy activation barriers as the exponential average is ill-conditioned and converges very slowly. He has pointed out that many computational studies based on minimal energies have very large error bars so that their conclusions are questionable. Therefore, we study the actual potential of mean force (PMF) for this system as a function of the reaction coordinate. We use Umbrella Sampling[28] and the same QM/MM setup as in our previous study to explore important regions of configuration space and evaluate the free energy as a function of the reaction coordinate by means of Multistate Bennett's Acceptance Ratio (MBAR)[29].

This manuscript starts with a brief introduction into the difficulty of predicting effective energy barriers using the exponential averaging and then outlines the equations employed to compute the PMF based on QM/MM Umbrella Sampling calculations. After reviewing the computational details in Section III, the obtained PMF of the initial NAD^+ exchange reaction and the resulting free energy activation barrier are compared to the previously determined minimum energy path and exponentially averaged effective barrier. Section IV is concluded with a detailed discussion of free energy reweighting, which was used to determine the PMF at a higher level of theory than that employed for the Umbrella Sampling.

II. THEORY

A. The Problem of the Ill-conditioned Exponential Average

If the minimal energy activation barrier of the single adiabatic mapping path i is denoted with $\Delta E_i^\ddagger$, then the average activation barrier for n samples is

$$\left\langle \Delta E^\ddagger \right\rangle = \frac{1}{n} \sum_i^n \Delta E_i^\ddagger \tag{1}$$

and its variance

$$\sigma^2 = \left\langle \left(\Delta E^\ddagger\right)^2 \right\rangle - \left\langle \Delta E^\ddagger \right\rangle^2 \ . \tag{2}$$

The exponential average (EA) for this set of energies is then computed as

$$\Delta E^\ddagger_{\mathrm{EA,num}} = -\beta^{-1} \ln \left(\frac{1}{n} \sum_i^n e^{-\beta \Delta E_i^\ddagger} \right) \ , \tag{3}$$

where $\beta = 1/k_\mathrm{B}T$, with k_B being the Boltzmann constant and T the absolute temperature, which is fixed to 300 K within the scope of this work. As there are several local minima along each degree of freedom (DoF) orthogonal to the reaction coordinate into which the system is minimized, and the number of these DoF is very large in extended biomolecular systems, one can assume that the minimal energy reaction barriers are normal distributed based on the central limit theorem.[30] The exponential average of normal distributed reaction barriers can be calculated analytically using the arithmetic mean $\left\langle \Delta E^\ddagger \right\rangle$ and the variance σ^2.

$$\Delta E^\ddagger_{\mathrm{EA,ana}} = \left\langle \Delta E^\ddagger \right\rangle - \frac{1}{2} \beta \sigma^2 \tag{4}$$

Ryde[27] performed numerical experiments, drawing random numbers from a normal distribution and computed the EA using eqs. (3) and (4). Ryde found that he needed more than an exponentially increasing large number of samples for increasing σ to converge the exponential average within 95 % confidence of the known result. This slow convergence of the exponential average impedes also the computation of absolute free energies. Mean and variance converge much faster than the exponential average, and thus the analytical expression (eq. (4)) using the first and second moment of the underlying distribution is more robust, but can only be employed if the distribution of activation barriers is indeed Gaussian.

B. Multistate Bennett's Acceptance Ratio

Each Umbrella Simulation i (called umbrella window) is associated with a biasing potential B_i, which modifies the original Born-Oppenheimer QM/MM potential energy surface (PES) U_0 to

$$U_i = U_0 + B_i \ . \tag{5}$$

In order to recover the unbiased data, we used MBAR to estimate the (relative) free energies A_i of each window which were introduced by the biasing potentials B_i. The free energy A_i of one window is implicitly defined as a function of all simulation frames and all free energies

$$e^{-\beta A_i} = \sum_j^S \sum_k^{n_j} \frac{e^{-\beta U_i(j,k)}}{\sum_l^S n_l \; e^{\beta A_l - \beta U_l(j,k)}} \; , \qquad (6)$$

where S is the number of windows, n_i the number of frames in window i, and $U_i(j,k)$ the value of the potential energy function of window i for frame k from simulation window j. The same QM/MM potential U_0 function was used in every window, and thus equation eq. (6) can be simplified to the biasing potentials only, increasing its numerical stability.

$$e^{-\beta A_i} = \sum_j^S \sum_k^{n_j} \frac{e^{-\beta B_i(j,k)}}{\sum_l^S n_l \; e^{\beta A_l - \beta B_l(j,k)}} \qquad (7)$$

Eq. (7) has to be solved self-consistently, but can alternatively be recast into a minimization problem

$$g_i = n_i - \sum_j^S \sum_k^{n_j} \frac{n_i \; e^{\beta A_i - \beta B_i(j,k)}}{\sum_l^S n_l \; e^{\beta A_l - \beta B_l(j,k)}} = 0 \; , \qquad (8)$$

where all g_i's have to be zero for the exact solution. The unbiased free energy as a function of the collective variable ξ is recovered using

$$\beta A_0(\xi) = -\ln \sum_j^S \sum_k^{n_j} \frac{\delta\left(\xi(j,k) - \xi\right)}{\sum_l^S n_l \; e^{\beta A_l - \beta B_l(j,k)}} \; . \qquad (9)$$

The Dirac delta function is evaluated with finite resolution using an indicator function $\mathbf{1}_{\xi \in [\xi_{\min}, \xi_{\max}]}$, which is equal to one if $\xi \in [\xi_{\min}, \xi_{\max}]$ and otherwise zero. We refer to $\delta\xi = \xi_{\max} - \xi_{\min}$ as the bin width at which we compute the free energy surface.

C. Free Energy Reweighting

Due to the immense number of QM energy and force calculations necessary for QM/MM-MD simulations and the connected high computational cost, free energy reweighting is proposed as a mean to obtain the PMF at a higher level theory without having to perform additional MD simulations. The goal is to sample the system's configuration space using a cost effective method and to subsequently reweight the resulting PMF based on single point calculations at the desired level of theory.

The PMF $A_0(\xi)$ is associated with the QM/MM PES U_0, whereas we desire to know $A_1(\xi)$, the free energy surface corresponding the potential energy function U_1.

$$\beta A_1(\xi) = -\ln \sum_j^S \sum_k^{n_j} \frac{\delta\left(\xi(j,k) - \xi\right) \;\; e^{-\beta \Delta U(j,k)}}{\sum_l^S n_l \; e^{\beta A_l - \beta B_l(j,k)}} \; , \qquad (10)$$

where $\Delta U = U_1 - U_0$. As the important regions within configuration space differ between the two PES, the Boltzmann weights for all frames will not be uniform. Therefore, we define the reweighting entropy in accordance with Li et al.[31] to estimate the fraction of frames lost for each value of ξ as

$$\mathcal{S}(\xi) = -\frac{\sum_j^S \sum_k^{n_j} \delta\left(\xi(j,k) - \xi\right) P(j,k,\xi) \ln P(j,k,\xi)}{\ln\left(\sum_j^S \sum_k^{n_j} \delta\left(\xi(j,k) - \xi\right)\right)} \qquad (11)$$

and

$$P(j,k,\xi) = e^{\beta A_1(\xi)} \frac{e^{-\beta \Delta U(j,k)}}{\sum_l^S n_l e^{\beta A_l - \beta B_l(j,k)}} \; . \qquad (12)$$

$\mathcal{S}(\xi)$ is bounded to [0,1] and assumes its maximum of 1 if all frames within bin ξ have equal reweighting probabilities. $\mathcal{S}(\xi)$ approaches 0 if very few frames dominate the reweighting probability. Similar information can be obtained by determining the maximum value of the reweighting probability ($P_{\max}(\xi)$) for each bin.

III. COMPUTATIONAL DETAILS

A. General QM/MM Setup

As reference geometries for the umbrella simulations, we used the adiabatic mapping path with the lowest activation barrier from our previous study[24]. We chose that frame in order to prove that the barrier is underestimated due to the minimizer identifying a local minimum with a high energy as reactant rather than the lower basin containing most reactant configurations. The same protein residues within the active site, namely Arg105, His158, Phe170, and Phe223, as well as succinyl-lysine (SLL) and the ribose-nictonamide part of NAD^+ were included in the QM region (113 atoms in total). We only modified the location of the QM/MM border, avoiding a cut through the peptide bonds along the protein backbone and placed it between C_α and C_β.

The QM region was described with HF-3c/minix[32], which has been shown to yield accurate chemistry but elevated energies for transition states[24]. The activation free energy will therefore be higher than one computed with a higher level method, but we expect that S_N1 and S_N2 can be correctly discerned nonetheless. The MM parameters for all standard protein residues were taken from AmberFF14[33], those for NAD^+ from the AMBER parameter database[34,35]. SLL was described with GAFF[36] parameters and AM1-BCC[37] charges. For the zinc finger we used ZAFF[38] parameters. The employed water model was TIP3P.[39] For the full original MM setup see von der Esch et al.[24] The QM/MM calculations were performed with our in-house program suite FermiONs++[40–42] which uses the OpenMM 7.3 library[43–45] to evaluate the MM subsystem.

B. Details of the QM/MM Umbrella Simulations

For the restrained QM/MM-MD simulations we used a Python interface for FermiONs++, which allows low-level access to the QM engine. The propagation of atomic coordinates, application of a thermostat, and evaluation of the umbrella potential was done within Python, only for the QM/MM energy and gradient evaluations the PyFermiONsInterface was used. The MD simulations were started from structures taken from the previously obtained minimal energy path with the lowest barrier height.[24] All residues within 10 Å of the QM subsystem were chosen to be active, thereby ensuring that there is always a layer of frozen atoms enclosing the active atoms. This ensures that no molecule can escape into the vacuum surrounding the simulation box, as no periodic boundary conditions were employed.

In order to distinguish between S_N1 and S_N2 reaction type, the sampling was conducted along two dimensions, the breaking C1'-N bond between ribose and nicotinamide and the forming bond O-C1' between the carbonyl oxygen and ribose. Hence, each umbrella window i was biased with two harmonic functions

$$B_i(\mathbf{x}) = \frac{1}{2}k_{1,i}(d_1 - d_{1,i})^2 + \frac{1}{2}k_{2,i}(d_2 - d_{2,i})^2 \ , \quad (13)$$

with $\mathbf{x}$ being a point in configuration space, $d_1 = d(\mathrm{O} - \mathrm{C1}')$, $d_2 = d(\mathrm{C1}' - \mathrm{N})$, as well as $k_{j,i}$ and $d_{j,i}$ being the force constant and equilibrium bond length of the respective bond in the biasing potential i. The force constants range from 200 to 600 kJ mol^{-1} Å^{-2}, adapting to the slope of the local PES. In total, 65 umbrella simulations were carried out. The force constants are shown pictorially in Fig. 6.

Each umbrella window simulation consists of three parts: i) heating, ii) equilibration, and iii) production. The system was propagated using the velocity Verlet algorithm[46] and the temperature was controlled using the Langevin thermostat[47]. The initial forces assigned to the active atoms where randomly chosen from the Maxwell-Boltzmann distribution at 1 K. During heating the time step was set to 0.1 fs and no thermostat was used. Every 10 time steps the velocities were rescaled in 1 K increments, reaching 300 K after 3000 time steps.

For equilibration and production, the time step was set to 0.5 fs and the Langevin friction constant to 1 ps^{-1}. For increased speed and stability, we used the fully converged extended Lagrangian method[48] implemented in FermiONs++[49]. The equilibration period was 1 ps long. The production runs were at least 10 ps and a maximum of 20 ps long. Simulations were terminated before the 20 ps limit, if the Mann-Kendall[50–52] test indicated that the mean of the two biased bond lengths had converged and therefore equilibrium within the window had been reached. Outputs were written every 2 fs.

C. Reweighting Simulations

To calculate the potential energy for frames at a different level, a similar Python setup as for the MD was used. The coordinates from the previously computed trajectories were employed by FermiONs++ for single point calculations. We computed B3LYP-D3/def2-SVP[53–58] and PBEh-3c/def2-mSVP[59] QM/MM energies.

D. MBAR Analysis

As only the relative values of the A_i's calculated with MBAR of each umbrella window are meaningful, the free energy of the first window is set to zero. The starting guess is zero for all windows. Eq. (7) was solved self-consistently, suggested minimization algorithms such as Newton-Raphson[29] or DIIS[60] did not improve the results. At the end of each self-consistent cycle the largest change in βA_i was determined and convergence was reached when it dropped under 10^{-7} and the norm of $\mathbf{g}^{\mathrm{T}} = (g_1, g_2, \ldots, g_S)$ (g_i's are defined in eq. (8)) was below 10^{-4}, ensuring that a stable minimum had been found.

The numerical errors of each bin were computed via bootstrapping[61] analysis. 10 bootstraping runs were performed, drawing random frames from each simulation with replacing and then performing 10 additional MBAR analyses. The standard deviation between the bootstrap samples of the free energy within each bin was used as statistical error estimate.

IV. RESULTS AND DISCUSSION

A. Free Energy Surface of the Initial Reaction Step

The umbrella sampling method allows for easy parallelization during the exploration of the free energy surface. However, we still performed these simulations in consecutive batches, filling in gaps between sampled areas that have been left by the previous set of simulations. In total 65 umbrella simulations were carried out.

The algorithms WHAM[62] or MBAR[29], which is often also called binless WHAM, assume that the input data describe the simulated system in equilibrium and that they are uncorrelated. We calculated the decorrelation times of the biasing potential of each umbrella window, the mean was 23 fs. Hence, the statistical inefficiency[29,63] was 47 fs. Based on these findings, we used data 40 fs apart to construct the free energy surface. For completeness, results based on the full data set can be found in the Appendix (Fig. 5). After determining the relative free energies A_i, we used a bin width of 0.075 Å for both bond lengths to evaluate eq. (9) (see Fig. 1A).

To obtain the free energy along a one-dimensional reaction coordinate for the nucleophilic substitution,

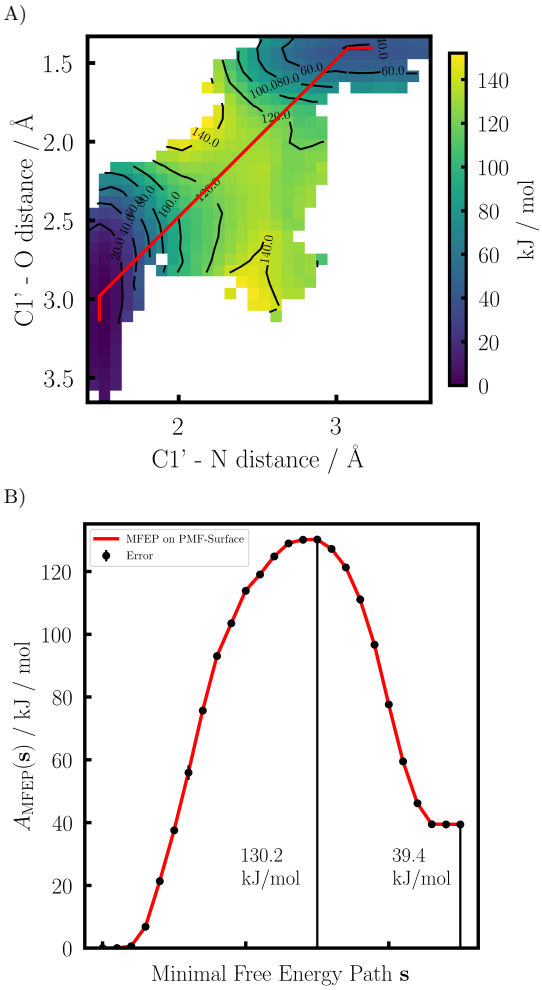


FIG. 1. A) Free energy surface of the first reaction step catalyzed by Sirt5 calculated with HF-3c/MM. The minimal free energy path (MFEP) connecting the reactant and intermediate state is shown in red. White areas were not visited during the simulations. B) The free energy profile along the MFEP (most likely reaction path), corresponding to the red line in A). The difference of well depths and barrier height along the MFEP are given explicitly.

we used Dijkstra's algorithm[64] to find the lowest energy path (Fig. 1B corresponding to the red line in Fig. 1A) connecting the lowest point of the reactant basin ($d(\mathrm{C1'-N}) < 2$ Å and $d(\mathrm{O-C1'}) > 2.5$ Å) with the lowest point of the intermediate basin ($d(\mathrm{C1'-N}) > 2.75$ Å and $d(\mathrm{O-C1'}) < 1.75$ Å).

The position on the free energy surface of the line connecting the educt and product of the investigated reaction step very clearly indicates a concerted mechanism. The energy changes first very little as the carbonyl oxygen approaches, but then the shortening of the (C1'-O)-bond length is directly proportional to the elongation of the (C1'-N)-bond in NAD^+. After the new bond has been formed, the energy decreases slightly further by nicotinamide moving away from the ribose. The reaction mechanism is therefore of S_N2 type disregarding of whether sirtuins catalyze a deacetylation or desuccinylation. The changes within the active site that lead to the different substrate specificity of the seven sirtuins do not change the overall conserved reaction mechanism.

B. Free Energy Paths vs. Minimal Energy Paths

The adiabatic mapping path that provided the starting configurations for the umbrella windows on the $d(\mathrm{O-C1'})$-$d(\mathrm{C1'-N})$-surface, had predicted an activation energy of 91.6 kJ/mol, which is around 40 kJ/mol smaller than the computed free energy barrier along the MFEP (130.2 kJ/mol). The low minimal energy barrier is very likely caused by the path starting off in a local minimum that is already much higher in energy than the majority of configurations forming the reactant basin. This result strongly suggests that predictions of reaction barriers or even reaction mechanisms based on minimal energy paths can be misleading, as has already been hinted by the strong scattering of minimal reaction barrier values in our previous paper.[24] The exponentially averaged barrier, $\Delta E^{\ddagger}_{\mathrm{EA,num}}$, which combines all 150 paths, is also lower than the free energy barrier (see Tab. I). Based on Ryde's results the numerical exponential average has, because of the large variance, an 95 % confidence interval of roughly 2000 kJ/mol.

TABLE I. The numerical results of our previous machine learning focused study on the reaction barriers of the first reaction are summarized by their mean ($\langle \Delta E^{\ddagger} \rangle$, eq. (1)), standard deviation (σ, eq. (2)), numerical exponential average ($\Delta E^{\ddagger}_{\mathrm{EA,num}}$, eq. (3)), exponential average assuming a Gaussian distribution ($\Delta E^{\ddagger}_{\mathrm{EA,ana}}$, eq. (4)), and the width of the 95 % confidence interval ($\Delta\Delta E^{\ddagger}_{\mathrm{EA,ana}}$). All numbers are given in kJ/mol. The values of $\Delta\Delta E^{\ddagger}_{\mathrm{EA,ana}}$ are estimated based on the results given in Ref. 27.

data set	$\langle \Delta E^{\ddagger} \rangle$	σ	$\Delta E^{\ddagger}_{\mathrm{EA,num}}$	$\Delta E^{\ddagger}_{\mathrm{EA,ana}}$	$\Delta\Delta E^{\ddagger}_{\mathrm{EA,ana}}$
150 paths	157.4	22.9	104.0	52.0	20
ML	157.3	13.4	138.4	120.8	3

The analytical EA, $\Delta E^{\ddagger}_{\mathrm{EA,ana}}$, is even lower than $\Delta E^{\ddagger}_{\mathrm{EA,num}}$, which is due to the large scattering of the computed barriers (large variance). The fact that the distribution of the 150 frames is bi-modal calls the applicability of the analytical formula, which assumes a normal distribution, into question. The distribution of energy barriers predicted by our ML model on the other hand is uni-modal and more narrow, as the fit underestimates

high and overestimates low barriers. Its EA result, as given in Tab. I, is much closer to the free energy barrier based on umbrella sampling. The low-dimensional ML model cannot incorporate the many DoF orthogonal to the reaction and effectively averages over them, yielding, to some degree surprisingly, a more realistic barrier estimate.

C. Going from HF-3c to B3LYP

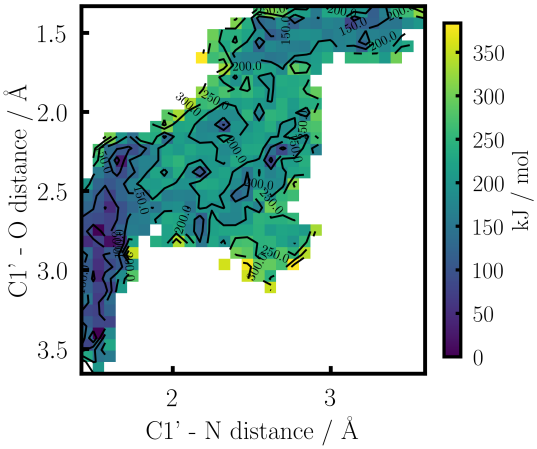


FIG. 2. QM/MM free energy at the level of B3LYP-D3/def2-SVP//MM reweighted from HF-3c//MM. The free energy surface is very uneven due to the large loss of information shown in Fig. 3.

After having obtained the PMF at the level of HF-3c/MM, we wanted to gain access to a PMF corresponding to a higher level of QM theory, for example, B3LYP-D3/def2-SVP, without having to explicitly simulate the system again on the more accurate and computationally more expensive PES. Therefore, we computed the potential energy for all 26,189 MD frames that were used to generate Fig. 1 with B3LYP-D3/def2-SVP//MM, and used all of them to reweight the PMF via eq. 10. The result, as seen in Fig. 2, is very uneven and does not resemble the HF-3c surface. To understand the cause of the erratic result, we computed the reweighting entropy (see Fig. 3). The analysis of the reweighting entropy showed that in every bin most of the sampled information was lost as one or a few frames had significantly larger Boltzmann weights than all other frames. This imbalance of weights constitutes a huge loss of sampling information and explains the noisy result of Fig. 2. It had been expected that the entropy would stay below 1, as some HF-3c configurations are more similar to those generated by B3LYP than others, but not that it is close to zero in nearly all bins.

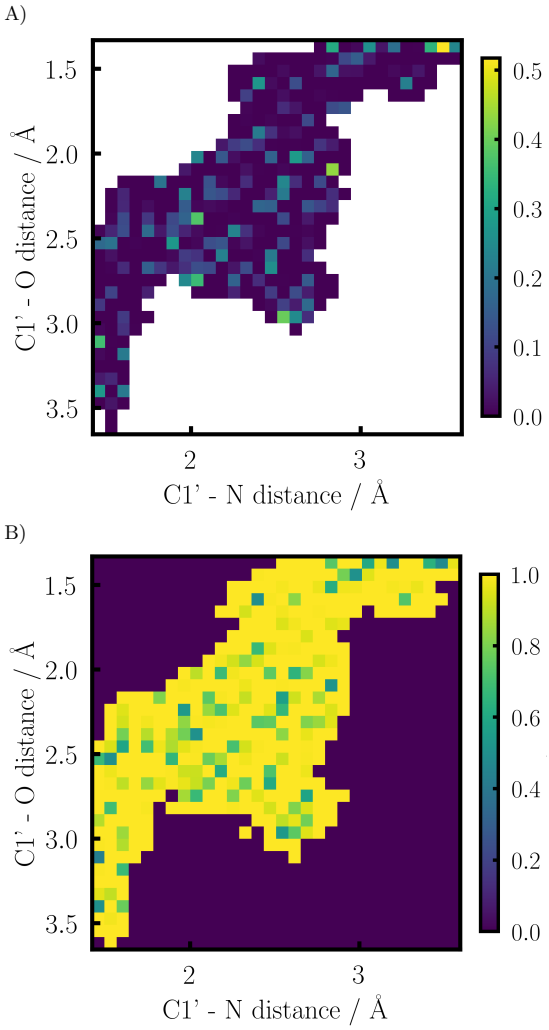


FIG. 3. A) Reweighting entropy and B) largest value of the reweighting probability $P_{\max}$ in each bin. Low entropy and large probability values indicate that one or few frames dominated all other frames within the bin, which is the case for nearly every bin. This explains the shape of the free energy surface shown in Fig. 10

We also tested reweighting to PBEh-3c/def2-mSVP in the reactant basin, as we expected PBEh-3c to be closer related to HF-3c, and therefore the loss of information to be less severe. The result is similarly noisy as those obtained for B3LYP.

To understand the reason for the large difference in reweighting weights, the umbrella window located at $d(\mathrm{C1'-N}) = 1.7$ Å / $d(\mathrm{O-C1'}) = 3.3$ Å was also simulated with PBEh-3c instead of HF-3c. This enabled us to

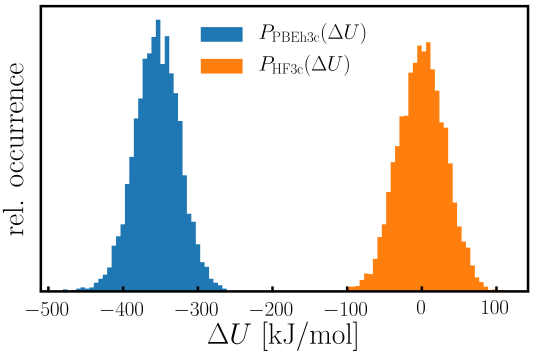


FIG. 4. Distribution of $\Delta U = U_{\mathrm{PBEh-3c}} - U_{\mathrm{HF-3c}}$ for the umbrella window $d(\mathrm{C1'-N}) = 1.7$ Å / $d(\mathrm{O-C1'}) = 3.3$ Å. The values of ΔU were shifted such that the mean of $P_{\mathrm{HF-3c}}(\Delta U)$ coincides with 0 kJ/mol. The two distributions $P_{\mathrm{HF-3c}}(\Delta U)$ and $P_{\mathrm{PBEh-3c}}(\Delta U)$ have no overlap as their respective means are about 350 kJ/mol apart, making the calculation of the free energy difference between the two QM methods impossible.

perform forward and backward free energy perturbations (FEP)[65] calculating the change in free energy for switching from one PES to the other. The results of the forward and backward perturbations differed by more than 300 kJ/mol and BAR[66] did not yield a reasonable result.

These large discrepancies of the FEP results can be explained when looking at the distribution of the potential energy difference based on either a HF-3c or a PBEh-3c simulation. Fig. 4 shows the two distributions of the energy difference $\Delta U = U_{\mathrm{PBEh-3c}} - U_{\mathrm{HF-3c}}$. For any free energy difference algorithm to work, the two distributions have to have some overlap, ideally they superimpose. It follows that the configurations spaces of HF-3c and PBEh-3c are significantly different for the chosen QM size, such that not even the general free energy difference for changing the PES can be calculated, let alone reweighting many small bins on a PMF. In general, the problem can be mitigated by decreasing the QM size and thereby making the PESs more similar.

However, even for the minimal QM region necessary for this system (SLL, ribose, nicotinamide, and Arg105, all in all 82 atoms) the two distributions are closer together but still have no overlap. Reweighting from PBEh-3c to B3LYP-D3/def2-SVP is not possible either for the minimal 82 atom-QM region (see Appendix C). It remains to be tested, what the maximal QM size is that would allow such a perturbation. Using alchemical intermediates, as they are often employed in MM simulations to connect two states with significantly different important configuration spaces, is not an economical option, since it would come at the cost of the low and high level QM method.

V. CONCLUSION AND OUTLOOK

Through computation of the PMF by means of QM/MM-MD simulations and subsequent evaluation using MBAR we have characterized the inital step of the dessucciylation reaction catalyzed by Sirt5. Our results indicate that analogously to the first step of the deacetylation reaction the NAD^+ transfer step of the dessucciylation reaction is of S_N2 type. This suggest that the differences in the active site, which give rise to varying substrate specificities within the sirtuin enzyme family, do not change the reaction mechanism.

The computation of the free energy reaction profile (minimal free energy path connecting reactant and intermediate) allowed us to evaluate the quality of free energy activation barriers estimated by means of exponential averaging. It was shown that the previously computed barrier based on 150 adiabatic mapping pathways underestimated the computed free energy barrier. This calls generally the reliability of reaction barriers and mechanisms based on minimal energy paths into question.

In order to provide the free energy surface at a higher level of theory reweighting was explored. However, when reweighting from HF-3c/minix to B3LYP-D3/def2-SVP and to PBEh-3c/def2-mSVP very noisy surfaces are obtained. It was found, by analysis of the reweighting entropy, that the reweighting process is dominated by a few frames, which leads to a huge loss of information and therefore erratic surfaces. Furthermore, it was shown that there is no overlap of the ΔU-distributions, which is the root cause for the reweighting not to work. Even when reducing the QM subsystem to the bare minimum for this system, reweighting remains impossible.

Currently, higher level PMFs remain inaccessible for extended biological systems. Future work has to explore the limiting QM size and the similarity of different levels of approximation.

The first of several desuccinylation reaction steps has now been shown to be independent of sirtuin specificity. A following study has to identify the exact mechanism of the remaining steps.

ACKNOWLEDGMENTS

Financial support was provided by “Deutsche Forschungsgemeinschaft” (DFG, German Research Foundation) - SFB 1309-325871075 “Chemical Biology of Epigenetic Modifications”. C.O. acknowledges further support as Max-Planck-Fellow at the MPI-FKF Stuttgart.

REFERENCES

[1] V. Uversky, in *Brenner's Encyclopedia of Genetics (Second Edition)*, edited by S. Maloy and K. Hughes (Academic Press, San Diego, 2013) second edition ed., pp. 425 – 430.

[2]C. Choudhary, B. T. Weinert, Y. Nishida, E. Verdin, and M. Mann, Nat. Rev. Mol. Cell Biol. **15**, 536–550 (2014).
[3]M. Schiedel and S. J. Conway, Curr. Opin. Chem. Biol. **45**, 166 (2018).
[4]K. Zhang and S. Y. Dent, J. Cell Biochem. **96**, 1137–1148 (2005).
[5]E. L. Greer and Y. Shi, Nat. Rev. Genet. **13**, 343 (2012).
[6]B. D. Sanders, B. Jackson, and R. Marmorstein, Biochim. Biophys. Acta Protein Proteonomics **1804**, 1604 (2010).
[7]R. A. Frye, Biochem. Biophys. Res. Commun. **260**, 273 (1999).
[8]R. A. Frye, Biochem. Biophys. Res. Commun. **273**, 793 (2000).
[9]J. Schemies, U. Uciechowska, W. Sippl, and M. Jung, Med. Res. Rev. **30**, 861 (2010).
[10]M. Schiedel, D. Robaa, T. Rumpf, W. Sippl, and M. Jung, Med. Res. Rev. **38**, 147 (2018).
[11]W. Dang, Drug Discov Today Technol **12**, e9 (2014), NIHMS150003.
[12]S. Michan and D. Sinclair, Biochem. J. **404**, 1 (2007).
[13]A. Chalkiadaki and L. Guarente, Nat. Rev. Canc. **15**, 608 (2015).
[14]J. Du, Y. Zhou, X. Su, J. J. Yu, S. Khan, H. Jiang, J. Kim, J. Woo, J. H. Kim, B. H. Choi, B. He, W. Chen, S. Zhang, R. A. Cerione, J. Auwerx, Q. Hao, and H. Lin, Science **334**, 806 (2011), NIHMS150003.
[15]R. H. Houtkooper, E. Pirinen, and J. Auwerx, Nature Reviews Molecular Cell Biology **13**, 225–238 (2012).
[16]K. J. Bitterman, R. M. Anderson, H. Y. Cohen, M. Latorre-Esteves, and D. A. Sinclair, J. Biol. Chem. **277**, 45099–45107 (2002).
[17]M. T. Schmidt, B. C. Smith, M. D. Jackson, and J. M. Denu, J. Biol. Chem. **279**, 40122–40129 (2004).
[18]J. Min, J. Landry, R. Sternglanz, and R.-M. Xu, Cell **105**, 269 (2001).
[19]M. D. Jackson, M. T. Schmidt, N. J. Oppenheimer, and J. M. Denu, Journal of Biological Chemistry **278**, 50985 (2003).
[20]Y. Zhou, H. Zhang, B. He, J. Du, H. Lin, R. A. Cerione, and Q. Hao, J. Biol. Chem. **287**, 28307 (2012).
[21]Y. Wang, Y. M. E. Fung, W. Zhang, B. He, M. W. H. Chung, J. Jin, J. Hu, H. Lin, and Q. Hao, Cell Chem. Biol. **24**, 339 (2017).
[22]P. Hu, S. Wang, and Y. Zhang, J. Am. Chem. Soc. **130**, 16721 (2008).
[23]Z. Liang, T. Shi, S. Ouyang, H. Li, K. Yu, W. Zhu, C. Luo, and H. Jiang, J. Phys. Chem. B **114**, 11927 (2010).
[24]B. von der Esch, J. C. B. Dietschreit, L. D. M. Peters, and C. Ochsenfeld, J. Chem. Theory Comput. **15**, 6660 (2019).
[25]K. E. Ranaghan and A. J. Mulholland, Int. Rev. Phys. Chem. **29**, 65 (2010).
[26]A. M. Cooper and J. Kästner, ChemPhysChem **15**, 3264 (2014).
[27]U. Ryde, J. Chem. Theory Comput. **13**, 5745 (2017).
[28]G. M. Torrie and J. P. Valleau, J. Comput. Phys. **23**, 187 (1977).
[29]M. R. Shirts and J. D. Chodera, J. Chem. Phys. **129**, 1 (2008), arXiv:0801.1426.
[30]Y. Dodge, ed., "Central limit theorem," in *The Concise Encyclopedia of Statistics* (Springer New York, New York, NY, 2008) pp. 66–68.
[31]P. Li, X. Jia, X. Pan, Y. Shao, and Y. Mei, J. Chem. Theory Comput. **14**, 5583 (2018).
[32]R. Sure and S. Grimme, J. Comput. Chem. **34**, 1672 (2013).
[33]J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, J. Chem. Theory Comput. **11**, 3696 (2015).
[34]J. J. Pavelites, J. Gao, P. A. Bash, and A. D. MacKerell, J. Comput. Chem. **18**, 221 (1997).
[35]R. C. Walker, M. M. De Souza, I. P. Mercer, I. R. Gould, and D. R. Klug, J. Phys. Chem. B **106**, 11658 (2002).
[36]J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, J. Comput. Chem. **25**, 1157 (2004), arXiv:z0024.
[37]A. Jakalian, B. L. Bush, D. B. Jack, and C. I. Bayly, J. Comput. Chem. **21**, 132 (2000).
[38]M. B. Peters, Y. Yang, B. Wang, L. Füsti-Molnár, M. N. Weaver, and K. M. Jr. Merz, J. Chem. Theory Comput. **6**, 2935 (2010), arXiv:NIHMS150003.
[39]W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, J. Chem. Phys. **79**, 926 (1983).
[40]J. Kussmann and C. Ochsenfeld, J. Chem. Phys. **138**, 134114 (2013).
[41]J. Kussmann and C. Ochsenfeld, J. Chem. Theory Comput. **11**, 918 (2015).
[42]J. Kussmann and C. Ochsenfeld, J. Chem. Theory Comput. **13**, 3153–3159 (2017).
[43]M. S. Friedrichs, P. Eastman, V. Vaidyanathan, M. Houston, S. LeGrand, A. L. Beberg, D. L. Ensign, C. M. Bruns, and V. S. Pande, J. Comp. Chem. **30**, 864 (2009).
[44]V. Pande and P. Eastman, "Computing in Science & Engineering" **12**, 34 (2010).
[45]P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks, and V. S. Pande, PLOS Computational Biology **13**, 1 (2017).
[46]W. C. Swope, H. C. Andersen, P. H. Berens, and K. R. Wilson, The Journal of Chemical Physics **76**, 637 (1982).
[47]M. Kröger, *Models for Polymeric and Anisotropic Liquids*, 1st ed., 675 (Springer-Verlag Berlin Heidelberg, 2005).
[48]A. M. N. Niklasson, P. Steneteg, A. Odell, N. Bock, M. Challacombe, C. H. Tymczak, E. Holmström, G. Zheng, and V. Weber, J. Chem. Phys. **130**, 214109 (2009).
[49]L. D. M. Peters, J. Kussmann, and C. Ochsenfeld, J. Chem. Theory Comput. **13**, 5479 (2017).
[50]H. Mann, Econometrica **13**, 163 (1945).
[51]M. Kendall, *Rank Correlation Methods*, 4th ed. (Charles Griffin, 1975).
[52]R. Gilbert, *Statistical Methods for Environmental Pollution Monitoring* (Wiley, 1987).
[53]A. D. Becke, J. Chem. Phys. **98**, 5648 (1993).
[54]C. Lee, W. Yang, and R. G. Parr, Phys. Rev. B **37**, 785 (1988).
[55]S. Vosko, L. Wilk, and M. Nusair, Can. J. Phys. **58**, 1200 (1980).
[56]P. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch, J. Phys. Chem. **98**, 11623 (1994).
[57]S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, J. Chem. Phys. **132** (2010), 10.1063/1.3382344.
[58]F. Weigend and R. Ahlrichs, Phys. Chem. Chem. Phys. **7**, 3297 (2005).
[59]S. Grimme, J. G. Brandenburg, C. Bannwarth, and A. Hansen, J. Chem. Phys. **143**, 054107 (2015).
[60]C. Zhang, C.-L. Lai, and B. M. Pettitt, Mol. Simul. **42**, 1079 (2016).
[61]B. Efron, Ann. Statist. **7**, 1 (1979).
[62]S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, J. Comput. Chem. **13**, 1011 (1992).
[63]W. Janke, in *Quantum Simulations Complex Many-Body Syst. From Theory to Algorithms*, Vol. 10, edited by J. Grotendorst, D. Marx, and A. Muramatsu (John von Neumann Institute for Computing, Jülich, 2002) pp. 423–445.
[64]E. W. Dijkstra, Numerische Mathematik **1**, 269 (1959).
[65]R. W. Zwanzig, J. Chem. Phys. **22**, 1420 (1954).
[66]C. H. Bennett, J. Comp. Phys. **22**, 245 (1976).

Appendix A: Influence of Bin Width and Sample Number

The influence of bin size on the free activation energy as well as the location of the minimal free energy path connecting the two minima on the surface was tested. Figure 5 indicates that there is no influence. Comparison with Fig. 1 shows that inclusion of additional correlated data points (closer together than 40 fs) does not improve or change the result either.

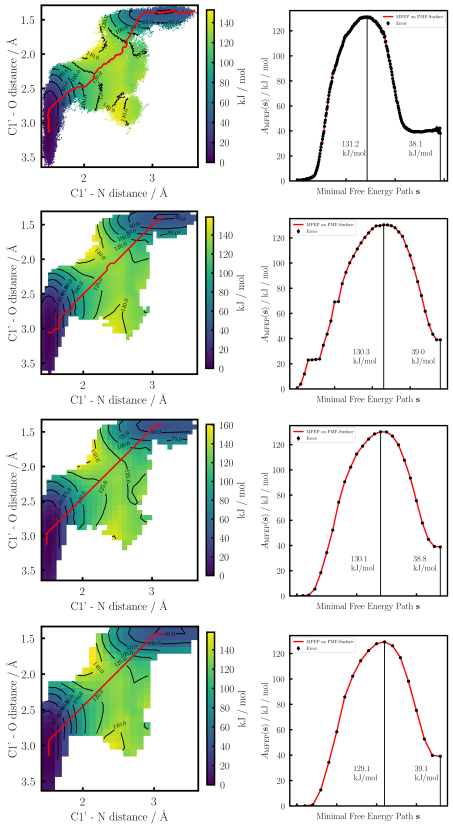


FIG. 5. All plots are based on the full data set (data points are 2 fs apart). The bin sizes used for the surfaces in the left column are the same along $d(\mathrm{O}-\mathrm{C1}')$ and $d(\mathrm{C1}'-\mathrm{N})$. The sizes are from top to bottom 0.01 Å, 0.05 Å, 0.075 Å, and 0.1 Å, respectively. The right column contains 1-D plots of the minimal energy path connecting the two minima on the free energy surface on the left. The difference of well depths and the barrier along the MFEP are given in each plot. The value of the free energy along the MFEP is clearly independent of bin size.

Appendix B: Deviation of Umbrella Window Mean from Bias Potential Minimum Position

Figure 6 visualizes the deviation of the mean along $d(\mathrm{C1}'-\mathrm{N})$ and $d(\mathrm{O}-\mathrm{C1}')$ within each umbrella window and the minimum of the biasing potential. Windows placed near the high energy transition state region or in one of the basins (either reactant or intermediate) show very little deviations between intended window mean (arrow base) and the computed mean (arrow tip). Windows placed in regions, where the PMF changes rapidly, deviate more strongly even if large force constants have been used. This is due to the overestimation of the transition barrier energy by HF-3c and corresponding large forces. In contrast, much lower force constants (160 kJ mol^{-1} Å^{-2}) were used by Hu et al.[22] for the umbrella simulations of Sir2Tm, where they employed B3LYP/6-31G* and calculated a free energy barrier of only 65.8 kJ/mol.

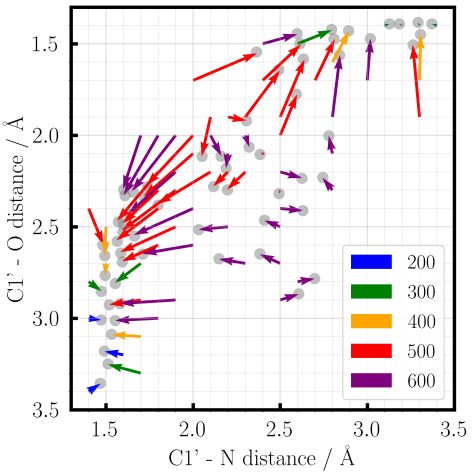


FIG. 6. The origin of each arrow indicates the original window placement, and therefore the center of each biasing potential $d_{j,i}$. The arrow's color corresponds to the force constant in kJ mol^{-1} Å^{-2}. The arrow head points to the mean d(C1'-O)/d(C1'-N) sampled in each umbrella simulation.

Appendix C: Reweighting from PBEh-3c to B3LYP

As PBEh-3c/def2-mSVP and B3LYP-D3/def2-SVP are both DFT functionals combined with a similar basis set, it was tested whether a reweighting from PBEh-3c to B3LYP was possible. This reweighting is not possible either, as a test performed for the same umbrella window as in the main text ($d(\mathrm{C1}'-\mathrm{N})$ = 1.7 Å / $d(\mathrm{O}-\mathrm{C1}')$ = 3.3 Å) proves, see Fig. 7.

At this point, it is not clear whether the perturbation from one functional to the other is prohibited by the accumulation of small difference over the entire QM region or whether there exist a few atoms within the system which are the main cause.

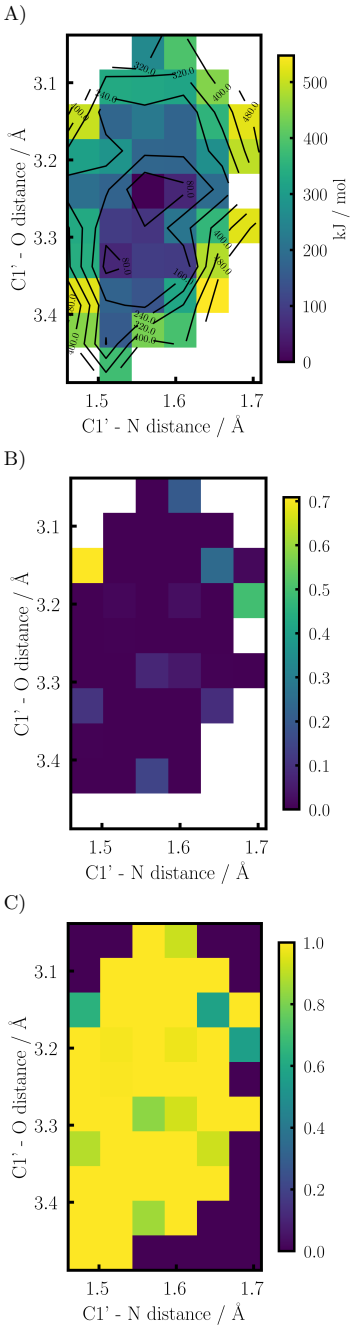


FIG. 7. A) Reweighted free energy from PBEh-3c to B3LYP-D3, B) reweighting entropy, $\mathcal{S}$, and C) largest value of the reweighting probability $P_{\max}$ in each bin. As in Fig. 3, the entropy and $P_{\max}$ clearly show the total dominance of one or few frames per bin.

3.5 Publication V: Predicting ^{19}F NMR chemical shifts: A combined computational and experimental study of a trypanosomal oxidoreductase-inhibitor complex

Johannes C. B. Dietschreit, A. Wagner, T. A. Le, P. Klein, H. Schindelin, T. Opatz, B. Engels, U. A. Hellmich, and C. Ochsenfeld

"Predicting ^{19}F NMR chemical shifts: A combined computational and experimental study of a trypanosomal oxidoreductase-inhibitor complex"
Angew. Chem. Int. Ed. **2020**, *56*, 12669-12673

Abstract: The absence of fluorine from most biomolecules renders it an excellent probe for NMR spectroscopy to monitor inhibitor–protein interactions. However, predicting the binding mode of a fluorinated ligand from a chemical shift (or *vice versa*) has been challenging due to the high electron density of the fluorine atom. Nonetheless, reliable ^{19}F chemical-shift predictions to deduce ligand-binding modes hold great potential for *in silico* drug design. Herein, we present a systematic QM/MM study to predict the ^{19}F NMR chemical shifts of a covalently bound fluorinated inhibitor to the essential oxidoreductase tryparedoxin (Tpx) from African trypanosomes, the causative agent of African sleeping sickness. We include many protein–inhibitor conformations as well as monomeric and dimeric inhibitor–protein complexes, thus rendering it the largest computational study on chemical shifts of 19F nuclei in a biological context to date. Our predicted shifts agree well with those obtained experimentally and pave the way for future work in this area.

https://doi.org/10.1002/anie.202000539

GDCh Communications Angewandte International Edition Chemie

How to cite: *Angew. Chem. Int. Ed.* **2020**, *59*, 12669–12673
International Edition: doi.org/10.1002/anie.202000539
German Edition: doi.org/10.1002/ange.202000539

NMR Spectroscopy

Predicting ^{19}F NMR Chemical Shifts: A Combined Computational and Experimental Study of a Trypanosomal Oxidoreductase–Inhibitor Complex

Johannes C. B. Dietschreit, Annika Wagner, T. Anh Le, Philipp Klein, Hermann Schindelin, Till Opatz, Bernd Engels, Ute A. Hellmich, and Christian Ochsenfeld**

Abstract: *The absence of fluorine from most biomolecules renders it an excellent probe for NMR spectroscopy to monitor inhibitor–protein interactions. However, predicting the binding mode of a fluorinated ligand from a chemical shift (or vice versa) has been challenging due to the high electron density of the fluorine atom. Nonetheless, reliable ^{19}F chemical-shift predictions to deduce ligand-binding modes hold great potential for in silico drug design. Herein, we present a systematic QM/MM study to predict the ^{19}F NMR chemical shifts of a covalently bound fluorinated inhibitor to the essential oxidoreductase tryparedoxin (Tpx) from African trypanosomes, the causative agent of African sleeping sickness. We include many protein–inhibitor conformations as well as monomeric and dimeric inhibitor–protein complexes, thus rendering it the largest computational study on chemical shifts of ^{19}F nuclei in a biological context to date. Our predicted shifts agree well with those obtained experimentally and pave the way for future work in this area.*

Fluorine is considered a "magic" element in medicinal and agricultural chemistry. It forms strong bonds to carbon, is the smallest biocompatible hydrogen substitute,[1] has the ability to form hydrogen bonds, and possesses a high electronegativity. Its introduction into small molecules can increase metabolic stability and allows the fine-tuning of physicochemical properties.[2] It is therefore not surprising that more than 20% of all FDA-approved drugs and more than 30% of all agrochemicals contain fluorine.[2] Replacing hydrogen by fluorine has been used successfully to, for example, investigate the interaction of inhibitors with proteases, explore their active site properties, and characterize inhibitors for neglected tropical diseases.[3]

With its 100% natural abundance, high gyromagnetic ratio, and the resulting high sensitivity, the spin-1/2 nucleus ^{19}F is of particular interest for NMR studies.[4] While practical advantages of fluorine for NMR spectroscopy have been exploited for many decades, the performance of corresponding quantum-chemical calculations for complex systems has gained momentum only lately.[5]

Chemical shifts of compounds containing fluorine have been calculated for many decades, from small molecules in the gas phase over biological systems in solution to solid-states.[6] The two most recent studies focusing on ^{19}F chemical shifts of biologically relevant molecules investigated crystals of fluorinated tryptophans[7] or monofluorinated phenylalanines in a protein (Brd4).[8] In the case of the tryptophan crystals, four molecules were used as a representation of the entire crystal. For Brd4, a quantum-mechanical/molecular-mechanical (QM/MM) setup was used with a buffer region of 4 Å and Boltzmann weighting of a few conformers. Nonetheless, the calculations differed from the measurements by between one and more than 20 ppm even after improving

[*] J. C. B. Dietschreit, Prof. Dr. C. Ochsenfeld
Theoretical Chemistry,
Department of Chemistry, University of Munich (LMU)
Butenandtstr. 7, 81377 Munich (Germany)
E-mail: christian.ochsenfeld@uni-muenchen.de
Prof. Dr. C. Ochsenfeld
Max Planck Institute for Solid State Research
70569 Stuttgart (Germany)
E-mail: c.ochsenfeld@fkf.mpg.de
A. Wagner, Prof. Dr. U. A. Hellmich
Dept. Chemistry, Section Biochemistry,
Johannes Gutenberg-Universität Mainz
55128 Mainz (Germany)
and
Centre for Biomolecular Magnetic Resonance (BMRZ),
Goethe-University Frankfurt
Max-von-Laue Str. 9, 60438 Frankfurt (Germany)
E-mail: u.hellmich@uni-mainz.de
T. A. Le, Prof. Dr. B. Engels
Institute for Physical and Theoretical Chemistry,
University of Würzburg
Emil-Fischer-Straße 42, 97074 Würzburg (Germany)
P. Klein, Prof. Dr. T. Opatz
Dept. Chemistry, Section Organic Chemistry,
Johannes Gutenberg-Universität Mainz, 55128 Mainz (Germany)
Prof. Dr. H. Schindelin
Institute of Structural Biology,
Rudolf Virchow Center for Experimental Biomedicine,
University of Würzburg, 97080 Würzburg (Germany)

Supporting information, which includes the experimental (NMR measurements) and computational (MM-MD simulation and QM/MM calculations) methods, NMR dilution curves, analysis of protein complex stability during the MD simulations, and figures highlighting the flexibility of CFT in complex with Tpx, and the ORCID identification number(s) for the author(s) of this article can be found under:
https://doi.org/10.1002/anie.202000539.

GDCh Communications Angewandte International Edition Chemie

predictions by linear regression to experimental data. Another study benchmarked different levels of quantum-chemical methods for fluorinated amino acids in implicit solvent, achieving at best a mean absolute error of 2.68 ppm with respect to the experiment.[9] Despite the impressive progress in the field, this is not sufficient to explain subtle differences in experimental spectra. Here, we use hundreds of frames from molecular dynamics (MD) simulations to ensure proper sampling of conformers and a significantly larger buffer region in our QM/MM calculations to increase the accuracy of our results.

Methods for computing NMR parameters range from empirical programs, such as SPARTA+,[10] to highly accurate QM calculations.[5,11,12] When using quantum-chemical methods, it has been shown that sufficiently large QM regions are necessary when describing complex systems.[13,14] However, the inclusion of many atoms is computationally very demanding. Thus, a plethora of methods has been devised to reduce the computational effort.[14,15] Here, we employ rigorous linear-scaling formulations that allow us to exploit the locality of the electronic structure within density-matrix-based theories. While this strongly reduces the computational scaling, for example, for the computation of NMR chemical shifts within density-functional theory from cubic to asymptotically linear, the accuracy is numerically unchanged and fully controlled.[5,16]

As a medically relevant test system, we selected the oxidoreductase tryparedoxin (Tpx), an essential enzyme of *Trypanosoma brucei*, the parasite that causes African sleeping sickness.[17] Tpx is inhibited by covalently binding to 2-(chloromethyl)-5-(4-fluorophenyl)thieno[2,3-d]pyrimidine-4-(3*H*)-one, CFT, which efficiently kills *T. brucei*.[18,19] CFT carries a 4-fluorophenyl moiety (Figure 1 A and Supporting Information, Figure S5). The chlorine leaving group facilitates the covalent interaction with Cys40 in the active site of Tpx.

In the asymmetric unit of our monoclinic crystals, three protein chains with two different inhibitor orientations are present (PDB: 6GXY, binding pose 1 for chains A and B, binding pose 2 for chain C, Figure 1 B).[19] In binding pose 1, the covalently bound CFT features extensive intramolecular interactions with the protein, including T-shaped π-stacking interactions with Trp70 and a weak hydrogen bond of the CFT fluorine with the backbone-Hα of Glu107. In binding pose 2, CFT is not in contact with the protein beyond the covalent bond to Cys40, and its fluorine atom is solvent exposed instead. In both, crystal and solution, CFT binding to the wild-type protein in pose 1 leads to Tpx dimerization mediated by extensive intermolecular inhibitor–inhibitor stacking and inhibitor–protein interactions.[19] The dissociation constant for the CFT-induced Tpx dimer is approximately 5 μM. In binding pose 2, dimerization is structurally not possible. We

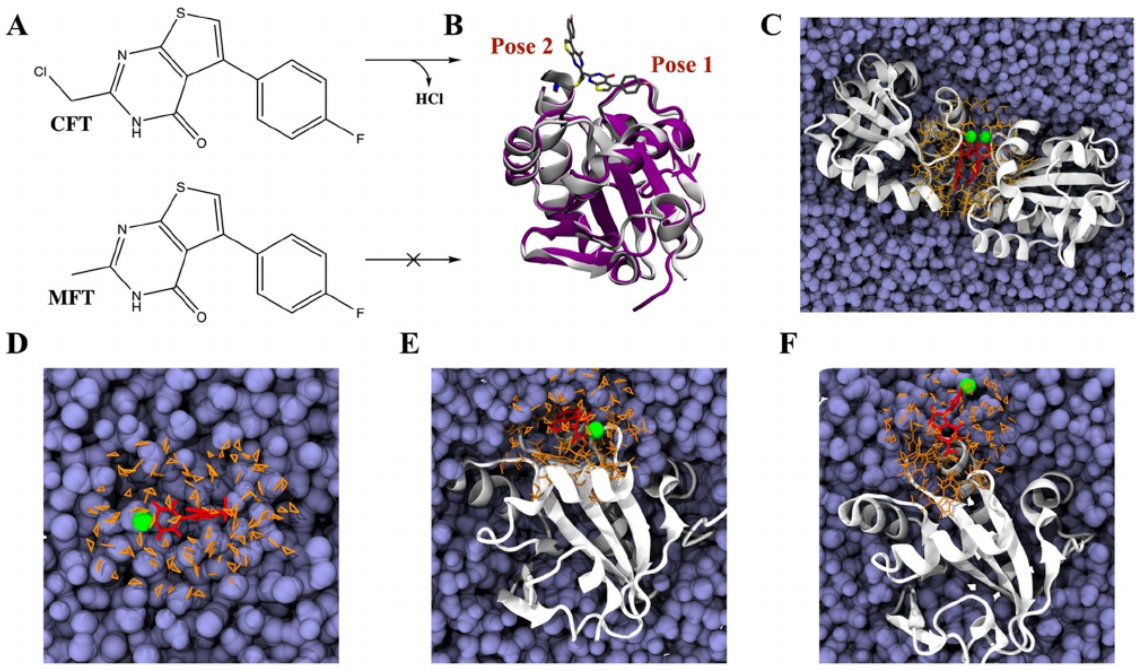


Figure 1. Interaction of *T. brucei* oxidoreductase tryparedoxin (Tpx) with a covalent inhibitor. A) cysteine-reactive CFT (top) and non-reactive MFT (bottom). B) Overlay of Tpx–CFT monomers in poses 1 and 2 as observed in our crystal structures (PDB: 6GXY).[19] C–F) Depiction of the QM region and MM embedding. Tpx is shown in white, water in blue, and all atoms in the QM region as orange sticks. The inhibitor is highlighted in red with its fluorine atom as green sphere. C) shows the Tpx–CFT dimer, D) the inhibitor in solution, E) the Tpx–CFT monomer in pose 1, and F) the Tpx–CFT monomer in pose 2.

identified residue Trp39 to be crucial for dimerization. Mutation of this active-site residue to alanine (Tpx-W39A) yields a protein that can still covalently interact with CFT, but dimerization upon inhibitor binding is extremely weak (Supporting Information, Figure S1).[19]

The ^{19}F signals for free CFT and its unreactive analogue 2-methyl-5-(4-fluorophenyl)thieno[2,3-d]pyrimidin-4(3*H*)-one, MFT, which is missing the chlorine leaving group (Figure 1A), were measured in solution at 298 K. Both ^{19}F chemical shifts were found to be very similar (−114.79 and −114.77 ppm, respectively). Upon binding of CFT to Tpx and the subsequent dimerization, a downfield chemical shift of approximately 0.3 ppm for the ^{19}F signal (−114.47 ppm) and substantial line broadening are observed, in agreement with incorporation of CFT into a high molecular weight, dimeric complex (Figure 2). Hence, the simultaneous availability of ^{19}F NMR and X-ray data for the Tpx–inhibitor system renders it exceptionally well-suited for systematic ^{19}F chemical shift studies. At high concentrations (greater than 500 μM), the ^{19}F chemical shift for the CFT–Tpx complex does not depend on the concentration of the protein–inhibitor complex. This suggests that under these conditions the ^{19}F chemical shift is not influenced by the monomer/dimer equilibrium since these concentrations are far above the K_D for dimerization. Only after significantly lowering the concentration, the ^{19}F signal starts to shift further downfield, indicating an increasing population of the monomer in exchange with the dimer (Supporting Information, Figure S2). However, due to limitations in the signal-to-noise ratio for the ^{19}F NMR measurements at high dilution, the concentration of the complex could not be reduced far enough below the K_D ($\ll 5$ μM) to reach conditions where the monomer is exclusively observed. In agreement with the above observation, the measured chemical shift for CFT bound to the W39A-mutant is shifted further downfield by an additional approximately 0.8 ppm compared to the WT complex (−113.7 ppm). This mutant does not significantly dimerize, and the distances from the inhibitor's fluorine group to the W39 Cα atom and to the W39 indole group are more than 16 Å and 11.5 Å, respectively, and thus should not affect the inhibitor's ^{19}F chemical shift. Ergo, this mutant can be used as a reference point for the ^{19}F chemical shift of CFT bound to a Tpx monomer.

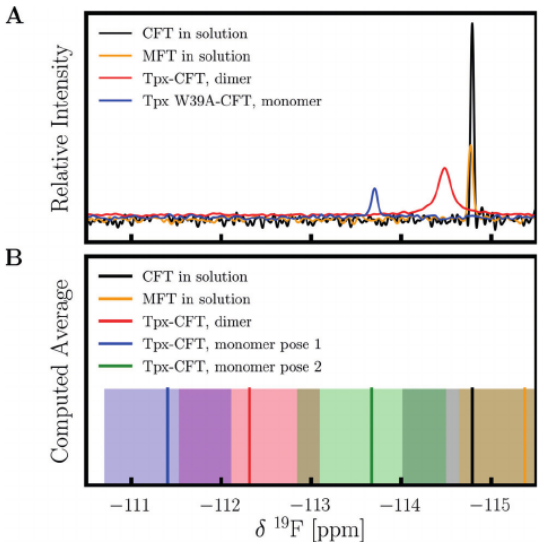


Figure 2. Comparison of experimentally measured (A) and calculated (B) ^{19}F NMR shifts. A) We compare CFT and MFT in solution, CFT in the dimeric complex and the W39A monomeric mutant. In the experiment, only one peak could be found for the monomeric protein. B) The computed averages are shown as vertical lines with the SEM indicated by an area shaded with decreased saturation. The calculations distinguish between two poses observed in the crystal structure (Figure 1). Importantly, the calculations and the experiment reveal the same ordering of peaks and indicate that pose 2 does not exist in solution.

In contrast to small organic molecules, the free energy landscape of solvated biomolecules typically does not possess one deep minimum, but rather a multitude of energetically close configurations that are thermodynamically accessible at physiological temperatures. In the NMR experiment, this implies the recording of an ensemble-averaged NMR chemical shift. Therefore, many different configurations of the system have to be taken into account to compute the observed chemical shift.[20] It has been demonstrated that inclusion of bond-length and bond-angle vibrations is often crucial for reliable chemical-shift computations.[21,22] Furthermore, the different relative orientations of molecules have to be accounted for as well. Here, we present a robust methodology based on MM-MD sampling and subsequent QM/MM calculations of ^{19}F NMR chemical shifts to identify CFT's binding poses relevant in solution.

Based on the crystal structure of the Tpx–CFT complex (PDB: 6GXY),[19] we computed separate MM-MD trajectories for free CFT and MFT in solution, for the monomeric complex with CFT bound in either pose 1 or 2, and the Tpx-dimer with both protomers binding to CFT in pose 1 (see the Supporting Information for setup details). An analysis of the complex stability during the simulations and inhibitor flexibility is given in Figures S4–S8 and Table S2 in the Supporting Information.

Subsequently, the ^{19}F chemical shifts of the inhibitor were calculated with our linear-scaling methods[5,16] for 200 snapshots taken evenly spaced in time from each MD trajectory. All interactions of either the protein or solvent atoms with the inhibitor were included explicitly. The aim was to perform high quality ^{19}F chemical-shift calculations on a large scale, based on an adequate description of the accessible phase space. The configurational ensembles included, for example, different solvation patterns or orientations of the inhibitor towards Tpx.

The high electron density of ^{19}F renders its spectroscopic properties, such as NMR chemical shifts, particularly challenging to calculate.[23] Motivated by previous studies,[12,13] we tested KT2[24] and B97-2[25] together with the NMR-specific basis set pcS-1,[26] and the necessary QM buffer size (Supporting Information, Figure S9). Both functionals show identical QM size convergence. As the radius of the QM region is increased from 4 to 5 Å, the calculated chemical shift changes by more than 0.5 ppm, highlighting the importance of

a sufficiently large QM sphere. A 7 Å QM buffer region around the inhibitor was found to be necessary to obtain size-converged shifts. Hence, we performed QM/MM-NMR calculations with KT2/pcS-1 for a 7 Å QM region. For the dimeric wild-type Tpx–CFT complex, this region (Figure 1 C) includes more than 1000 atoms. In combination with the 200 sampling points taken from each of the individual MD trajectories, this makes it one of the largest QM-based ^{19}F NMR chemical-shift calculations reported so far.[7,8,27] The QM size and the combined number of calculations in this context are unprecedented and are at the frontier of what is currently possible for such large biological systems. The embedded QM regions are shown in Figure 1 C–F. A detailed description is given in the Supporting Information (Supporting Information, Table S3).

It is important to note that the values of the calculated ^{19}F shifts for individual MD frames are scattered over a large shift range of about 60 ppm (Supporting Information, Figure S10). This is, however, not unexpected due to bond-length and bond-angle vibrations in the MD simulations.[21] The vibrations causing these distributions, especially those of the C–F bond, are fast processes on the spectroscopic time scale, and thus are not observed experimentally. We use the experimentally measured ^{19}F signal of free CFT in solution as reference for our calculated values, as relative shieldings are much more accurate.

As expected, our calculated ^{19}F chemical shifts correctly predict that the chemical shifts for free CFT and MFT are very similar to each other. Importantly, the predicted chemical shift for the inhibitor bound to Tpx is calculated to be downfield shifted compared to free CFT. An even more pronounced downfield shift compared to free CFT is predicted for CFT bound to monomeric Tpx in binding pose 1. This is in excellent agreement with what is observed experimentally for CFT bound to the monomeric W39A Tpx-mutant and in our dilution experiments (Supporting Information, Figures S1 and S2). In contrast, for the monomeric complex in pose 2, a chemical shift is calculated that lies between free CFT and the dimeric complex. Thus, our calculations qualitatively predict the correct order of the ^{19}F chemical shifts for CFT in the different states, as well as the true direction of chemical-shift changes induced by protein binding and complex dissociation. They further suggest that binding pose 2 of CFT observed in chain C of the crystal structure is not relevant in solution, as one would then expect, for CFT bound to the monomeric W39A mutant, a ^{19}F signal with a chemical shift in between those of free CFT and CFT bound to dimeric wild-type Tpx. This agrees well with the extended degree of solvent exposure of the inhibitor in this binding pose (Supporting Information, Figure S8), thus rendering it more similar to the free inhibitor. However, our calculations overestimate the chemical-shift differences between the different states. Already the calculated chemical shift difference between free CFT and free MFT (0.58 ppm) is larger than the measured one (0.02 ppm). This pattern continues for the other pairwise chemical-shift differences (CFT vs. CFT–WT: 0.32 ppm/2.48 ppm; Dimer vs. Monomer (pose 1): 0.77 ppm/0.97 ppm). Nevertheless, the computed trend allows us to discriminate between the different structures observed experimentally and to assign the measured shifts to a given conformer.

The accuracy of the prediction could be further increased by using a higher level of theory, a larger basis set, or more accurate dynamics (QM/MM-MD instead of MM-MD) improving the description of bond lengths, vibrations, and non-covalent interactions, which would entail, however, significantly higher computational costs.

Our study underlines the usefulness of ^{19}F NMR for the investigation of complex protein–inhibitor interactions, showcases current computational possibilities, and illustrates the power of predicting ^{19}F NMR chemical shifts in a complex biological system as a prerogative for further biomedical applications and drug design.

Acknowledgements

C.O. acknowledges funding by the "Deutsche Forschungsgemeinschaft" (DFG, German Research Foundation)—SFB 1309-325871075 and support as a Max-Planck Fellow at the Max-Planck Institute for Solid-State Research in Stuttgart. U.A.H. acknowledges support by the Carl Zeiss Foundation and the JGU Mainz Inneruniversitäre Forschungsförderung. This work was supported by the Rhineland-Palatinate Natural Products Research Center and the Center for Biomolecular Magnetic Resonance (BMRZ), Frankfurt University which is funded by the state of Hesse. H.S. acknowledges support from the Rudolf Virchow Center for Experimental Biomedicine. We thank Elke Duchardt-Ferner, Benedikt Goretzki, and Luise Krauth-Siegel for stimulating discussions.

Conflict of interest

The authors declare no conflict of interest.

Keywords: African sleeping sickness · covalent inhibitors · NMR spectroscopy · quantum chemistry · structural biology

[1] E. Neil, G. Marsh, *Chem. Biol.* **2000**, *7*, R153 – R157.
[2] A. Strunecká, J. Patočka, P. Connett, *J. Appl. Biomed.* **2004**, *2*, 141 – 150; K. Müller, C. Faeh, F. Diederich, *Science* **2007**, *317*, 1881 – 1886.
[3] M. Giroud, M. Harder, B. Kuhn, W. Haap, N. Trapp, W. B. Schweizer, T. Schirmeister, F. Diederich, *ChemMedChem* **2016**, *11*, 1042 – 1047; J. A. Olsen, D. W. Banner, P. Seiler, U. O. Sander, A. D'Arcy, M. Stihle, K. Müller, F. Diederich, *Angew. Chem. Int. Ed.* **2003**, *42*, 2507 – 2511; *Angew. Chem.* **2003**, *115*, 2611 – 2615; M. Berninger, C. Erk, A. Fuß, J. Skaf, E. Al-Momani, I. Israel, M. Raschig, P. Güntzel, S. Samnick, U. Holzgrabe, *Eur. J. Med. Chem.* **2018**, *152*, 377 – 391.
[4] C. Kang, *Curr. Med. Chem.* **2019**, *26*, 4964; D. Rose-Sperling, M. A. Tran, L. M. Lauth, B. Goretzki, U. A. Hellmich, *Biol. Chem.* **2019**, *400*, 1277.
[5] C. Ochsenfeld, J. Kussmann, F. Koziol, *Angew. Chem. Int. Ed.* **2004**, *43*, 4485 – 4489; *Angew. Chem.* **2004**, *116*, 4585 – 4589; M. Beer, J. Kussmann, C. Ochsenfeld, *J. Chem. Phys.* **2011**, *134*, 074102.
[6] U. Sternberg, M. Klipfel, S. L. Grage, R. Witter, A. S. Ulrich, *Phys. Chem. Chem. Phys.* **2009**, *11*, 7048 – 7060; M. E. Harding,

M. Lenhart, A. A. Auer, J. Gauss, *J. Chem. Phys.* **2008**, *128*, 244111; E. Y. Lau, J. T. Gerig, *J. Am. Chem. Soc.* **2000**, *122*, 4408 – 4417; N. K. Mishra, A. K. Urick, S. W. J. Ember, E. Schönbrunn, W. C. Pomerantz, *ACS Chem. Biol.* **2014**, *9*, 2755 – 2760; C. Kasireddy, J. G. Bann, K. R. Mitchell-Koch, *Phys. Chem. Chem. Phys.* **2015**, *17*, 30606 – 30612; T. Tanuma, J. Irisawa, *J. Fluorine Chem.* **1999**, *99*, 157 – 160; J. Augspurger, J. G. Pearson, E. Oldfield, C. E. Dykstra, K. D. Park, D. Schwartz, *J. Magn. Reson.* **1992**, *100*, 342 – 357; A. C. de Dios, J. G. Pearson, E. Oldfield, *Science* **1993**, *260*, 1491 – 1496; A. Zheng, S.-B. Liu, F. Deng, *J. Am. Chem. Soc.* **2009**, *131*, 15018 – 15023.

[7] M. Lu, S. Sarkar, M. Wang, J. Kraus, M. Fritz, C. M. Quinn, S. Bai, S. T. Holmes, C. Dybowski, G. P. A. Yap, J. Struppe, I. V. Sergeyev, W. Maas, A. M. Groneborn, *J. Phys. Chem. B* **2018**, *122*, 6148 – 6155.

[8] W. C. Isley, A. K. Urick, W. C. K. Pomerantz, C. J. Cramer, *Mol. Pharm.* **2016**, *13*, 2376 – 2386.

[9] J. N. Dahanayake, C. Kasireddy, J. M. Ellis, D. Hildebrandt, O. A. Hull, J.-P. Karnes, D. Morlan, K. R. Mitchell-Koch, *J. Comput. Chem.* **2017**, *38*, 2605 – 2617.

[10] Y. Shen, A. Bax, *J. Biomol. NMR* **2010**, *48*, 13 – 22.

[11] T. Helgaker, M. Jaszunski, K. Ruud, *Chem. Rev.* **1999**, *99*, 293 – 352; J. Vaara, *Phys. Chem. Chem. Phys.* **2007**, *9*, 5399 – 5418; M. Bühl, V. G. Malkin, M. Kaupp, *Calculation of NMR and EPR Parameters, Theory and Applications*, Wiley-VCH, Weinheim, **2004**; F. A. A. Mulder, M. Filatov, *Chem. Soc. Rev.* **2010**, *39*, 578 – 590; J. Gauss, J. F. Stanton, *J. Chem. Phys.* **1995**, *103*, 3561 – 3577; J. Gauss, J. F. Stanton, *Chem. Phys. Lett.* **1997**, *276*, 70 – 77; J. Gauss in *Modern Methods and Algorithms of Quantum Chemistry NIC Series, Vol. 3*, 2nd ed. (Ed.: J. Grotendorst), John von Neumann Inst. for Computing, Jülich **2000**, pp. 541 – 592.

[12] D. Flaig, M. Maurer, M. Hanni, K. Braunger, L. Kick, M. Thubauville, C. Ochsenfeld, *J. Chem. Theory Comput.* **2014**, *10*, 572 – 578.

[13] D. Flaig, M. Beer, C. Ochsenfeld, *J. Chem. Theory Comput.* **2012**, *8*, 2260 – 2271.

[14] C. Steinmann, J. M. H. Olsen, J. Kongsted, *J. Chem. Theory Comput.* **2014**, *10*, 981 – 988; J. D. Hartman, T. J. Neubauer, B. G. Caulkins, L. J. Mueller, G. J. O. Beran, *J. Biomol. NMR* **2015**, *62*, 327 – 340.

[15] D. B. Chesnut, K. D. Moore, *J. Comput. Chem.* **1989**, *10*, 648 – 659; J. M. H. Olsen, N. H. List, K. Kristensen, J. Kongsted, *J. Chem. Theory Comput.* **2015**, *11*, 1832 – 1842; C. Steinmann, L. A. Bratholm, J. M. H. Olsen, J. Kongsted, *J. Chem. Theory Comput.* **2017**, *13*, 525 – 536; X. Jin, T. Zhu, J. Z. H. Zhang, X. He, *Front. Chem.* **2018**, *6*, 1 – 11; T. Zhu, J. Z. H. Zhang, X. He, *J. Chem. Theory Comput.* **2013**, *9*, 2104 – 2114; M. Svensson, S. Humbel, R. D. J. Froese, T. Matsubara, S. Sieber, K. Morokuma, *J. Phys. Chem.* **1996**, *100*, 19357 – 19363; A. Zheng, M. Yang, Y. Yue, C. Ye, F. Deng, *Chem. Phys. Lett.* **2004**, *399*, 172 – 176; S. Moon, D. A. Case, *J. Comput. Chem.* **2006**, *27*, 825 – 836.

[16] J. Kussmann, M. Beer, C. Ochsenfeld, *WIREs Comput. Mol. Sci.* **2013**, *3*, 614 – 636; J. Kussmann, C. Ochsenfeld, *J. Chem. Phys.* **2013**, *138*, 134114.

[17] M. A. Comini, R. L. Krauth-Siegel, L. Flohé, *Biochem. J.* **2007**, *402*, 43 – 49; A. Wagner, E. Diehl, R. L. Krauth-Siegel, U. A. Hellmich, *Biomol. NMR Assignments* **2017**, *11*, 193 – 196.

[18] F. Fueller, B. Jehle, K. Putzker, J. D. Lewis, R. L. Krauth-Siegel, *J. Biol. Chem.* **2012**, *287*, 8792 – 8802.

[19] A. Wagner, T. A. Le, M. Brennich, P. Klein, N. Bader, E. Diehl, D. Paszek, A. K. Weickhmann, N. Dirdjaja, R. L. Krauth-Siegel, B. Engels, T. Opatz, H. Schindelin, U. A. Hellmich, *Angew. Chem. Int. Ed.* **2019**, *58*, 3640 – 3644; *Angew. Chem.* **2019**, *131*, 3679 – 3683.

[20] S. Grimme, C. Bannwarth, S. Dohm, A. Hansen, J. Pisarek, P. Pracht, J. Seibert, F. Neese, *Angew. Chem. Int. Ed.* **2017**, *56*, 14763 – 14769; *Angew. Chem.* **2017**, *129*, 14958 – 14964.

[21] M. Dračínský, H. M. Möller, T. E. Exner, *J. Chem. Theory Comput.* **2013**, *9*, 3806 – 3815.

[22] E. E. Kwan, R. Y. Liu, *J. Chem. Theory Comput.* **2015**, *11*, 5083 – 5089; S. Vogler, J. C. B. Dietschreit, L. D. M. Peters, C. Ochsenfeld, *Mol. Phys.* **2020**, accepted.

[23] A. M. Teale, O. B. Lutnæs, T. Helgaker, D. J. Tozer, J. Gauss, *J. Chem. Phys.* **2013**, *138*, 024111; G. L. Stoychev, A. A. Auer, R. Izsak, F. Neese, *J. Chem. Theory Comput.* **2018**, *14*, 619 – 637.

[24] T. W. Keal, D. J. Tozer, *J. Chem. Phys.* **2003**, *119*, 3015 – 3024.

[25] P. J. Wilson, T. J. Bradley, D. J. Tozer, *J. Chem. Phys.* **2001**, *115*, 9233 – 9242.

[26] F. Jensen, *J. Chem. Theory Comput.* **2008**, *4*, 719 – 727.

[27] C. Di Pietrantonio, A. Pandey, J. Gould, A. Hasabnis, R. S. Prosser in *Methods in ENZYMOLOGY Biological NMR Part B* (Ed.: A. J. Wand), Academic Press, Cambridge, **2019**, pp. 103 – 130.

Manuscript received: January 13, 2020
Revised manuscript received: March 22, 2020
Accepted manuscript online: April 2, 2020
Version of record online: May 25, 2020

Supporting Information

Predicting ^{19}F NMR chemical shifts: A combined computational and experimental study of a trypanosomal oxidoreductase-inhibitor complex

Johannes C. B. Dietschreit[1], Annika Wagner[2,3], T. Anh Le[4], Philipp Klein[5], Hermann Schindelin[6], Till Opatz[5], Bernd Engels[4], Ute A. Hellmich[2,3], Christian Ochsenfeld[1,7]

[1] Theoretical Chemistry, Department of Chemistry, University of Munich (LMU), Butenandtstr. 7, D-81377 Munich, Germany

[2] Dept. Chemistry, Section Biochemistry, Johannes Gutenberg-Universität Mainz, D-55128 Mainz, Germany

[3] Centre for Biomolecular Magnetic Resonance (BMRZ), Goethe-University Frankfurt, Max-von-Laue Str. 9, D-60438 Frankfurt/M, Germany

[4] Institute for Physical and Theoretical Chemistry, University of Würzburg, Emil-Fischer-Straße 42, D-97074 Würzburg, Germany

[5] Dept. Chemistry, Section Organic Chemistry, Johannes Gutenberg-Universität Mainz, D-55128 Mainz, Germany

[6] Institute of Structural Biology, Rudolf Virchow Center for Experimental Biomedicine, University of Würzburg, D-97080 Würzburg, Germany

[7] Max-Planck-Institute for Solid-State Research, D-70569 Stuttgart, Germany

Correspondence should be addressed to: u.hellmich@uni-mainz.de; c.ochsenfeld@fkf.mpg.de

Materials and Methods:

Inhibitor synthesis:
Synthesis of 2-(chloromethyl)-5-(4-fluorophenyl)thieno[2,3-d]pyrimidine-4(3*H*)-one (CFT) and 2-methyl-5-(4-fluorophenyl)thieno[2,3-d]pyrimidine-4(3*H*)-one (MFT) was described previously [1].

NMR spectroscopy:
Tpx WT and Tpx W39A were prepared as described previously [1, 2]. For ^{19}F NMR measurements, the Tpx variants were incubated for 30 min with 4 mM TCEP in a 1:3 molar protein:CFT ratio. Free CFT was removed via size exclusion chromatography on a 24 mL high-resolution column (Enrich^{TM} 70 10 x 300 SEC Bio-Rad Laboratories GmbH). Fractions containing Tpx-CFT were combined and concentrated to 1 mM (Tpx WT-CFT) or 700 µM (Tpx W39A-CFT) in 25 mM NaP_i pH 7.5, 150 mM NaCl containing 10% D_2O.

^{19}F-NMR spectra of isolated inhibitors (100 µM, 1k scans) or protein-inhibitor complexes (see Table S1) were recorded on a Bruker AVANCE 3 600 MHz spectrometer with a Prodigy TCI cryoprobe (Bruker, Karlsruhe). All measurements were carried out at 298 K in Tpx buffer (25 mM NaP_i pH 7.5, 150 mM NaCl) supplemented with 10% D_2O and 3% DMSO.

Table S1: Parameters used for ^{19}F NMR measurements of Tpx in complex with inhibitor.

Construct	Concentration	scans
Tpx WT-CFT	750 µM	512
	500 µM	512
	250 µM	512
	100 µM	2k
	50 µM	4k
	10 µM	50k
Tpx W39A-CFT	700 µM	64
	350 µM	64
	200 µM	128
	100 µM	256
	50 µM	256
	25 µM	512

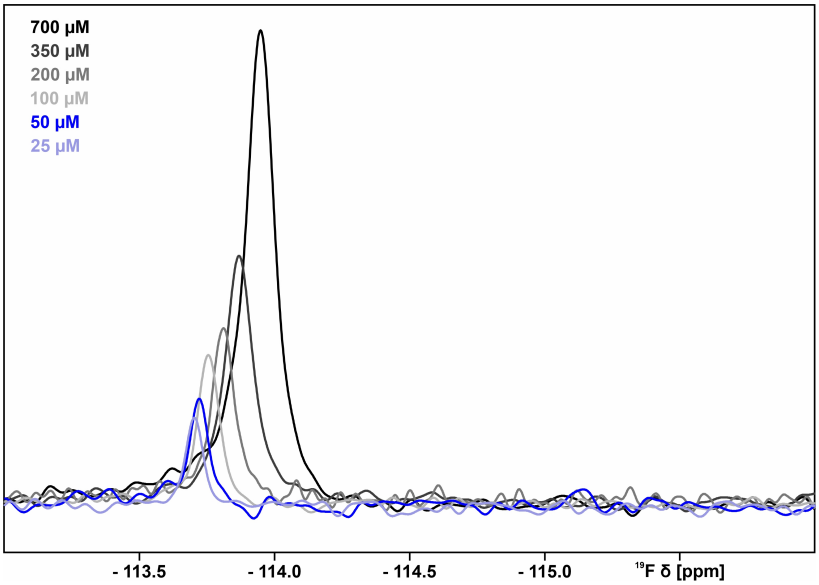


Figure S1: Evolution of the CFT-Tpx W39A ^{19}F NMR peak upon dilution. Only small shifts are observed indicating a residual, low affinity interaction between two CFT-bound Tpx W39A monomers. Therefore, the resulting peak at 25 µM protein was assumed as the representative of the true monomeric CFT-Tpx complex. Importantly, and in agreement with the effects observed for the WT protein and with some residual dimerization, the chemical shift displays a slight shift at lower protein concentrations.

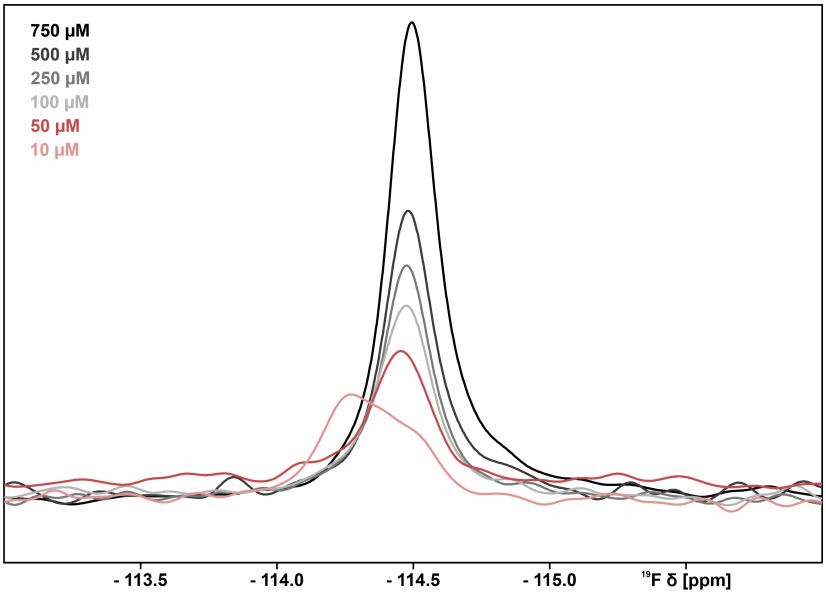


Figure S2: Evolution of the CFT-Tpx dimer peak upon dilution. At concentrations below 100 µM, the ^{19}F peak of CFT-Tpx develops a shoulder shifted towards the direction of the chemical shift observed for the monomeric CFT-Tpx W39A bound construct, and thus indicates the increasing presence of CFT-Tpx monomers with decreasing protein concentration.

MD simulations: For the MD simulations, the protein was described by the AMBER force field ff14SB and CFT with the generalized AMBER force field (GAFF) [3, 4]. Partial charges were computed with the Gaussian03 program package on the HF/6-31G(d) level by fitting the electrostatic potential [5, 6]. For the protein, hydrogen atoms were added and the charge of the protein-inhibitor complex was neutralized with Na^+ ions using tleap [7]. The covalent protein-inhibitor complex and both isolated compounds, CFT or MFT, were solvated in a truncated octahedral shell with the TIP3P water model with a 10 Å buffer around the complex or a 18 Å buffer around isolated CFT or MFT, respectively [8]. In the crystal structure (6GXY) of the Tpx-CFT complex [1], chains A and B in the unit cell are comparable (see Fig. S3). Therefore, CFT-Tpx pose 1 refers to chain A of the crystal structure, CFT-Tpx pose 2 to chain C and the CFT-Tpx dimer to chains A and B.

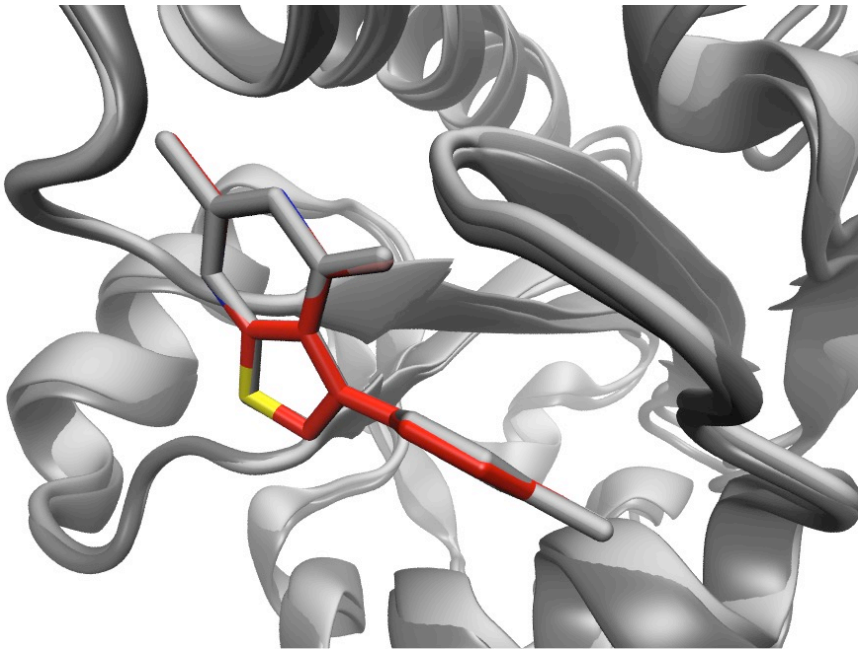


Figure S3: Superposition of chains A and B of the crystal structure of the Tpx-CFT complex (6GXY). Due to the high resolution of 1.6 Å, no non-crystallographic symmetry restraints were employed during refinement. CFT of chain A is shown with carbon depicted in red, sulfur in yellow, nitrogen in blue, oxygen in pink and fluorine in white. Tpx of chain A is shown in dark gray. All atoms of CFT of chain B and Tpx of chain B are shown in light gray. The two structures are virtually indistinguishable.

The MD simulations were performed with the AMBER14 software package [7]. First, the solvent was minimized for 1000 cycles with the simulation engine sander (Simulated Annealing with NMR-Derived Energy Restraints) with restraints on the protein-inhibitor complex with a force constant of 500 kcal/(mol·Å^2), followed by 2500 cycles for the whole system. Afterwards, the system was gradually heated to 300 K over 100 ps. All production runs had a duration of 30 ns.

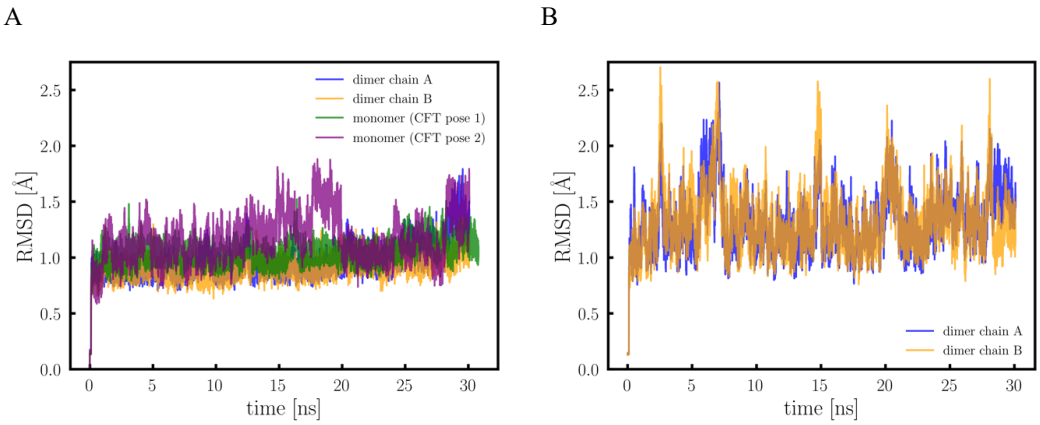


Figure S4: A) Backbone RMSD of the Tpx monomer (CFT in pose 1 and 2) and the Tpx dimer (chain A and B). In the dimeric case, each Tpx monomer was superimposed on itself. Tpx is more flexible in the monomeric pose 2 as the inhibitor does not reside in the binding site. B) Backbone RMSD of the Tpx dimer (chain A and B), where all frames were superimposed on the full dimer. These values are higher than in (A) as the dimeric complex is more flexible. Calculated mean and standard deviations are given in Table S2.

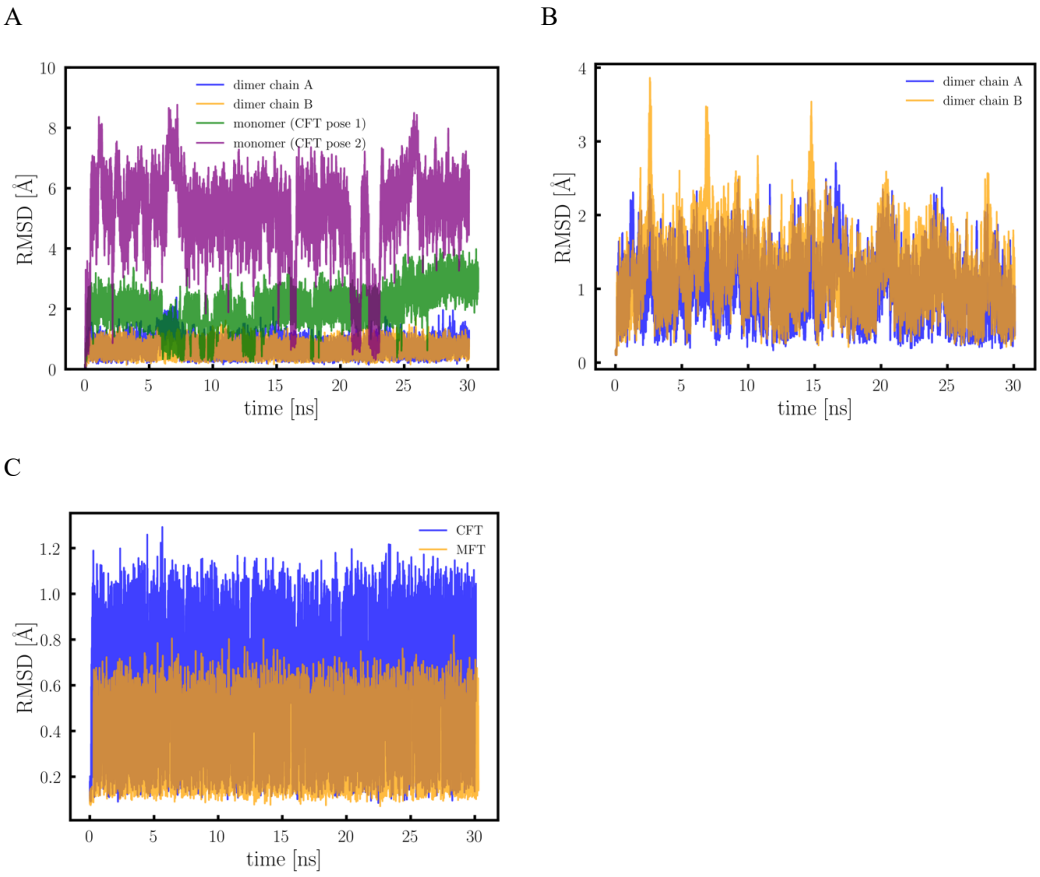


Figure S5: Heavy atom RMSD of the inhibitor CFT in complex with a TPX monomer (pose 1 and 2) or in complex with the TPX dimer (chains A and B). In A) the atoms were aligned with the backbone of the respective Tpx chain, in B) with the backbone of the full dimer. C) shows the heavy atom RMSD of CFT and MFT in solution. The additional chlorine in CFT clearly increases its RMSD. Calculated mean and standard deviations are given in Table S2.

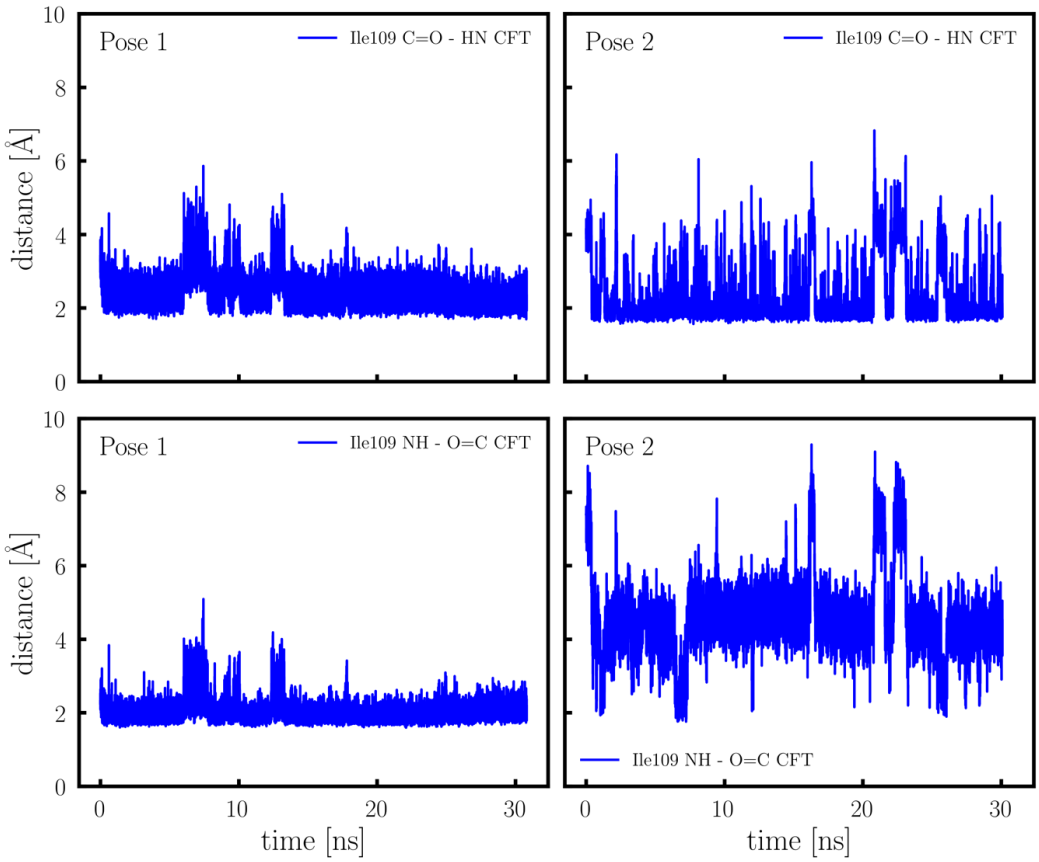


Figure S6: Two intermittent H-bond interactions between the backbone atoms of Ile109 of the TPX monomer and CFT in pose 1 (left column) and pose 2 (right column). The top row shows the H-bond with CFT as donor, in the bottom row CFT is the acceptor. Calculated mean and standard deviations are given in Table S2.

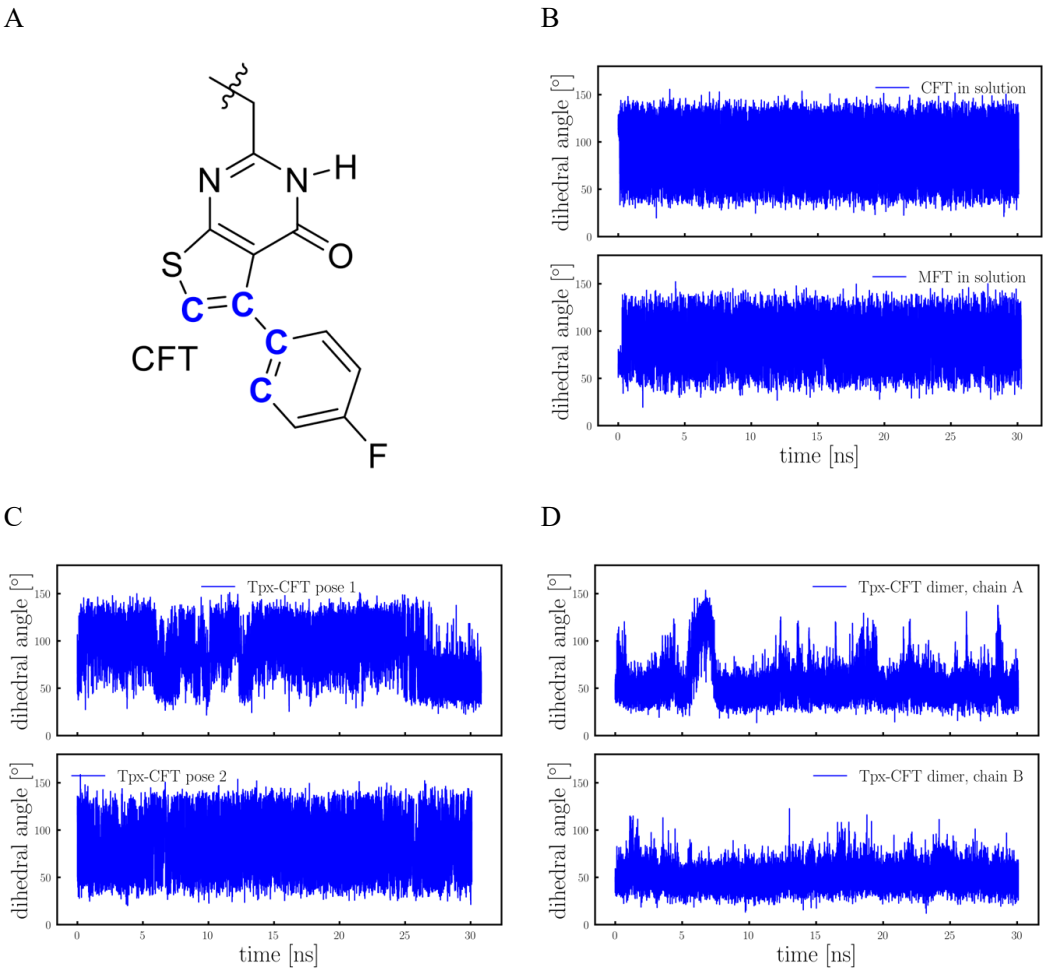


Figure S7: A) Selected atoms in blue for the calculation of the dihedral angle analyzed in B-D. B) CFT (top) and MFT (bottom) in solution, C) CFT bound to Tpx monomer (poses 1 or 2) or D) CFT in the Tpx dimer. Calculated mean and standard deviations are given in Table S2.

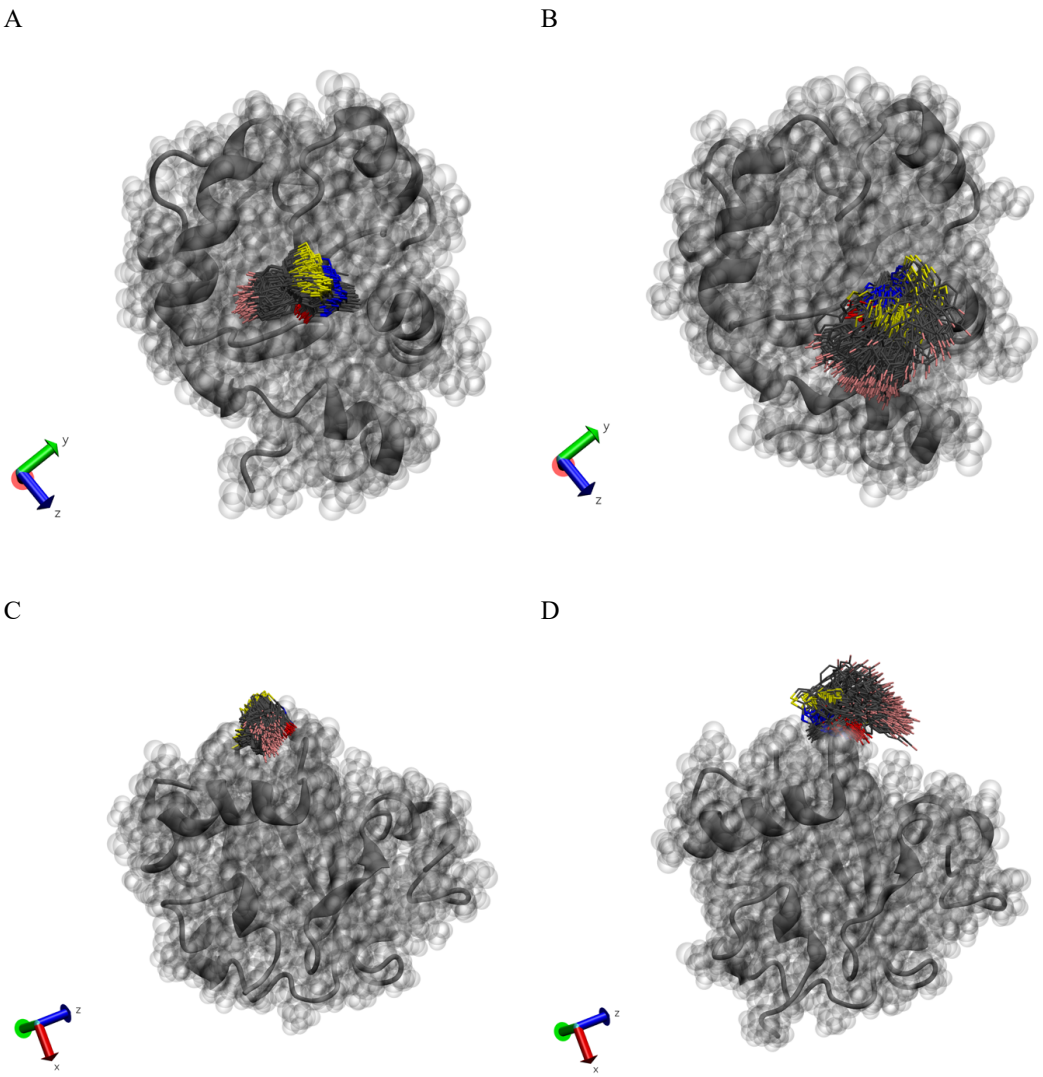


Figure S8: Comparison of the conformational space explored by the inhibitor CFT bound to Tpx in a monomeric complex during the MD simulations of poses 1 and 2. A/C display pose 1 and B/D pose 2. The angle of the view is indicated by the axes in the corner of each picture. The conformational space explored by the solvent exposed inhibitor in pose 2 is much greater than in pose 1. Yet, the inhibitor is nonetheless restricted in its motion by the protein surface as shown in D. Carbon atoms are shown in dark grey, nitrogen in blue, sulfur in yellow, and fluorine in pink. The protein backbone is shown in cartoon representation, the protein surface is visible as van-der-Waals spheres of the outer atoms.

Table S2: Calculated mean and standard deviations of the RMSD over the course of the MD simulations for the Tpx monomer and dimer, for CFT in solution or in complex with the Tpx monomer and dimer (see Figure S3, S4). To calculate the RMSD for Tpx all backbone atoms and for CFT all heavy atoms were considered and calculated in relation to the starting structure. Furthermore, mean and standard deviations were calculated for CFT-I109 distances (referring to Figure S5) or dihedrals of CFT (see Figure S6). 10,000 data points were included in the calculations. RMSD values and standard deviations (sdv) as well as distance values are given in Å, dihedrals in [°].

CFT / MFT	**CFT in solution**	**MFT in solution**	**Tpx-CFT monomer pose 1**	**Tpx-CFT monomer pose 2**	**TPX-CFT dimer chain A**	**TPX-CFT dimer chain B**
mean RMSD	0.6	0.3	2.0	5.1	1.0	1.2
sdv	0.2	0.2	0.6	1.4	0.4	0.5
Tpx backbone	**CFT in solution**	**MFT in solution**	**Tpx-CFT monomer pose 1**	**Tpx-CFT monomer pose 2**	**TPX-CFT dimer chain A**	**TPX-CFT dimer chain B**
mean RMSD	-	-	1.0	1.2	1.3	1.3
sdv	-	-	0.1	0.2	0.3	0.3
CFT to I109	**CFT in solution**	**MFT in solution**	**I109 C=O to HN CFT pose 1**	**I109 NH to O=C CFT pose 1**	**I109 C=O to HN CFT pose 2**	**I109 NH to O=C CFT pose 2**
mean distance	-	-	2.1	2.6	4.6	2.3
sdv	-	-	0.4	0.6	1.1	0.8
CFT dihedral	**CFT in solution**	**MFT in solution**	**Tpx-CFT monomer pose 1**	**Tpx-CFT monomer pose 2**	**TPX-CFT dimer chain A**	**TPX-CFT dimer chain B**
mean dihedral	88.6	89.1	90.3	81.6	55.2	50.2
sdv	27.9	27.6	25.1	28.0	18.6	11.5

^{19}F NMR chemical shift predictions:

Initial decision on the setup

Before calculating the NMR shifts for hundreds of frames, a small benchmark study was performed to estimate the optimal setup for this investigation. Frames from the simulation of the Tpx-CFT complex in pose 1 were selected. The QM/MM chemical shift calculations were performed at the DFT level for the QM part of the system (see definition below). The test calculations were performed with respect to DFT functional, QM-size convergence, and convergence criteria.

The QM/MM calculations were performed with the ChemShell[9] package. The MM part was described with the exact same parameters as used for the molecular dynamics simulations. The software package FermiONs++[10-12] was used for the QM calculations. The QM/MM interactions were described by electrostatic embedding in an additive scheme.

To study the size convergence behavior of the system, we performed calculations where the QM region included only the inhibitor, the inhibitor and all protein residues/water molecules surrounding it within 4, 5, 6, 7, and 8 Å, respectively (see Fig. S9).

The use of a large basis set, like pcS-2 [13], increased the computational cost (when using KT2) by an order of magnitude compared to pcS-1, and was thus deemed too expensive. The higher computational time for B97-2 [14] than KT2 [15] necessitated the use of KT2 instead of the hybrid functional B97-2. A high number of grid points was used in order to provide reliable chemical shifts (150 radial points and 974 angular points in the Lebdev grids). In general, very conservative settings were used to ensure the correctness of the computed chemical shifts (integral threshold: 10^{-10}, PreLinK threshold: 10^{-4}, SCF convergence threshold: 10^{-7}).

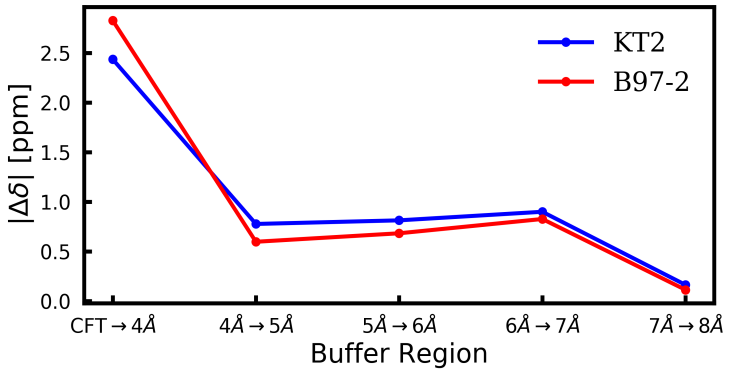


Figure S9. Convergence study of two DFT functionals, KT2 and B97-2, with respect to the QM-size (buffer region around CFT). The graph displays the difference between the absolute ^{19}F chemical shifts of the Tpx-CFT complex in pose 1 for increasing buffer regions around CFT. The convergence behavior of both DFT functionals is almost identical.

Calculation of ^{19}F NMR shifts

From every MD simulation 200 frames were selected, which were evenly spaced in time, so that a representation of the Boltzmann distribution could be obtained. For every frame all atoms of protein residues and water molecules within 7 Å of any atom belonging to the inhibitor molecule were selected. Due to the different atom configurations in every MD frame, the actual number of atoms in the QM region fluctuated. The isotropic NMR shifts were computed for every frame. The shift distributions are reported in Figure S8 while we refer to the calculated means in the main article. The convergence of the computed shifts is controlled by looking at the convergence of the mean as well as the standard deviation (measured for the width of the underlying distribution) with the number of frames (Figure S9).

Table S3: QM system sizes within the QM/MM chemical shift calculations.

System	CFT in solution	MFT in solution	Tpx-CFT pose 1	Tpx-CFT pose 2	Tpx-CFT dimer
Mean No. of QM-Atoms	491	481	893	877	1385

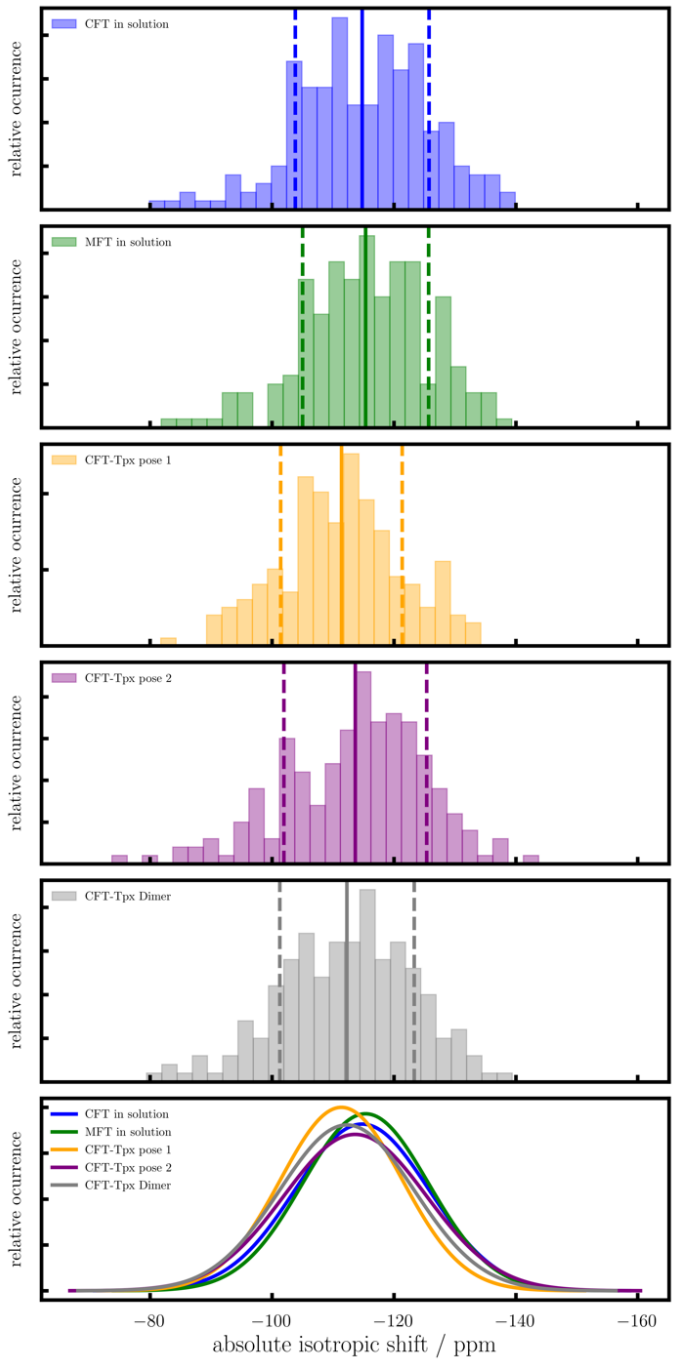


Figure S10: Comparison of ^{19}F chemical shifts of CFT and MFT alone in solution, the inhibitor in complex with a Tpx monomer in the two distinct poses (pose 1: buried in the binding pocket; pose 2: solvent exposed), and the inhibitor in the dimeric complex with the Tpx WT. Distributions of ^{19}F-chemical shifts computed for all MD frames. The bottom panel shows a gaussian function approximating the distributions above. The solid lines indicate the mean, the dashed lines the peak width at half height.

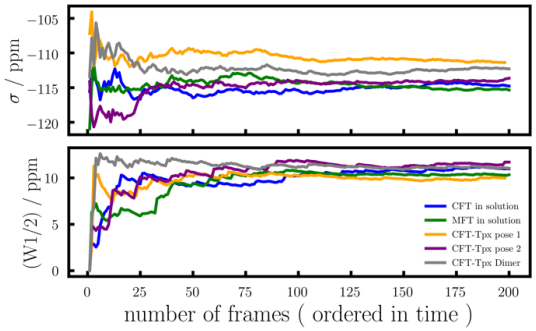


Figure S11: Changes in the mean of the ^{19}F chemical shift and the standard deviation with increasing number of frames. Both quantities are indicators of convergence.

Bibliography

1. Wagner, A., et al., *Inhibitor-Induced Dimerization of an Essential Oxidoreductase from African Trypanosomes.* Angew. Chem. Int. Ed., 2019. **58**(11): p. 3640-3644.
2. Wagner, A., et al., *Backbone NMR assignments of tryparedoxin, the central protein in the hydroperoxide detoxification cascade of African trypanosomes, in the oxidized and reduced form.* Biomol. NMR Assign., 2017. **11**(2): p. 193-196.
3. Wang, J., et al., *Automatic atom type and bond type perception in molecular mechanical calculations.* J. Mol. Graph. Model., 2006. **25**: p. 247-260.
4. Maier, J.A., et al., *ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB.* J. Chem. Theory Comput., 2015. **11**: p. 3696-3713.
5. Ditchfield, R., W.J. Hehre, and J.A. Pople, *Consistent Molecular Orbital Methods. 9. Extended Gaussian-type basis for molecular-orbital studies of organic molecules.* J. Chem. Phys., 1971. **54**: p. 724-728.
6. Frisch, M.J., et al., *Gaussian 03*. 2004, Gaussian, Inc.: Wallingford CT.
7. Case, D.A., et al., *AMBER14*. 2014, University of California: San Francisco.
8. Jorgensen, W.L., et al., *Comparison of simple potential functions for simulating liquid water.* J. Chem. Phys., 1983. **79**: p. 926-935.
9. Sherwood, P., et al., *QUASI: A general purpose implementation of the QM/MM approach and its application to problems in catalysis.* J. Mol. Struct.: THEOCHEM, 2003. **632**(1): p. 1-28.
10. Ochsenfeld, C., J. Kussmann, and F. Koziol, *Ab initio NMR spectra for molecular systems with a thousand and more atoms: A linear-scaling method.* Angew. Chem. Int. Ed., 2004. **43**: p. 4485-4489.
11. Kussmann, J., M. Beer, and C. Ochsenfeld, *Linear-scaling self-consistent field methods for large molecules.* WIREs Comput. Mol. Sci., 2013. **3**(6): p. 614-636.
12. Kussmann, J. and C. Ochsenfeld, *Pre-selective screening for matrix elements in linear-scaling exact exchange calculations.* J. Chem. Phys., 2013. **138**(13): p. 134114.
13. Jensen, F., *Basis set convergence of nuclear magnetic shielding constants calculated by density functional methods.* J. Chem. Theory Comput., 2008. **4**: p. 719-727.
14. Wilson, P.J., T.J. Bradley, and D.J. Tozer, *Hybrid exchange-correlation functional determined from thermochemical data and ab initio potentials.* J. Chem. Phys., 2001. **115**: p. 9233-9242.
15. Keal, T.W. and D.J. Tozer, *The exchange-correlation potential in Kohn-Sham nuclear magnetic resonance shielding calculations.* J. Chem. Phys., 2003. **119**: p. 3015-3024.

3.6 Publication VI: Important Components for Accurate Hyperfine Coupling Constants: Electron Correlation, Dynamic Contributions, and Solvation Effects

Sigurd Vogler, Johannes C. B. Dietschreit, Laurens D. M. Peters, and C. Ochsenfeld
"Important Components for Accurate Hyperfine Coupling Constants:
Electron Correlation, Dynamic Contributions, and Solvation Effects"
Mol. Phys. **2020**, e1772515

Abstract: The calculation of hyperfine coupling constants is a challenging task in balancing accuracy and computational effort. While previous work has shown the importance of electron correlation and molecular dynamic contributions, we present a systematic study simultaneously analyzing the influence of both effects on hyperfine coupling constants. To this end, we thoroughly study two organic radicals, namely the dimethylamino radical and ethanal radical cation, proving the need to account for conformational flexibility as well as the large influence of electron correlation. Based on these results, we analyze the effect of electron correlation and dynamic simulations on a set of 12 organic radicals, illustrating that both effects are vital for an accurate *in silico* description on the same scale. Furthermore, we study the influence of solvation using the efficient nuclei-selected algorithm to obtain hyperfine coupling constants with electron correlation for large systems, indicating the necessity to include explicit solvent molecules. Finally, we introduce a composite approach to incorporate all contributions for hyperfine coupling of radicals in solution at comparatively low computational cost. This is successfully tested on the hydroxylated TEMPO radical in aqueous solution, where we are able to compute a ^{14}N-HFCC of 44.4 MHz compared to the experimentally measured 47.6 MHz.

Important Components for Accurate Hyperfine Coupling Constants: Electron Correlation, Dynamic Contributions, and Solvation Effects

Sigurd Vogler[a,†], Johannes C. B. Dietschreit[a,†], Laurens D. M. Peters[a], and Christian Ochsenfeld[a,*]

[a]Chair of Theoretical Chemistry, Department of Chemistry, University of Munich (LMU), Butenandtstr. 7, 81377 Munich, Germany

ARTICLE HISTORY
Compiled August 23, 2020

ABSTRACT
The calculation of hyperfine coupling constants is a challenging task in balancing accuracy and computational effort. While previous work has shown the importance of electron correlation and molecular dynamic contributions, we present a systematic study simultaneously analyzing the influence of both effects on hyperfine coupling constants. To this end, we thoroughly study two organic radicals, namely the dimethylamino radical and ethanal radical cation, proving the need to account for conformational flexibility as well as the large influence of electron correlation. Based on these results, we analyse the effect of electron correlation and dynamic simulations on a set of 12 organic radicals, illustrating that both effects are vital for an accurate *in silico* description on the same scale. Furthermore, we study the influence of solvation using the efficient nuclei-selected algorithm to obtain hyperfine coupling constants with electron correlation for large systems, indicating the necessity to include explicit solvent molecules. Finally, we introduce a composite approach to incorporate all contributions for hyperfine coupling of radicals in solution at comparatively low computational cost. This is successfully tested on the hydroxylated TEMPO radical in aqueous solution, where we are able to compute a ^{14}N-HFCC of 44.4 MHz compared to the experimentally measured 47.6 MHz.

KEYWORDS
Computational Chemistry; Hyperfine Coupling Constants; Molecular Dynamics; Electron Correlation; Solvation Effects

1. Introduction

Electron paramagnetic resonance (EPR) spectroscopy is an important tool for studying radicals. As a non-invasive method, it is indispensable in research tackling many biological systems [1, 2]. However, the *ab initio* computation of the EPR parameters, namely the hyperfine coupling constants (HFCCs) and the g-tensors, remains challenging. These open-shell properties can be calculated using the unrestricted approach which incorporates spin polarization and delocalization, but can lead to erratic results due to spin contamination. This can be improved by using a spin restriction within the restricted-unrestricted ansatz by Rinkevicius *et al.* [3], where spin contamination is overcome while still including spin polarization. Nonetheless, the computationally

* CONTACT C. Ochsenfeld. Email: christian.ochsenfeld@uni-muenchen.de
† Contributed equally to this work

less demanding unrestricted approach often leads to reliable results, especially for well localized radicals and when using density functional theory (DFT) [4–8].

Based on an unrestricted framework, further aspects need to be considered: EPR-specific basis sets have been demonstrated to be important [9, 10], and while DFT is often highly accurate, higher-order levels of theory systematically taking electron correlation into account are often necessary. This includes methods such as second-order Møller-Plesset perturbation theory (MP2), double hybrid (DH)-DFT [6, 11], and coupled cluster approaches [5, 12–16], as well as different multi-reference ansätze [17–19].

Even with elaborate methods, extensive basis sets, and in absence of spin contamination, the results can deviate from experimental values due to the neglect of dynamic contributions, i.e. vibrational averaging. This was shown in recent work by Massolle *et al.* [20] on verdazyl radicals, where computational results at the DFT-level are improved by averaging over frames from a molecular dynamics (MD) simulation based on a quantum mechanically derived force field [21]. Similarly, studies on nitroxide radicals show that considering both vibrational averaging and solvent effects leads to more accurate results within the DFT framework [22, 23]. The influence of molecular and intermolecular motion of explicitly solvated benzosemiquinone was studied in detail by Asher and Kaupp [24]. Nonetheless, solvent effects are often small, and work by Rinkevicius *et al.* [25] shows that a description of the environment by means of molecular mechanics theory can be sufficient. Very recent work from the Cappelli group has used successfully a polarizable QM/MM ansatz for the Proxyl and TEMPO (2,2,6,6-Tetramethylpiperidinyloxyl) radical [26].

A thorough investigation of the effect of the bending angle of the methyl radical, its incorporation within an *ab initio* MD (AIMD) simulation [27], and its solvation [28] also mandates the correct description of dynamic contributions. Similarly, significant ro-vibrational contributions were shown in the analysis of out-of-plane bending in H_2NO [29], of dimethyl nitroxide [30], and of other organic radicals [31–33]. Here, we also want to mention recent corresponding work in the computation of nuclear magnetic resonance shielding tensors by Grimme *et al.* [34]. that considers a set of conformers or rotamers to accurately describe flexible molecules in solution.

While both, the effect of electron correlation and dynamic contributions significantly improve the *in silico* results, their combined description is computationally cumbersome. A straightforward approach is to perform an MD simulation and compute the EPR parameters for a set of frames. This requires the cost for computing the EPR parameters per frame to be small, allowing a sufficient number of frames to be computed for accurately incorporating the vibrational and rotational motion. The description of the correlation contribution by the cheapest wavefunction-based ansatz, MP2, is still expensive due to its conventionally large scaling behavior of $\mathcal{O}(N^5)$ as well as its large prefactor. This also applies to DH-DFT that contains a second-order perturbation theory term analogous to MP2 [11]. The prefactor can be reduced by the resolution-of-the-identity (RI) approximation [35–40], whereas linear scaling behavior can be achieved by a reformulation in, e.g. atomic orbitals (AO) [41], using distance-including integral estimates [42, 43]. The required analytic energy gradients for the computation of molecular properties at the MP2-level have been developed in the AO-basis [44]. By introducing the RI approximation and a Cholesky decomposition [45, 46] of the (pseudo-)density matrices (CDD) in the computation of AO-MP2 energy gradients [44], we recently presented a low-scaling, low-prefactor implementation to compute HFCCs [47]. By computing only selected nuclei, the computational cost can be reduced further [48]. Using these methods, large-scale computations of

HFCCs based on multiple frames from an MD simulation have become possible.

This work simultaneously analyses the effect of both dynamic contributions and electron correlation on the HFCCs, thus providing a computational protocol towards the calculation of accurate HFCCs for large molecular systems using the efficient quantum chemical methods introduced above. We highlight the importance of accounting for dynamic contributions by investigating in depth the dependence of the HFCC on geometric parameters, such as bond lengths, and bond and dihedral angles of two organic radicals, ethanal and dimethylamine. Subsequently, we analyse the contribution of electron correlation and dynamic contributions on isotropic HFCCs of a set of 12 organic radicals which also includes the two aforementioned examples. We chose organic molecules as their single-configurational character and the appearance of only light nuclei allow for a good description using DFT-based methods. Dynamic contributions are considered by computing a set of snapshots from an AIMD simulation (e.g. Refs. [49–51]) of the radicals using the fast small basis set Hartree-Fock method (HF3c) [52] and the efficient three-fold corrected Perdew-Burke-Ernzerhoff generalized-gradient-approximation (PBEh3c) [53]. The effect of electron correlation is shown by comparing the isotropic HFCCs both at the Hartree-Fock (HF) and DFT-level, as well as using RI-CDD MP2 and DH-DFT. Finally, solution effects are considered both with continuum solvation models [54] and with explicit solvent molecules by means of a hybrid quantum mechanical/molecular mechanics (QM/MM) ansatz. Based on these results, we present a pragmatic composite approach to incorporate electron correlation, dynamic contributions, and solvent effects.

2. Theory

The isotropic HFCC can be calculated in the absence of spin-orbit coupling by [55]

$$a_{\mathrm{iso}}(k) = \frac{\mu_0}{3} g_{\mathrm{e}} g_{\mathrm{N}}(k) \beta_{\mathrm{e}} \beta_{\mathrm{N}} \langle S_z \rangle^{-1} \rho(k), \tag{1}$$

where μ_0 is the permeability of the vacuum, g_e is the electronic and $g_N(k)$ is the nuclear g-factor of nucleus k, β_{e} is the Bohr magneton, β_{N} is the nuclear magneton, and $\langle S_z \rangle$ is the mean value of S_z in the current electronic state. $\rho(k)$ is the Fermi-contact integral (real-space spin density), which is

$$\rho(k) = \sum_{\mu\nu} P_{\mu\nu}^{\alpha-\beta} \langle \phi_\mu(r) | \delta(r - r_{\mathrm{k}}) | \phi_\nu(r) \rangle, \tag{2}$$

where $P_{\mu\nu}^{\alpha-\beta}$ is the difference between the α- and β-electron density matrices of the basis functions ϕ_μ and ϕ_ν, and the nucleus is located at r_k.

For methods beyond Hartree-Fock or DFT, the respective energy equation needs to be perturbed with respect to the nuclear magnetic moment M_k of nucleus k [55]. Thus, analytic gradients of the MP2 expression are required both for HFCCs at the MP2-level and at the DH-DFT level. We resort to our RI-CDD MP2 gradients [47] and the selected-nuclei ansatz [48] in the present work to efficiently obtain the correlation contributions also for large systems.

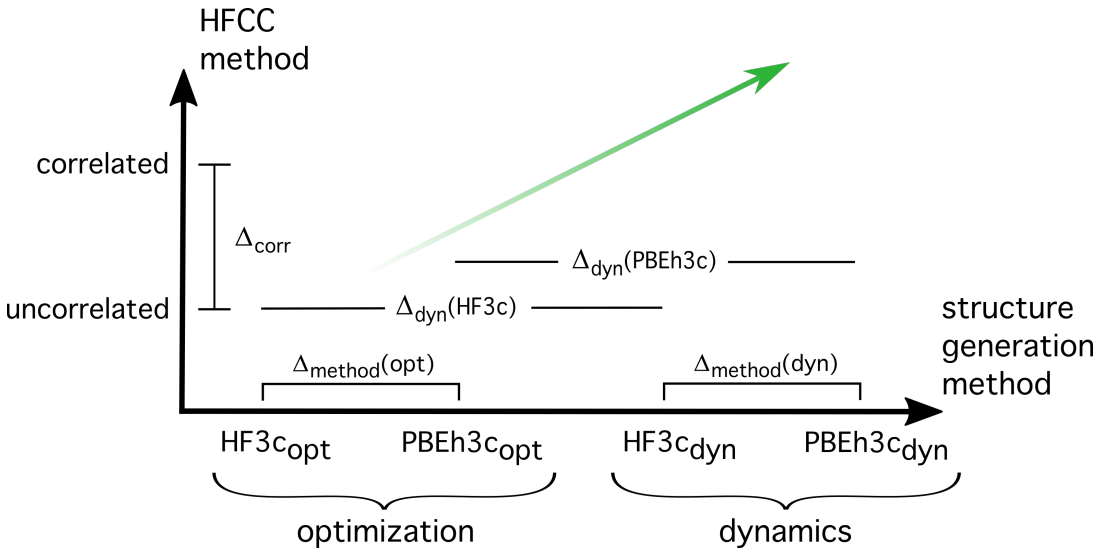


Figure 1. Graphical definitions of different contributions to the observable. We define the dynamic contribution Δ_{dyn} as the difference between optimization and a dynamics simulation at the same level of theory. The influence of the level of theory used for structure generation is called Δ_{method}; in this paper it is the difference in influence between HF3c and PBEh3c. Δ_{corr} denotes the effect of correlation on the computation of HFCCs.

3. Computational Details

The isotropic HFCCs were obtained at the HF-, DFT-, DH-DFT-, and MP2-level using the respective implementation in the program package FermiONs++ [56–58]. The MP2-contributions are hereby computed using the aforementioned RI-CDD ansatz [47] and a QQR-based integral screening [42, 43]. The Laplace expansion coefficients are selected based on the minimax-approximation [59]. The extents of the QQR-type integral estimates are determined with the same thresholds as in Ref. [43]. The QQR-screening threshold was set to 10^{-8} and seven Laplace expansion points were chosen based on the study of the accuracy in Ref. [47]. The Density matrix-based Laplace-transformed Unrestricted Coupled-Perturbed Self-Consistent-Field Theory (DL-UCPSCF) [60] equations were converged to a threshold of 10^{-4}. Deviations of less than 1 MHz can be expected with these thresholds [47]. The frozen core approximation was not used in our MP2 computations. All computations were performed with the highly accurate basis set EPR-III [9, 10], with the exception of both the reference coupled cluster computations with singles and doubles excitations (CCSD) [61], which were obtained with the program package Cfour [62] using the basis set def2-TZVPP [63]. The auxiliary basis set def2-TZVPP-RI/JK by Weigend [64] was chosen for the computations using the basis set EPR-III [9, 10]. For efficient DFT calculations on the TEMPO radical we used the newly implemented, exact semi-numerical exchange to gain the necessary speed up needed for such a large number of basis functions [65–67].

The AIMD simulations at the HF3c- [52] and PBEh3c-level [53] were performed as canonical (NVT-)ensembles with the Velocity Verlet propagator [68, 69] at 298.15 K, using the Bussi-Donadio-Parrinello thermostat [70]. Each simulation included a 100 fs equilibration period and a 10 ps production run with a time step of 0.1 fs. Geometries were saved every 1 fs. Furthermore, the fully converged extended Lagrangian Born-Oppenheimer MD (XL-BOMD) method [71] was used to speed up SCF convergence. HFCCs were computed for every 100 fs of the trajectory.

Our study of solvation effects on the hydroxylated TEMPO radical involved two additional kinds of simulations: one in which the solvent (water) was modelled as a polarizable continuum, and another which considered explicit solvent molecules in a QM/MM ansatz. All solvent simulations were performed with HF3c. We chose the conductor-like polarizable continuum model (C-PCM) SWIG (Switching/Gaussian) [72–74], which uses Gaussian charge distributions instead of point charges centered on all tesserae. The dielectric constant of the continuum was set to 78.355 to model water and the additive constant in the denominator to 0.5 in order to be analogous to the conductor-like screening model COSMO [75]. The volume of the molecule was determined at every time step using the atomic radii defined by Bondi [76]. The remaining settings are identical to the vacuum simulations.

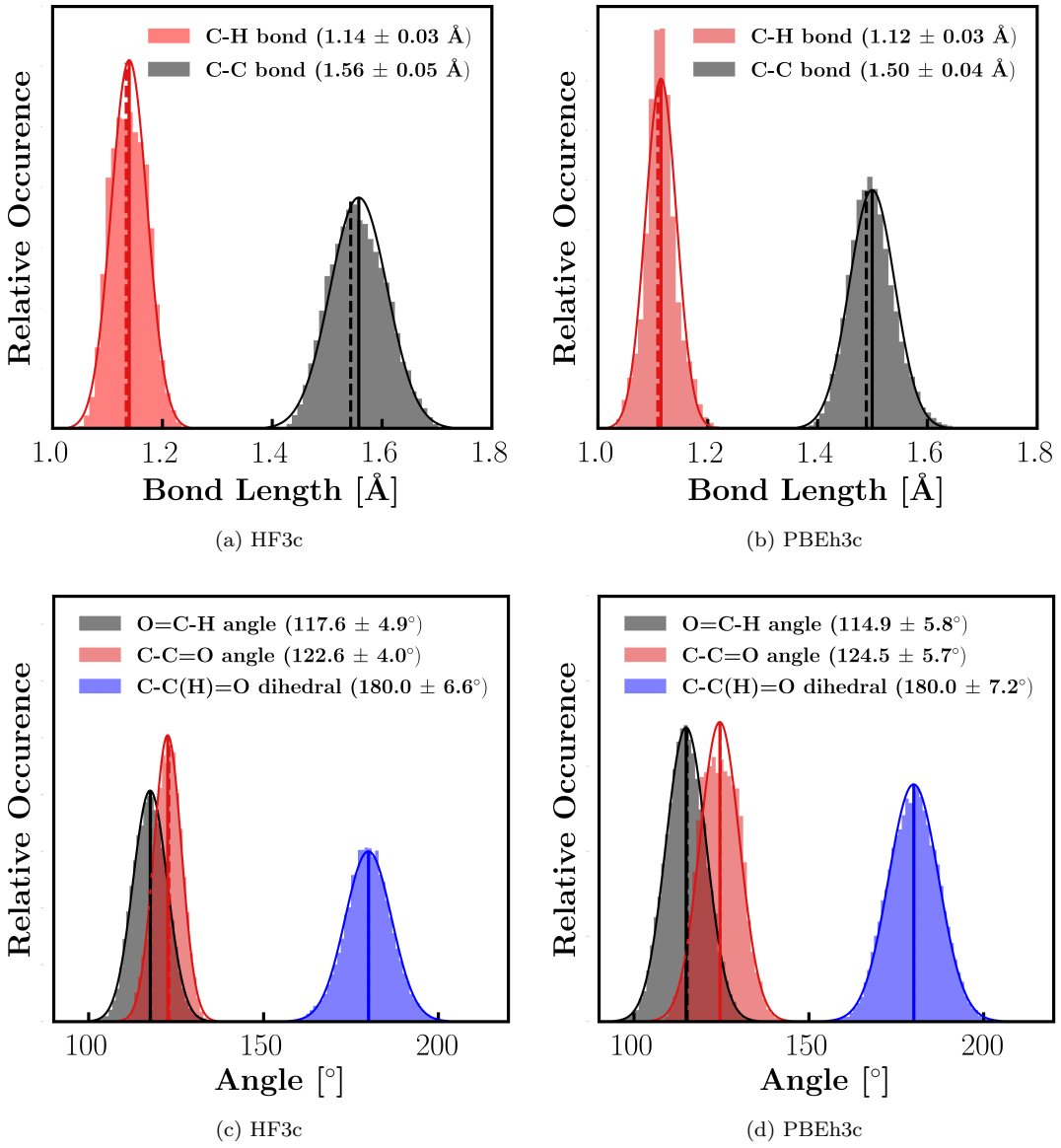
For the explicit solvent simulation, TEMPO was parametrized with Antechamber [77], which is part of AMBER16 [78], assigning GAFF parameters [79] and AM1 charges [80]. A charge of -1 had to be assumed as Antechamber can only handle singlet configurations. TEMPO was then solvated in a box of SPC/E solvent [81] with a 15 Å distance between the solute and the box faces. The system was neutralized with one sodium ion, placed far away from the TEMPO molecule. We then carried out 20000 steps of steepest decent energy minimization, 30 ps of heating, and a 200 ps equilibration run under NVT conditions using the NAMD engine [82] and periodic boundary conditions. All bonds to hydrogen atoms were kept fixed with SHAKE [83] and the time step was 2 fs. The temperature was kept constant using the Langevin-Piston thermostat with a damping constant of 1 ps^{-1}. Non-bonded interactions were cut-off at 12 Å and smoothly switched off starting at 10 Å. A Verlet-nearest neighbor list was used with a radius of 13.5 Å. Periodic electrostatic interactions were evaluated with the Particle Mesh Ewald method and a polynomial interpolation of order 6. The non-bonded interactions were evaluated at every step and a full electrostatic calculation was performed every second step.

The last frame of the equilibration run was centered on the TEMPO radical. In the subsequent QM/MM-AIMD all water molecules within 4 Å of the hydroxylated TEMPO molecule were treated quantum mechanically (TEMPO: 30 atoms, water: 198 atoms), whereas all other water molecules were kept frozen to prevent a mixing of QM and MM waters. A 5 ps QM/MM-AIMD was performed using the same settings as before, with the exception of the time step of 0.25 fs. The first picosecond was discarded as equilibration phase. We extracted again 100 frames equally spaced in time for HFCC computations and all presented results are normal averages over these frames.

4. Results

A variety of contributions to the accuracy of *in silico* HFCCs were studied in this work, namely the effect of dynamics (Δ_{dyn}), electron correlation (Δ_{corr}), and solvation (Δ_{solv}). Figure 1 illustrates our definition of the two former contributions, a definition of the latter can be found in Figure 7. The dynamic contribution Δ_{dyn} is defined as the difference of the average HFCCs from a set of structures obtained from an AIMD-simulation to the static HFCC of the optimized structure both in vacuum.

We start our analysis by illustrating the influence of Δ_{dyn} and Δ_{corr} independent of solvation effects, first in-depth on two systems and subsequently on our benchmark set of 12 organic radicals. Solvation will be studied in Sec. 4.3.



(a) HF3c (b) PBEh3c

(c) HF3c (d) PBEh3c

Figure 2. Distributions of bond lengths, bond angles, and dihedral angles in the AIMD simulation of the ethanal radical cation at room temperature at the HF3c- and PBEh3c-level. Average bond lengths, bond angles, and dihedral angles are given with their standard deviation in the legend. The average is indicated with a straight line, a normalized Gaussian distribution with the same standard deviation is centred around the mean of the distribution. The value of the minimum energy geometry is shown as a dashed line.

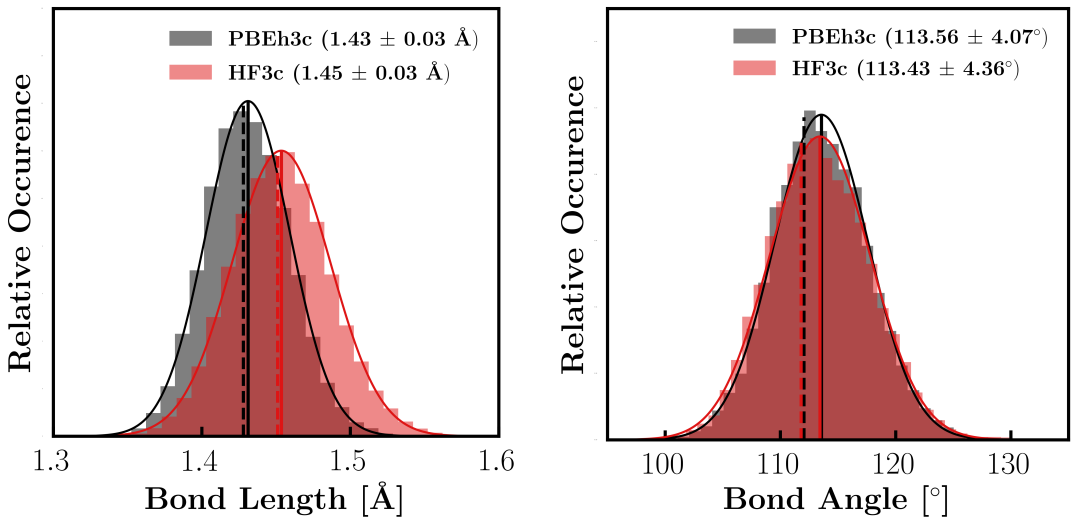


Figure 3. Distributions of the N-C bond length and and the ∠(C-N-C) bond angle in the AIMD simulation of the dimethylamino radical at room temperature at the HF3c- and PBEh3c-level. The average bond length and bond angle and the standard deviation is indicated. The average is indicated with a straight line, a normalized Gaussian distribution with the same standard deviation is centred around the mean of the distribution. The value of the minimum energy geometry is shown as a dashed line.

4.1. The dimethylamino radical and the ethanal radical cation

In this section, we study in detail the dependence of the HFCCs on bond lengths, bond angles, and dihedral angles of both the dimethylamino radical and the ethanal radical cation.

We first analyse the distribution of the bond angles, bond lengths, and the dihedral angle in the ethanal radical cation as obtained from both a HF3c- and a PBEh3c-based AIMD. The results are shown in Figure 2. Anharmonicity can clearly be seen by visual comparison with Gaussian distributions.

In general, the distributions are similar between the HF3c- and the PBEh3c-based AIMD. In the case of the ethanal radical cation, the distribution of all bond angles are broader with PBEh3c. The bond length variations are less than 0.1 Å for all bonds and both methods. The changes in the mean bond angles are significantly larger than for the dimethylamino radical which shows only deviations less than 1°.

Irrespective of the close similarities of the mean bond lengths and angles, specific care has to be taken as to what level of theory is employed to describe the dynamic contributions, i.e. an accurate description of the potential energy surface of the system is paramount. This is shown in Table 1, where the isotropic HFCCs of both systems obtained from the HF3c- and PBEh3c-optimized structures as well as from averaging over the respective AIMD simulations are compared. As can be seen, considerable deviations (larger than 20 MHz) occur in the case of the hydrogen atom between the results obtained with structures obtained at the HF3c-level in comparison to HFCCs of PBEh3c-optimized structures. The deviation is larger than the influence of the dynamics, as shown in Table 1, where the HFCCs obtained from the respective AIMD simulation are depicted. It has to be noted, though, that (i) the deviation due to using the HF3c- instead of PBEh3c-MD simulations is partially less severe than the effect of differing methods used for the HFCC computation shown in Table 1, and that (ii) the description of dynamic contributions between HF3c and PBEh3c is comparable,

Table 1. Isotropic HFCCs in MHz computed at the three levels of theory (B3LYP, B2PLYP, and MP2) of the dimethylamino radical and the ethanal radical cation using the HF3c-optimized and the PBEh3c-optimized structures as well as structures obtained from respective AIMD-simulations. Dynamic contributions Δ_{dyn} (see Figure 1) are also shown. Experimental results are available for three nuclei. ^{1}H (HCO): 381 [84], ^{14}N: 41.4 [85], ^{1}H: 76.7 [85]

Nucleus	Opt. structure			AIMD simulation			Δ_{dyn}	
	HF3c	PBEh3c	Δ_{method}	HF3c	PBEh3c	Δ_{method}	HF3c	PBEh3c
B3LYP								
Ethanal radical								
^{17}O	-40.8	-43.6	-2.8	-40.1	-43.1	-3.0	0.7	0.5
^{13}C (HCO)	40.7	42.9	2.1	40.5	43.3	2.8	-0.2	0.4
^{13}C (CH_3)	-83.1	-73.4	9.7	-85.7	-75.4	9.7	-2.6	-2.0
^{1}H (HCO)	364.8	335.5	-29.3	377.7	348.6	-29.1	12.9	13.1
Dimethylamine radical								
^{14}N	35.4	34.7	-0.7	36.0	35.4	-0.6	0.6	0.7
^{13}C	-29.9	-32.0	-2.1	-30.4	-32.5	-2.1	-0.5	-0.5
^{1}H	72.7	80.0	7.3	75.2	82.6	7.4	2.5	2.6
RI-CDD B2PLYP								
Ethanal radical								
^{17}O	-50.6	-54.6	-4.0	-49.3	-53.7	-4.4	1.3	0.9
^{13}C (HCO)	44.3	47.3	3.0	43.5	47.1	3.6	-0.8	-0.2
^{13}C (CH_3)	-97.7	-85.5	12.2	-100.4	-87.9	12.5	-2.7	-2.4
^{1}H (HCO)	374.5	339.6	-34.9	388.6	355.0	-33.6	14.1	15.4
Dimethylamine radical								
^{14}N	40.8	40.1	-0.7	41.5	41.0	-0.5	0.7	0.9
^{13}C	-33.5	-36.0	-2.5	-34.1	-36.5	-2.4	-0.6	-0.5
^{1}H	71.2	78.4	7.2	73.7	80.9	7.2	2.5	2.5
RI-CDD MP2								
Ethanal radical								
^{17}O	-45.6	-49.3	-3.7	-43.7	-48.4	-4.7	1.9	0.9
^{13}C (HCO)	48.2	51.0	2.8	46.3	50.4	4.1	-1.9	-0.6
^{13}C (CH_3)	-102.5	-88.3	14.2	-105.6	-91.1	14.5	-3.1	-2.8
^{1}H (HCO)	337.8	298.6	-39.2	355.1	315.0	-40.1	17.3	16.4
Dimethylamine radical								
^{14}N	36.9	36.8	-0.1	37.7	37.7	0.0	0.8	0.9
^{13}C	-31.7	-34.2	-2.5	-31.6	-34.0	-2.4	0.1	0.2
^{1}H	63.5	69.8	6.3	65.5	71.9	6.4	2.0	2.1

as can be seen from the Δ_{dyn} values in Table 1. The latter motivates the computation of the dynamic contributions based on a HF3c-based AIMD trajectory.

Though dynamic contributions result in changes in the HFCCs, these are on the same order as the differences between the results obtained from DFT, post-Kohn-Sham (KS), and post-HF methods in Table 1. While this is not always the case, as shown in Section 4.2, where the neglect of dynamic contributions leads to results strongly deviating from experimental findings for a variety of radicals, the results in Table 1 motivate a careful choice of the method with which the isotropic HFCCs are computed. Furthermore, it has to be noted that the results obtained from the AIMD simulation are not always closer to the experimental results. Two reasons can be named for this: first and foremost the error of the approximation to the Schrödinger equation itself and, second, the exact experimental conditions were not sufficiently replicated, therefore, including solvent effects can be crucial.

The reason for the significant dynamic contributions, apart from the anharmonicity in the distribution of the structural parameters in the AIMD simulation, is the strong and non-linear dependency of the HFCCs on the bond lengths, bond angles, and

dihedral angles. In order to investigate this relationship systematically, we start with a PBEh3c-optimized structure of both the dimethylamino radical and the ethanal radical cation. Subsequently, we modify one structural parameter, i.e. the bond length, bond angle, or dihedral angle, at a time, and compute the isotropic HFCCs of all nuclei for each structure using the hybrid functional B3LYP, the DH-DFT method RI-CDD B2PLYP, HF, RI-CDD MP2, and CCSD.

The results for the ethanal radical cation are shown in Figure 4. Noteworthy is the strong dependence of the isotropic HFCCs on the geometry, especially for the hydrogen nuclei. While for most nuclei the dependency of the HFCC on the structural parameters is consistent throughout the different methods studied in this work, the importance of correlation can be deduced from the differing results obtained with HF. This is most apparent in the dependency of the ^{13}C (H$\underline{\text{C}}$O)-HFCC on the angle ∠(C-C=O) and the angle ∠(O=C-H), as well as on both bond lengths. In general, the HF results deviate substantially from the results obtained with the other methods. Thus, irrespective of dynamic contributions, HF computations are incapable of correctly describing the spin density in the ethanal radical cation, so that methods incorporating electron correlation are required, which is reflected in the high amount of spin contamination found in HF calculations. Furthermore, strong non-linear dependencies indicate that dynamic contributions will change the *in silico* results considerably.

Similar results are obtained with the dimethylamino radical in Figure 5, except for the description of the hydrogen atoms, where HF is found in between the other methods which take electron correlation into account. HFCCs of both the carbon and the nitrogen nucleus strongly depend, however, on electron correlation.

One can make two general observations in Figures 4 and 5. For one, HF and B3LYP almost always give extreme values for the HFCCs and electron correlation methods lie in-between. Secondly, no method is unfortunately the closest to CCSD for every nucleus, where we would expect CCSD to be the method that describes electron correlation best. The fact that DFT is the other extreme compared to HF is in agreement with findings that HF yields very localized and DFT usually very delocalized singly occupied molecular orbitals (SOMOs) [86].

4.2. Study of a set of organic radicals

In order to determine both the influence of electron correlation and dynamic contributions on the accuracy of HFCCs, we computed the HFCCs based on AIMD simulations at the PBEh3c-level of theory for a set of organic radicals shown in Figure 6.

4.2.1. Convergence with the number of frames

Prior to comparing the results of all radicals, we analysed in detail the convergence with respect to the number of MD frames for which the HFCCs were computed. This is shown in SFigure 1 for a selection of three radicals (**2**, **3**, and **11**). Converged results require more than 5 ps (here 50 frames). Noteworthy is radical **3**, where a significant standard deviation is apparent, especially in the HFCCs of the hydrogen atoms. This can be explained by a Jahn-Teller distortion of the CH-bonds, which will be discussed in detail in Section 4.2.2.

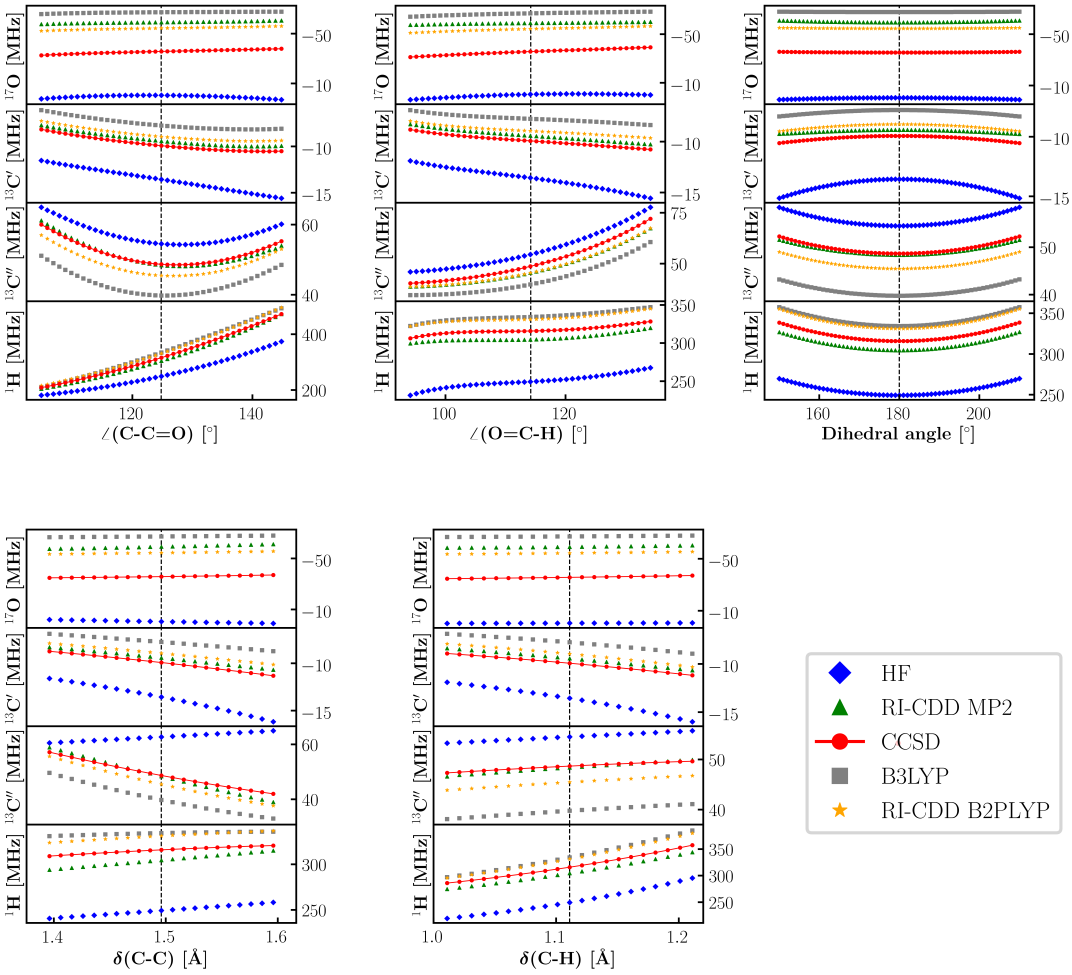


Figure 4. Dependency of the isotropic HFCCs of the different nuclei in the ethanal radical cation on the bond angles $\angle(C-C=O)$ and $\angle(O=C-H)$, the bond lengths δ(C-C) and δ(C-H), and the dihedral angle. The PBEh3c-based geometry optimization (vertical dashed line) leads to $\angle(C\text{-}C{=}O) = 124.8°$, $\angle(O=C-H) = 114.2°$, $\delta(C-C) = 1.496$ Å, $\delta(C-H) = 1.111$ Å, and $\angle$(dihedral) $= 180.0°$. All computations were performed with the def2-TZVPP basis set. The CCSD values were joined by a line to guide the eye. On the y-axis, ^{13}C is a shorthand for ^{13}C (C̲HO) and ^{13}C for ^{13}C (C̲H$_3$).

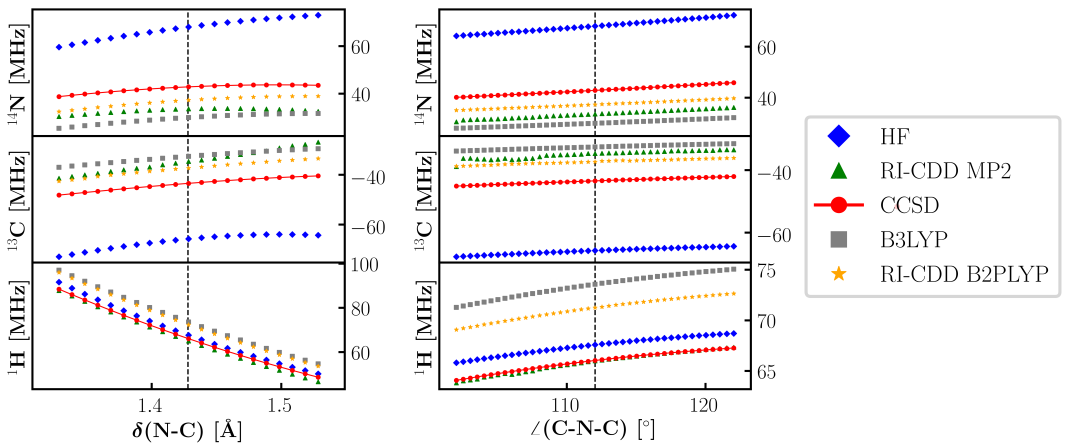


Figure 5. Dependency of the isotropic HFCCs of the different nuclei (carbon, nitrogen, and hydrogen) in the dimethylamino radical on the δ(N-C) bond length and the $\angle$(C-N-C) bond angle. The PBEh3c-based geometry optimization (vertical dashed line) results in δ(N-C) of 1.428 Å and $\angle$(C-N-C) of 112.04°. All computations were performed with the def2-TZVPP basis set. The CCSD values were joined by a line to giude the eye.

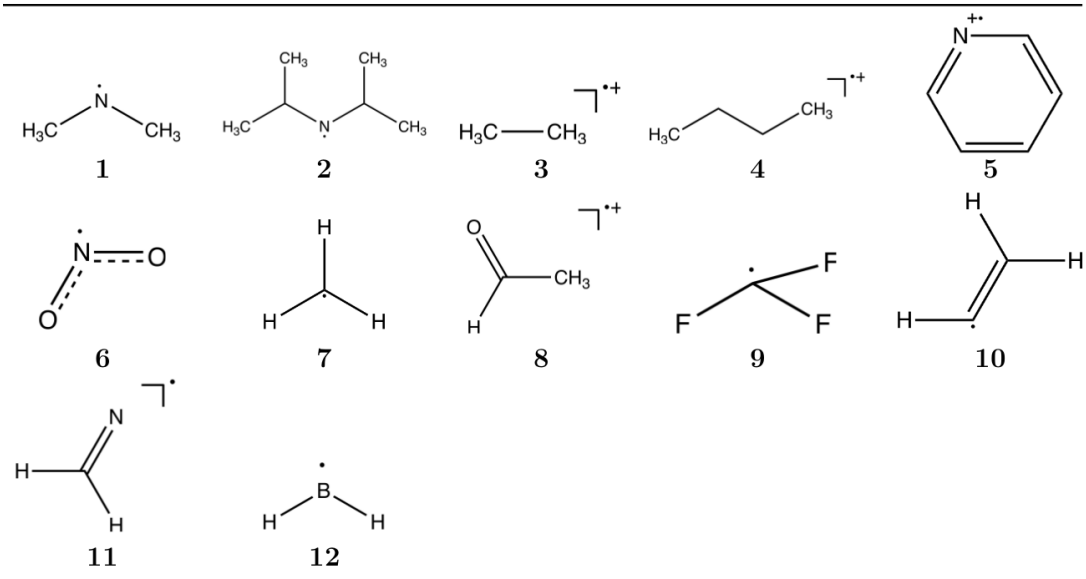


Figure 6. Organic radicals investigated in this work.

Table 2. Average and maximum correlation contributions Δ_{corr} (defined as the difference to the respective KS- or HF-result) and dynamic contributions Δ_{dyn} as indicated in Figure 1 in the computation of isotropic HFCCs of the radicals in Figure 6 excluding erratic RI-CDD-MP2 results due to spin contamination. All results are in MHz.

	B3LYP	RI-CDD-B2PLYP	RI-CDD-MP2
$< \Delta_{corr} >$	—	9.4	26.9
$\max(\Delta_{corr})$	—	39.8	102.2
$< \Delta_{dyn} >$	10.7	11.3	30.9
$\max(\Delta_{dyn})$	157.9	154.3	151.4

4.2.2. Molecular dynamic and electron correlation contributions

The isotropic HFCCs of all organic radicals shown in Figure 6 were calculated with the PBEh3c-optimized structure and from the respective PBEh3c-AIMD simulation with B3LYP, RI-CDD B2PLYP, and RI-CDD MP2. The average dynamic and correlation contributions are shown in Table 2, highlighting that both are significant for the overall obtained HFCCs. Taking electron correlation into account significantly changes the computed HFCCs. This is most apparent by comparing the HF- and RI-CDD-MP2 results, which deviate by 26.9 MHz on average. Similarly, the correlation contribution at the DH-DFT-level is significant. The larger magnitude of the correlation effect at the RI-CDD-MP2-level than at the RI-CDD-B2PLYP-level can be explained by the known overestimation of correlation effects by MP2, the scaling of 0.27 present in the B2PLYP functional, and by the fact that electron correlation is also partially included in the hybrid-DFT ansatz. Overall, molecular dynamic contributions change the computed HFCCs on average by 10.7 MHz for B3LYP, by 11.3 MHz for RI-CDD-B2PLYP, and by 30.9 MHz for RI-CDD-MP2. For the latter method, we disregarded the systems with high spin contamination. While the dynamic contributions are on the same order as the correlation contributions, a detailed analysis discloses that for a variety of systems the inclusion of dynamic effects is indispensable for the accurate description of the system. The larger magnitude of the dynamic contribution in the case of RI-CDD-MP2 can be attributed to the inaptitude of HF-determinants for the description of the potential energy surface of these organic radicals. It is further instructive to look at the maximum contribution of both effects, as shown in Table 2. It clearly shows, that the neglect of those contributions can in severe cases lead to meaningless results, which is especially valid for the dynamic contributions.

A detailed presentation of the results of all radicals is shown in Table 3, where we also state experimental results if available. While a comparison to the experimental results is instructive, caution is warranted as the AIMD simulations do not necessarily replicate the experimental conditions. The experimental conditions include a variety of temperatures and solvents/matrices, whereas all AIMD simulations were performed at room temperature in the gas phase for the sake of reliably comparing the influence of dynamic and correlation contribution across all radicals in Figure 6.

At this point, we want to mention that the signs indicated in the experimental results are not directly obtained from the experiment, but are assigned afterwards using theoretical results. In the cases where the sign obtained with B3LYP, RI-CDD-B2PLYP, and RI-CDD-MP2 deviates from the corresponding reported values based in the experimental column, our absolute values match with the experimental results and are consistent. We thus conclude that the original assignment of the sign could be erroneous in line with previous theoretical work [87, 88]. We hence indicate our

proposed sign in parentheses.

It has to be noted that some of the radicals exhibit energetically close lying minima. In radical **5**, e.g. geometry optimizations with PBEh3c lead to either a radical with A_1 or with B_1 symmetry. The A_1 symmetry structure has a high localization of the spin density at the nitrogen nucleus with an isotropic HFCC of 110.3 MHz versus -4.3 MHz obtained from the B_1 structure (both results were obtained with the B3LYP functional). Our PBEh3c-AIMD simulation confirms that we stay on the A_1 potential energy surface throughout the simulation, as the HFCCs obtained from 100 AIMD frames range from 73.9 MHz to 151.5 MHz. Such artefacts can be circumvented by verifying that the potential energy of the optimized start structure is the lower bound within the whole simulation.

In most cases, including electron correlation via RI-CDD-MP2 significantly improves the agreement between the *in silico* HFCCs and the experimental results (see ^{14}N-HFCC in radical **1** and **2**, ^{19}F-HFCC in radical **9**, ^{11}B-HFCC in radical **12**, and the ^{1}H-HFCCs in radicals **3**, **8**, and **12**). In this analysis, we disregarded the radicals with extensive spin contamination, i.e. radicals **5**, **10**, and **11**. These systems are more accurately described at the DH-DFT-level. In general, DH-DFT seems to yield better results than MP2 with respect to the experimental values (see, e.g. ^{14}N-HFCC in radicals **1**, **2**, **5**, **6**, and **11**, ^{19}F-HFCC in radical **9**, ^{13}C-HFCC in radicals **7**, **10**, and **11**, and ^{1}H-HFCCs in radicals **7**, **8**, **10**, **11**, and **12**).

While the choice of optimal double hybrid functionals is challenging, recent work [89] showcases that a correct determination of the optimal HF exchange contribution as well as spin component scaling in the second-order correlation contribution can significantly improve the computational results. In this work, we chose the two well-established functionals B3LYP and B2PLYP. Comparing the two functionals is complicated, as they vary in their HF-exchange contributions. Thus, differences can not directly be related to the additional treatment of electron correlation. However, B2PLYP significantly improves the *in silico* results in most cases with respect to the B3LYP results (see, e.g. for the ^{14}N-HFCC in radicals **1**, **2**, **5**, and **11**, the ^{19}F-HFCC in radical **9**, the ^{13}C-HFCCs in radicals **10** and **11**, and the ^{1}H-HFCCs in radicals **1**, **2**, **3**, **4**, **7**, and **10**).

As shown in Table 2, the neglect of dynamic contributions can be severe. The large magnitude of the dynamic contribution can be seen in the hyperfine splitting of the alkane radical cations, namely of the ethane radical **3** and the butane radical **4**. These represent special cases, where a Jahn-Teller distortion breaks the symmetry of the six CH-bonds at the terminal methyl groups [90–92]. This can be verified experimentally by looking at the low temperature EPR spectrum of the ethane radical cation, which exhibits a triplet splitting due to a localization of the spin density at two equivalent hydrogen atoms [90, 91]. Moving to higher temperatures, dynamic contributions lead to a septet splitting, as on average all six hydrogen atoms have become equivalent [90, 91]. This is not due to a disappearance of the Jahn-Teller distortion, but due to this effect becoming dynamic. Our results confirm this behavior, as two C-H bonds are shorter (1.07 $\pm$ 0.03 Å) than the remaining four (1.13 $\pm$ 0.03 Å) in the AIMD simulation. We thus expect a significant change in the obtained HFCCs when dynamics are considered via an AIMD simulation. We can confirm the low-temperature splitting in our computations of radical **3**, where we obtain HFCCs close to the experimental results. When moving towards higher temperatures, our computations confirm the septet splitting.

Table 3.: Influence of dynamic contributions and electron correlation on the HFCCs in MHz of radicals **1-12** in Figure 6. HFCCs are obtained (1) using the PBEh3c-optimized structure and (2) from averaging 100 frames from PBEh3c-AIMD simulations at 298.15 K. Experimental results are shown where available obtained from varying experimental conditions. Erroneously attributed signs according to our computations are indicated in parentheses. The electron correlation contribution estimated via second-order perturbation theory is shown in parentheses in the *in silico* columns. Erratic RI-CDD MP2 results due to spin contamination are omitted, indicated by the respective average $< \hat{S}^2 >$ value (the threshold is 0.8, the mean spin contamination of the accepted calculations is 0.76).

Nucleus	B3LYP		RI-CDD B2PLYP				RI-CDD MP2				exptl.
	1	2	1		2		1		2		
					Radical **1**						
^{14}N	34.7	35.3	40.8	(-9.4)	41.6	(-9.6)	30.8	(-41.1)	37.6	(-36.3)	41.4 [85]
^{13}C	-32.0	-32.7	-36.4	(6.4)	-37.3	(6.6)	-34.1	(30.8)	-34.3	(32.4)	—
^{1}H[a]	119.9 0.1	82.6	116.7 0.0	(3.7) (-1.8)	80.3	(1.9)	100.8 -8.0	(-5.1) (-13.8)	71.8	(-3.4)	76.7 [85]
					Radical **2**						
^{14}N	34.5	34.9	40.8	(-9.1)	41.2	(-9.4)	63.7	(-8.5)	34.3	(-39.0)	40.1 [85]
^{13}C (C$\underline{\text{H}}$)	-26.8	-27.1	-30.0	(7.4)	-30.3	(7.8)	-16.7	(42.1)	-25.8	(35.0)	—
^{13}C[b] (C$\underline{\text{H}}_3$)	70.0 14.0	41.1	69.9 13.9	(1.5) (-0.6)	40.7	(0.3)	91.9 50.0	(24.5) (-9.8)	37.3	(-4.2)	— —
^{1}H (C$\underline{\text{H}}$)	34.2	46.9	33.1	(0.0)	46.0	(1.2)	0.3	(-33.7)	37.7	(-5.6)	40.1 [85]
^{1}H[b] (C$\underline{\text{H}}_3$)	22.2 -2.8	2.2	21.1 -3.0	(2.2) (0.5)	1.8	(0.9)	8.5 -18.3	(-5.4) (-13.1)	2.2	(3.8)	1.8 [85]
					Radical **3**						
^{13}C	20.5	8.6	17.7	(-2.6)	4.7	(-0.2)	6.3	(-14.1)	-5.7	(-7.1)	—
^{1}H[c]	471.0 -19.7	149.5	461.6 -23.3	(4.8) (5.0)	145.7	(1.0)	445.4 -26.2	(23.2) (17.6)	136.8	(12.7)	427.4 141.0 [90, 91]
					Radical **4**						
^{13}C (C$\underline{\text{H}}_3$)	-9.9	-10.9	-12.1	(0.5)	-13.0	(9.4)	-10.2	(8.2)	-13.8	(6.7)	—
^{13}C (C$\underline{\text{H}}_2$)	8.9	13.1	7.3	(2.5)	11.6	(1.5)	6.4	(0.1)	0.0	(-12.3)	—
^{1}H (C$\underline{\text{H}}_2$)	9.4	13.6	7.7	(3.1)	11.8	(4.0)	10.3	(20.1)	8.5	(15.6)	—
^{1}H[d] (C$\underline{\text{H}}_3$)	214.2	154.3	197.5	(14.0)	142.3	(9.5)	179.7	(34.5)	125.9	(19.6)	171.8 [90, 91]
	20.9	39.8	18.6	(1.2)	36.1	(3.0)	14.0	(0.4)	30.5	(6.9)	22.4 [90, 91]
					Radical **5**						
^{14}N	110.3	110.6	117.2	(-20.6)	116.4	(-20.9)					114.9 [93]
o-^{13}C	35.1	33.8	38.5	(-11.5)	37.5	(-14.2)					—
m-^{13}C	-14.0	-11.9	-15.9	(10.2)	-14.5	(12.1)					—
p-^{13}C	-10.5	-12.1	-10.7	(8.3)	-12.1	(9.2)	$< \hat{S}^2 > =$ 1.1997		$< \hat{S}^2 > =$ 1.2345		—
o-^{1}H	93.9	91.4	89.9	(4.6)	88.0	(3.5)					82.1 [93]
m-^{1}H	35.5	31.2	32.2	(9.5)	28.4	(10.9)					31.3 [93]
p-^{1}H	22.5	23.7	16.4	(1.8)	16.7	(-2.1)					24.1 [93]
					Radical **6**						
^{14}N	143.5	142.5	146.8	(0.5)	146.4	(2.5)	145.9	(-7.7)	134.1	(-15.8)	152.0 [94]
^{17}O	-58.2	-57.0	-68.1	(-7.0)	-66.9	(-7.7)	-70.9	(3.5)	-69.5	(-3.5)	(-)62.2 [94]

Continued on next page

[a]The non-dynamic computations exhibit four and two equivalent hydrogen atoms, leading to two separate hydrogen HFCCs. Averaging over these HFCCs results in 80.0 MHz (B3LYP), 77.8 MHz (RI-CDD-B2PLYP), and 63.9 MHz (RI-CDD-MP2).

[b]Both ^{13}C (C$\underline{\text{H}}_3$) and ^{1}H (C$\underline{\text{H}}_3$) exhibit differences between the HFCCs of the nuclei in the non-dynamic computations. Averaging results in 42.0 MHz (B3LYP), 41.9 MHz (RI-CDD-B2PLYP), and 71.0 MHz (RI-CDD-MP2) for ^{13}C (C$\underline{\text{H}}$3) and 2.7 MHz (B3LYP), 2.2 MHz (RI-CDD-B2PLYP), and -9.8 MHz (RI-CDD-MP2) for ^{1}H (C$\underline{\text{H}}_3$).

[c]At low temperatures (4 K), the experimental results show a triplet splitting of 427.4 MHz, whereas at higher temperatures (77 K) a septett splitting of 141.0 MHz is obtained. The *in silico* results match this splitting pattern.

[d]Similar to radical **10**, a Jahn-Teller distortion can be observed at low temperatures. From the PBEh3c-optimized structure, two distinct hydrogen HFCCs of the methyl groups can be observed.

Table 3 – *Continued from previous page*

Nucleus	B3LYP		RI-CDD B2PLYP		RI-CDD MP2		exptl.
	1	2	1	2	1	2	
			Radical **7**				
^{13}C	79.8	114.6	85.1 (-26.0)	120.0 (-26.3)	59.3 (-101.5)	94.8 (-102.2)	107.4 [95]
^{1}H	-64.4	-58.4	-70.1 (13.0)	-64.0 (12.5)	-71.9 (49.8)	-65.9 (48.6)	(-)64.6 [95]
			Radical **8**				
^{17}O	-43.6	-43.1	-57.2 (20.5)	-60.8 (16.7)	-48.8 (79.4)	-45.5 (83.2)	—
^{13}C (H$\underline{\text{C}}$O)	42.9	43.2	48.4 (-1.8)	47.6 (-2.3)	51.3 (-5.5)	52.7 (-4.2)	—
^{13}C ($\underline{\text{C}}$H$_3$)	-73.4	-75.4	-84.6 (13.3)	-84.1 (18.1)	-89.3 (41.1)	-111.1 (26.9)	—
^{1}H ($\underline{\text{H}}$CO)	335.5	356.3	329.3 (33.4)	355.0 (37.3)	301.0 (51.9)	334.2 (65.7)	381 [91]
^{1}H (C$\underline{\text{H}}_3$)	-3.0	-0.6	-5.3 (3.4)	0.5 (7.9)	-8.9 (-5.7)	-23.8 (-9.5)	-8 [91]
			Radical **9**				
^{13}C	728.0	737.1	745.6 (-22.5)	754.9 (21.8)	750.1 (-67.6)	762.2 (-64.9)	—
^{19}F	393.3	390.4	408.7 (-3.4)	406.5 (-4.5)	370.1 (-47.9)	398.0 (-16.9)	405.0 [96]
			Radical **10**				
^{13}C ($\underline{\text{C}}$H)	314.5	296.5	321.6 (-38.4)	304.1 (-39.8)			301.5 [97, 98]
^{13}C ($\underline{\text{C}}$H$_2$)	-11.9	-17.0	-17.3 (16.6)	-22.8 (18.0)			-24.1 [97, 98]
^{1}H (C$\underline{\text{H}}$)	53.2	44.6	46.0 (12.8)	37.4 (13.5)	$< \hat{S}^2 > =$ 0.9399	$< \hat{S}^2 > =$ 0.9459	38.7 [97, 98]
^{1}H Z-(C$\underline{\text{H}}_2$)	178.7	182.5	177.9 (2.8)	181.9 (2.2)			184.8 [97, 98]
^{1}H E-(C$\underline{\text{H}}_2$)	113.8	120.7	112.7 (-0.4)	119.6 (0.1)			111.0 [97, 98]
			Radical **11**				
^{13}C	-68.0	-70.2	-77.7 (19.3)	-80.0 (21.8)			81.1 [99]
^{14}N	23.1	23.1	27.2 (-9.1)	27.2 (-9.4)	$< \hat{S}^2 > =$ 0.9356	$< \hat{S}^2 > =$ 0.9410	28.6 [99]
^{1}H	235.7	240.0	235.4 (8.3)	239.7 (7.3)			244.8 [99]
			Radical **12**				
^{11}B	363.7	342.0	366.5 (-14.9)	345.0 (-15.5)	357.7 (-43.1)	336.0 (-45.8)	358 [100]
^{1}H	42.5	35.6	39.5 (5.2)	32.4 (5.8)	26.1 (17.9)	19.4 (19.9)	38 [100]

In the case of the butane radical cation, a triplet splitting is also observed experimentally at 77 K. We can confirm this splitting in both our PBEh3c-optimized structure and in the results based on our AIMD simulation, where we see two discrete hydrogen-HFCCs with the larger component arising from two hydrogen atoms leading to the observed triplet splitting. While the inclusion of dynamic contributions leads to a reduction of the HFCCs in the direction of the experimental values, they overshoot, which is in line with our room temperature simulation resulting in a larger reduction.

The HFCCs obtained from the PBEh3c-optimized structures of radicals **1** and **2** show that the spin density is primarily located at four of the six hydrogen atoms. An analysis of the C-H bond lengths shows in analogy to the ethane radical that two C-H bonds are shorter than the remaining four (1.089 Å vs. 1.098 Å). When MD contributions are considered, this distortion is averaged resulting in HFCCs that closely match the experimental results. Notwithstanding, similar results can be obtained from averaging the six hydrogen-HFCCs obtained with the PBEh3c-optimized structure. We therefore considered dynamic contributions in our overall analysis in Table 2 only by comparing to the averaged results. The effect is analogous in radical **2**.

Another example of extensive dynamic contributions is the methyl radical **7**. Optimizing the structure leads to a planar configuration. When vibrational averaging is considered, non-planar configurations also contribute to the overall HFCCs, which leads to considerable changes of up to 40 % and to a better agreement with the experimental findings. It is apparent, that the spin density at the carbon atom increases, whereas the spin density at the hydrogen nuclei decreases considerably for B3LYP, RI-CDD-B2PLYP, and RI-CDD-MP2.

Table 4. Comparison of the HFCCs in MHz obtained from the HF3c- (opt_{HF3c}) and PBEh3c-optimized structure (opt_{PBEh3c}) and from HF3c- (dyn_{HF3c}) and PBEH3-simulations (dyn_{PBEh3c}) to the combined HF3c/PBEh3c-approach (dyn_{comb}). The latter HF3c/PBEh3c approach consists of adding the dynamical correction as obtained from an HF3c-AIMD simulation to the HFCCs obtained with the PBEh3c-optimized structure.

Nucleus	opt_{HF3c}	opt_{PBEh3c}	dyn_{HF3c}	dyn_{PBEh3c}	Δ_{method}(opt)	Δ_{method}(dyn)	Δ_{dyn}(HF3c)	Δ_{dyn}(PBEh3c)	dyn_{comb}
				Radical **1**					
				B3LYP					
^{14}N	35.4	34.7	35.7	35.3	-0.7	-0.4	0.3	0.6	35.0
^{13}C	-29.9	-32.0	-30.6	-32.7	-2.1	-2.1	-0.7	-0.7	-32.8
^{1}H	72.7	80.0	75.1	82.6	7.3	7.5	2.4	2.6	82.4
				B2PLYP					
^{14}N	41.5	40.8	41.9	41.6	-0.7	-0.3	0.4	0.8	41.2
^{13}C	-33.9	-36.4	-34.8	-37.3	-2.5	-2.5	-0.9	-0.9	-37.3
^{1}H	70.7	77.8	73.1	80.3	7.1	7.2	2.4	2.5	80.2
				Radical **8**					
				B3LYP					
^{17}O	-40.8	-43.6	-40.1	-43.1	-2.8	-3.0	0.7	0.5	-42.9
^{13}C (H$\underline{C}$O)	40.7	42.9	40.9	43.2	2.2	2.3	0.2	0.3	43.1
^{13}C ($\underline{C}H_3$)	-83.1	-73.4	-86.3	-75.4	9.7	10.9	-3.2	-2.0	-76.6
^{1}H ($\underline{H}$CO)	364.8	335.5	379.9	356.3	-29.3	-23.6	15.1	20.8	350.5
^{1}H (C$\underline{H}_3$)	-0.4	-3.0	-0.1	-0.6	-2.6	-0.5	0.3	2.4	-2.7
				B2PLYP					
^{17}O	-53.7	-57.2	-56.2	-60.8	-3.5	-4.6	-2.5	-3.6	-59.7
^{13}C (H$\underline{C}$O)	45.9	48.4	45.5	47.6	2.5	2.1	-0.4	-0.8	48.0
^{13}C ($\underline{C}H_3$)	-96.7	-84.6	-98.4	-84.1	12.1	14.3	-1.7	0.5	-86.3
^{1}H ($\underline{H}$CO)	363.0	329.3	383.8	355.0	-33.7	-28.8	20.8	25.7	350.1
^{1}H (C$\underline{H}_3$)	-2.8	-5.3	-0.3	0.5	-2.3	0.8	2.5	5.8	-2.6
				Radical **9**					
				B3LYP					
^{13}C	530.9	728.0	535.8	737.1	197.1	201.3	4.9	9.1	732.9
^{19}F	492.8	393.3	489.1	390.4	-99.5	-98.7	-3.7	-2.9	389.6
				B2PLYP					
^{13}C	548.8	745.6	554.3	754.9	196.8	200.6	5.5	9.3	751.0
^{19}F	508.5	408.7	503.4	406.5	-99.8	-96.9	-5.1	-2.2	403.6

4.2.3. Composite HF3c/PBEh3c approach

The results in Table 3 include dynamic contributions using the PBEh3c-AIMD simulations. Especially when moving towards larger molecular systems, the computational cost for such high-level MD simulations will become prohibitive. While the results in Table 1 indicate that the computationally cheaper HF3c method leads to deviating results, these can mostly be attributed to a differing optimized ground state structure (see Figures 2 and 3). The dynamic contributions, however, are approximated decently using the HF3c-based AIMD simulations. Thus, we investigate whether adding the Δ_{dyn} contribution at the HF3c-AIMD level to the HFCCs obtained with a PBEh3c-optimized structure can be considered a viable pragmatic approach. This is shown in the following Table 4.2.3, where we compare the results for radicals **1**, **8**, and **9** using this combined HF3c/PBEh3c-approach (dyn_{comb}) to the respective HFCCs obtained from HF3c- and PBEh3c-AIMD simulations, respectively.

The results obtained directly from the HF3c-AIMD simulation deviate significantly

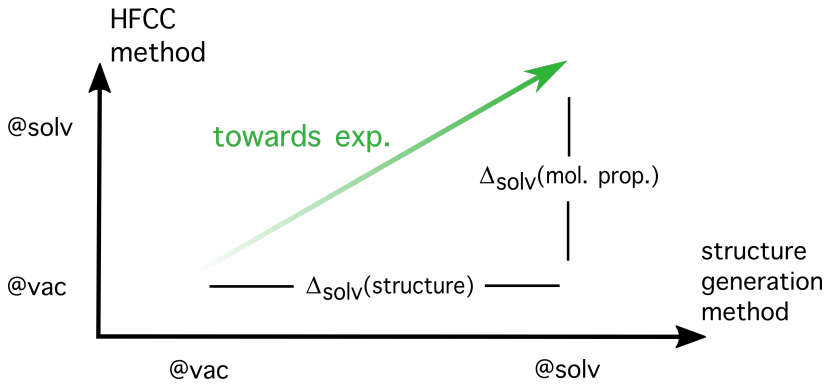


Figure 7. Graphical definitions of solvent effects, similar to Figure 1. We split the solvent effect Δ_{solv} into two components, Δ_{solv}(mol. prop.) and Δ_{solv}(structure). The structural component denotes the effect of solvation on the molecular structure. The molecular property effect describes the influence of solvation on the HFCC calculation for any given configuration.

from the respective PBEh3c-based results. This confirms the findings in Table 1. However, by adding the dynamic correction as obtained from the HF3c-AIMD simulation to the HFCCs of the PBEh3c-optimized structure, this error can be removed, and results decently incorporating the dynamical effect are achieved. This composite HF3c/PBEh3c ansatz is analogous to respective work on nuclear magnetic resonance shielding constants [101], where the dynamic contribution is obtained at a lower level of theory and added to the stationary high-level results. That such a composite approach is applicable shows a small statistical analysis performed on data from Table 3 (shown in section 4 of the SI). A drastic example where this combined approach leads to improved results in the computation of the HFCCs is radical **9**, where the HF3c geometries exhibit significantly smaller δ(C-F), and thus lead to strongly deviating results. The dynamic contribution, however, is correctly incorporated, leading to good results using the combined HF3c/PBEh3c ansatz. The results of radical **9** reinforce the necessity to perform high-level geometry optimizations in order to obtain comparable results.

4.3. Solvation

In addition to correlation effects and dynamic contributions, we consider the influence of solvation. The change in the obtained geometry by incorporation of solvent effects during geometry optimisations and dynamic simulations is called *structural* solvation effects henceforth (Δ_{solv}(structure)); the effect of including solvation (via PCM or explicit solvent molecules) in the computation of the HFCCs is termed *molecular property* solvation effect (Δ_{solv}(mol. prop.)).

When the effect of the solvent on the HFCCs is to be considered, two effects can be distinguished: (i) specific interactions of the molecule with the solvent requiring explicit solvent molecules and (ii) broader contributions that can be captured by continuum solvation models [54].

To study the influence of the solvent on the *in silico* isotropic HFCCs, we analysed the ^{14}N-HFCC of a hydroxylated TEMPO-radical. TEMPO has previously been studied in detail by Barone *et al.* [22], where including dynamic contributions as well as continuum solvent contributions by the means of COSMO was vital for the accurate description. In a very recent study, they also used explicit solvation in a QM/MM

ansatz, where TEMPO was treated quantum mechanically and the water molecules with a polarizable force field [26]. In our study of the ^{14}N-HFCC of the hydroxylated TEMPO-radical, we analyse both the influence of the solvent and dynamic contributions, considering both *molecular property* and *structural* solvation effects as defined in Figure 7 with either the C-PCM or with explicit solvent molecules within the QM/MM ansatz, where we include several solvation shells in the QM-sphere. To increase the efficiency in the computation of the ^{14}N-HFCCs in the frames obtained from the AIMD in the presence of explicit solvent molecules, our new selected-nuclei ansatz was used for the RI-CDD B2PLYP computations [48]. We obtain both Δ_{dyn} and Δ_{solv} as defined in Figures 1 and 7 at the HF3c-level and improve our results by the composite HF3c/PBEh3c scheme Δ_{method} introduced in Section 4.2.3. Other than in the work by Barone *et al.* [22, 102], we do not employ the basis set N07D or the auxiliary basis set def2-SVP-RI. We use for the sake of consistency and generality the same triple-ζ basis EPR-III and RI-basis def2-TZVPP-RI/JK as we have throughout this paper.

All solvation results are summarized in detail in STables 1 and 2. Distributions of the HFCCs computed for the MD frames within the different settings can be found in SFigures 2-4. STable 1 presents the HFCCs calculated for optimized structures, STable 2 contains averages over configurations taken from the three different AIMDs. Both tables clearly outline the solvation contributions Δ_{solv}(structure) and Δ_{solv}(mol. prop.), as well as Δ_{corr}. Also, STable 1 contains Δ_{method}(opt) and STable 2 Δ_{dyn}. The central result presented in those tables is that the contributions Δ_{solv}(mol. prop.) and Δ_{corr} are independent of the degree of solvation used during the structure generation, and Δ_{solv}(structure) and Δ_{dyn} are independent of the level of theory used to compute the HFCC. The only effect that varies with every combination of settings is Δ_{method}(opt) (maximum difference 1.5 MHz).

Unfortunately, the dissection of the various contributions to the HFCCs is incomplete as one cannot estimate the dynamic effects present in the QM/MM system, as there is no clearly defined global minimum. Thus, we can only comment on the fact that solvation (in the case of PCM) seems to dampen the dynamics, as one would guess based on chemical intuition, and we expect the same effect for the QM/MM case. The size of the *structural* solvent effect (Δ_{solv}(structure)) is in this case smaller for QM/MM than for PCM. Far more important and again aligned with our intuition is the size of the *molecular property* solvent effect. It is considerably larger in the QM/MM setting than the PCM case, as the quantum mechanically treated solvent molecules around the hydroxylated TEMPO significantly influence the electron density at the radical.

We can use these surprisingly constant contributions to extrapolate high level results from lower level calculations. Due to its costs, we have not performed PBEh3c AIMDs, but use Δ_{method}(opt) for extrapolating the result easily within 1 MHz of accuracy. Table 4.3 shows how the different increments can be added together to extrapolate an B3LYP HFCC that would result from a PBEh3c-QM/MM-MD.

One can go even further, and correct the B3LYP result with a mean Δ_{corr} of 4.4 MHz taken from the PBEh3c optimized structures to extrapolate the B2PLYP result (see STable 1). That means, a result that requires a PBEh3c-QM/MM-MD and hundreds of B2PLYP calculations can be estimated by performing one PBEh3c optimization and one B2PLYP calculations on the solute alone. This is very important as performing double-hybrid calculations on large QM systems is prohibitively expensive with regard to computation time and memory requirements.

Using this composite approach, we improve the agreement with experiment significantly from the B3LYP@vac//HF3c@vac opt. HFCC, which differs from the experi-

Table 5. Assembly of high level ^{14}N-HFCCs of the hydroxylated TEMPO-radical in MHz, taking into account Δ_{method} and Δ_{dyn} as defined in Figure 1 and the *structural* and *molecular property* solvent contributions as defined in Figure 7. A detailed analysis of the data can be found in STables 1 and 2. The $< \Delta_{\text{corr}} >$ is extracted from the optimized PBEh3c structures as difference between B3LYP and B2PLYP. The experimental result in an aqueous solution is 47.6 MHz according to Ref. [103].

Contribution	Explanation	B3LYP	B2PLYP
HF3c@vac opt.		25.5	
$< \Delta_{\text{method}} >$		$\approx +8.4$	
Δ_{solv}(structure)	$\text{QM/MM}_{\text{dyn}} - \text{vac}_{\text{dyn}}$	-0.8	
Δ_{solv}(mol. prop.)	B3LYP@QM/MM − B3LYP@vac	+4.5	
Δ_{dyn}	HF3c_{dyn}@vac − HF3c_{opt}@vac	+2.4	
$< \Delta_{\text{corr}} >$	$\text{B2PLYP//PBEh3c}_{\text{opt}} - \text{B3LYP//PBEh3c}_{\text{opt}}$	–	+4.4
PBEh3c@QM/MM		40.0	44.4

mental result by 22 MHz, to a value which is only about 3 MHz lower than the experimental result (B2PLYP@QM/MM//HF3c@QM/MM). This shows how important proper solvation (Δ_{solv}(structure) and Δ_{solv}(mol. prop.)), dynamics, electron correlation, but also the method used for structure generation are, when trying to compare with experimental results.

Thus, a thorough study on the HFCC of a radical in solution requires a set of computations following our protocol. The effect of the solvent in this specific example exceeds the effect of the dynamic contribution. Based on the results in Sec. 4.2.2 this hints at a significant contribution that needs to be evaluated with care. While computationally cheap continuum solvation models can indicate the effect of the solvent, it is also apparent that for a reliable description the inclusion of explicit water molecules is necessary.

5. Conclusion

In this work, we studied a variety of organic radicals analyzing both the effect of electron correlation and dynamics simultaneously on the accuracy of the *in silico* HFCCs. In our test set, electron correlation was shown to be a significant contributor, strongly improving the accuracy. The importance of electron correlation can especially be seen when comparing the HF results to the respective RI-CDD MP2 HFCCs. Despite the functional B3LYP leading to results agreeing reasonably well with experimental results, our findings show that further inclusion of electron correlation as within DH-DFT is beneficial and must not be neglected.

While electron correlation has to be considered for accurate results, neglecting dynamic contributions can in some cases lead to wrong results. In our test set, this especially applies to alkane radical cations where the Jahn-Teller distortion turns dynamic. Therefore, we conclude that for reliable *in silico* HFCCs both effects must be considered. When moving towards larger molecular systems using our established methodology, the cost to compute the HFCCs taking electron correlation into account can be reduced with our recently introduced efficient AO-based approach.

The last step when comparing to experiment is modelling the same molecular surrounding in the computations as were used during the measurement, e.g. the inclusion of explicit solvent molecules in a QM/MM approach. In this way, solvent effects can simultaneously and accurately be described.

The results suggest that obtaining the dynamic correction at a computationally cheaper AIMD-level using our composite HF3c/PBEh3c approach can be sufficient in many cases to capture most of the dynamic contribution, providing a good compromise between accuracy and computational cost. This also applies to the inclusion of solvation effects, where costs can be cut without loss of accuracy.

Acknowledgements

We dedicate this work to Professor Jürgen Gauß on the occasion of his 60$^{\text{th}}$ birthday. The authors thank Henryk Laqua (LMU Munich) for help and discussions with respect to the seminumerical calculations.

Disclosure statement

The authors declare no conflict of interest.

Funding

S.V. thanks the Studienstiftung des Deutschen Volkes for a graduate fellowship. C.O. acknowledges funding by the "Deutsche Forschungsgemeinschaft" (DFG, German Research Foundation) - SFB 1309 - 325871075 and Germany's Excellence Strategy - EXC 2089/1-390776260 (excellence cluster e-conversion), as well as financial support as a Max-Planck Fellow at the Max Planck Institute for Solid State Research in Stuttgart.

References

[1] G. Jeschke, Biochim. Biophys. Acta **1707**, 91–102 (2005).
[2] I.D. Sahu, R.M. McCarrick and G.A. Lorigan, Biochemistry **52**, 5967–5984 (2013).
[3] Z. Rinkevicius, L. Telyatnyk, O. Vahtras and H. Ågren, J. Chem. Phys. **121**, 7614 (2004).
[4] N. Rega, M. Cossi and V. Barone, J. Chem. Phys. **105**, 11060 (1996).
[5] M. Munzarová and M. Kaupp, J. Phys. Chem. A **103**, 9966–9983 (1999).
[6] S. Kossmann, B. Kirchner and F. Neese, Mol. Phys. **105**, 2049–2071 (2007).
[7] V. Barone and P. Cimino, Chem. Phys. Lett. **454**, 139–143 (2008).
[8] B. Kirste, Magn. Reson. Chem. **54**, 835–841 (2016).
[9] V. Barone, in *Recent Advances in Density Functional Methods* (, , 1995), p. 287.
[10] S. Kossmann and F. Neese, J. Phys. Chem. A **114**, 11768–11781 (2010).
[11] S. Grimme, J. Chem. Phys. **124**, 034108 (2006).
[12] J. Čížek, J. Chem. Phys. **45**, 4256–4266 (1966).
[13] J. Čížek, J. Paldus and L. Šroubková, Int. J. Quantum Chem. **3**, 149–167 (1967).
[14] J. Čížek and J. Paldus, Int. J. Quantum Chem. **5**, 359–379 (1967).
[15] P. Verma, A. Perera and J.A. Morales, J. Chem. Phys. **139**, 174103 (2013).
[16] M. Saitow and F. Neese, J. Chem. Phys. **149**, 034104 (2018).
[17] H.U. Suter, M.B. Huang and B. Engels, J. Chem. Phys. **101**, 7686 (1994).
[18] F. Neese, Magn. Reson. Chem. **42**, S187–S198 (2004).
[19] T. Shiozaki and T. Yanai, J. Chem. Theory Comput. **12**, 4347–4351 (2016).
[20] A. Massolle, T. Dresselhaus, S. Eusterwiemann, C. Doerenkamp, H. Eckert, A. Studer and J. Neugebauer, Phys. Chem. Chem. Phys. **20**, 7661–7675 (2018).
[21] S. Grimme, J. Chem. Theory Comput. **10**, 4497–4514 (2014).

[22] V. Barone, P. Cimino and A. Pedone, Magn. Reson. Chem. **48**, S11–S22 (2010).
[23] E. Stendardo, A. Pedone, P. Cimino, M.C. Menziani, O. Crescenzi and V. Barone, Phys. Chem. Chem. Phys. **12**, 11697–11709 (2010).
[24] J.R. Asher and M. Kaupp, Chem. Phys. Chem. **8**, 69–79 (2007).
[25] Z. Rinkevicius, N.A. Murugan, J. Kongsted, B. Frecus, A.H. Steindal and H. Agren, J. Chem. Theory Comput. **7**, 3261–3271 (2011).
[26] T. Giovannini, P. Lafiosca, B. Chandramouli, V. Barone and C. Capelli, J. Chem. Phys. **150**, 124102 (2019).
[27] H. Tachikawa, M. Igarashi and T. Ishibashi, Chem. Phys. Lett. **352**, 113–119 (2002).
[28] M. Igarashi, T. Ishibashi and H. Tachikawa, J. Mol. Struct. **594**, 61–69 (2002).
[29] J. Neugebauer, M.J. Louwerse, P. Belanzoni, T.A. Wesolowski and E.J. Baerends, J. Chem. Phys. **123**, 114101 (2005).
[30] M. Pavone, C. Benzi, F. De Angelis and V. Barone, Chem. Phys. Lett. **395**, 120–126 (2004).
[31] X. Chen, Z. Rinkevicius, Z. Cao, K. Ruud and H. Ågren, Phys. Chem. Chem. Phys. **13**, 696–707 (2011).
[32] X. Chen, Z. Rinkevicius, , K. Ruud and H. Ågren, J. Chem. Phys. **138**, 054310 (2013).
[33] A.Y. Adam, A. Yachmenev, S.N. Yurchenko and P. Jensen, J. Chem. Phys. **143**, 244306 (2015).
[34] S. Grimme, C. Bannwarth, S. Dohm, A. Hansen, J. Pisarek, P. Pracht, J. Seiber and F. Neese, Angew. Chem. Int. Ed. **56**, 14763–14769 (2017).
[35] J.L. Whitten, J. Chem. Phys. **58**, 4496–4501 (1973).
[36] B.I. Dunlap, J.W.D. Connolly and J.R. Sabin, J. Chem. Phys. **71** (8), 3396–3402 (1979).
[37] M. Feyereisen, G. Fitzgerald and A. Komornicki, Chem. Phys. Lett. **208**, 359–363 (1993).
[38] O. Vahtras, J. Almlöf and M. Feyereisen, Chem. Phys. Lett. **213**, 514–518 (1993).
[39] F. Weigend and M. Häser, Theor. Chem. Acc. **97**, 331–340 (1997).
[40] F. Weigend, M. Häser, H. Patzelt and R. Ahlrichs, Chem. Phys. Lett. **294**, 143–152 (1998).
[41] B. Doser, J. Zienau, L. Clin, D.S. Lambrecht and C. Ochsenfeld, Z. Phys. Chem. **224**, 397 (2010).
[42] S.A. Maurer, D.S. Lambrecht, D. Flaig and C. Ochsenfeld, J. Chem. Phys. **136**, 144107 (2012).
[43] S.A. Maurer, D.S. Lambrecht, J. Kussmann and C. Ochsenfeld, J. Chem. Phys. **138**, 014101 (2013).
[44] S. Schweizer, B. Doser and C. Ochsenfeld, J. Chem. Phys. **128** (15), 154101 (2008).
[45] H. Koch, A. Sánchez de Meras and T.B. Pedersen, J. Chem. Phys. **118**, 9481–9484 (2003).
[46] J. Boström, M. Pitoňák, F. Aquilante, P. Neogrády, T.B. Pedersen and R. Lindh, J. Chem. Theory Comput. **8**, 1921–1928 (2012).
[47] S. Vogler, M. Ludwig, M. Maurer and C. Ochsenfeld, J. Chem. Phys. **147**, 024101 (2017).
[48] S. Vogler, G. Savasci, M. Ludwig and C. Ochsenfeld, J. Chem. Theory Comput. **14**, 3014–3024 (2018).
[49] A. Warshel and M. Karplus, Chem. Phys. Lett. **32**, 11–17 (1975).
[50] C. Leforestier, J. Chem. Phys. **68**, 4406 (1978).
[51] L.D.M. Peters, J. Kussmann and C. Ochsenfeld, J. Chem. Theory Comput. **13**, 5479–5485 (2017).
[52] R. Sure and S. Grimme, J. Comput. Chem. **34**, 1672–1685 (2013).
[53] S. Grimme, J.G. Brandenburg, C. Bannwarth and A. Hansen, J. Chem. Phys. **143**, 054107 (2015).
[54] J. Tomasi, B. Mennuci and R. Cammi, Chem. Rev. **105**, 2999–3094 (2005).
[55] J. Gauss, in *Modern Methods and Algorithms of Quantum Chemistry* (, , 2000), pp. 541–592.
[56] J. Kussmann and C. Ochsenfeld, J. Chem. Phys. **138**, 134114 (2013).
[57] J. Kussmann and C. Ochsenfeld, J. Chem. Theory Comput. **11**, 918–922 (2015).

[58] J. Kussmann and C. Ochsenfeld, J. Chem. Theory Comput. **13**, 3153–3159 (2017).
[59] A. Takatsuka, S. Ten-no and W. Hackbusch, J. Chem. Phys. **129**, 044112 (2008).
[60] M. Beer and C. Ochsenfeld, J. Chem. Phys. **128**, 221102 (2008).
[61] G.D. Purvis III and R.J. Bartlett, J. Chem. Phys. **76**, 1910 (1982).
[62] J.F. Stanton, J. Gauss, L. Cheng, M.E. Harding, D.A. Matthews and P.G. Szalay, CFOUR, Coupled-Cluster techniques for Componentutational Chemistry, a quantum-chemical program package With contributions from A.A. Auer, R.J. Bartlett, U. Benedikt, C. Berger, D.E. Bernholdt, Y.J. Bomble, O. Christiansen, F. Engel, R. Faber, M. Heckert, O. Heun, M. Hilgenberg, C. Huber, T.-C. Jagau, D. Jonsson, J. Jusélius, T. Kirsch, K. Klein, W.J. Lauderdale, F. Lipparini, T. Metzroth, L.A. Mück, D.P. O'Neill, D.R. Price, E. Prochnow, C. Puzzarini, K. Ruud, F. Schiffmann, W. Schwalbach, C. Simmons, S. Stopkowicz, A. Tajti, J. Vázquez, F. Wang, J.D. Watts and the integral packages MOLECULE (J. Almlöf and P.R. Taylor), PROPS (P.R. Taylor), ABACUS (T. Helgaker, H.J. Aa. Jensen, P. Jørgensen, and J. Olsen), and ECP routines by A. V. Mitin and C. van Wüllen. For the current version, see http://www.cfour.de.
[63] F. Weigend and R. Ahlrichs, Phys. Chem. Chem. Phys. **7**, 3297–3305 (2005).
[64] F. Weigend, Phys. Chem. Chem. Phys. **4**, 4285–4291 (2002).
[65] H. Laqua, J. Kussmann and C. Ochsenfeld, J. Chem. Theory Comput. **14**, 3451 (2018).
[66] H. Laqua, J. Kussmann and C. Ochsenfeld, J. Chem. Phys. **149**, 204111 (2018).
[67] H. Laqua, T.H. Thompson, J. Kussmann and C. Ochsenfeld, J. Chem. Theory Comput. **16**, 1456–1468 (2020).
[68] L. Verlet, Phys. Rev. **159**, 98 (1967).
[69] W.C. Swope, H.C. Andersen, P.H. Berens and K.R. Wilson, J. Chem. Phys. **76**, 637 (1982).
[70] G. Bussi, D. Donadio and M. Parrinello, J. Chem. Phys. **126**, 014101 (2007).
[71] A.M.N. Niklasson, P. Steneteg, A. Odell, N. Bock, M. Challacombe, C.H. Tymczak, E. Holmström, G. Zheng and V. Weber, J. Chem. Phys. **130**, 214109 (2009).
[72] A.W. Lange and J.M. Herbert, J. Phys. Chem. Lett. **128**, 556–561 (2010).
[73] A.W. Lange and J.M. Herbert, J. Chem. Phys. **134**, 204110 (2011).
[74] A.W. Lange and J.M. Herbert, J. Chem. Phys. **134**, 117101 (2011).
[75] A. Klamt and G. Schüürmann, J. Chem. Soc., Perkin Trans. 2 pp. 799–805 (1993).
[76] A. Bondi, J. Phys. Chem. **68**, 441–451 (1964).
[77] J. Wang, W. Wang and P.A. Kollman, J. Mol. Graph. Model **25**, 247–260 (2005).
[78] D. Case, D. Cerutti, T. Cheatham, T. Darden, R. Duke, T. Giese, H. Gohlke, A. Goetz, D. Greene, N. Homeyer, S. Izadi, A. Kovalenko, T. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D. Mermelstein, K. Merz, G. Monard, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D. Roe, A. Roitberg, C. Sagui, C. Simmerling, W. Botello-Smith, J. Swails, R. Walker, J. Wang, R. Wolf, X. Wu, L. Xiao, D. York and P. Kollman, AMBER 2017 University of California, San Francisco, 2017.
[79] J. Wang, R.M. Wolf, J.W. Caldwell, P.A. Kollman and D.A. Case, J. Comput. Chem. **25** (9), 1157–1174 (2004).
[80] M.J.S. Dewar, E.G. Zoebisch, E.F. Healy and J.J.P. Stewart, J. Am. Chem. Soc. **107** (13), 3902–3909 (1985).
[81] H.J. Berendsen, J.R. Grigera and T.P. Straatsma, J. Phys. Chem. **91** (24), 6269–6271 (1987).
[82] J.C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R.D. Skeel, L. Kalé and K. Schulten, J. Comput. Chem. **26** (16), 1781–1802 (2005).
[83] J.P. Ryckaert, G. Ciccotti and H.J. Berendsen, J. Comput. Phys. **23** (3), 327–341 (1977).
[84] J. Boon, Philip, M.C.R. Symons, K. Ushida and T. Shida, J. Chem. Soc. Perkin Trans. 2 **0**, 1213 (1984).
[85] W.C. Danen and T.T. Kensler, J. Am. Chem. Soc. **92**, 5235 (1970).
[86] T. Bally and T. Borden, in , edited by Kenny B. Lipkowitz and Donald B. Boyd (, , 1999), *Reviews in Computational Chemistry*, Vol. 13, Chap. Calculations on Open Shell Molecules: A Beginners Guide.

[87] L. Hermosilla, P. Calle, J.M. Garcia de la Vega and C. Sieiro, J. Phys. Chem. A **109**, 1114–1124 (2005).
[88] L. Hermosilla, P. Calle, J.M. Garcia de la Vega and C. Sieiro, J. Phys. Chem. A **110**, 13600–13608 (2006).
[89] S. Kozuch, D. Gruzman and J.M.L. Martin, J. Phys. Chem. C **114**, 20801–20808 (2010).
[90] K. Toriyama, K. Nunome and M. Iwasaki, J. Chem. Phys. **77**, 5891 (1982).
[91] M.C.R. Symons, Chem. Soc. Rev. **13**, 393–439 (1984).
[92] U. Jacovella, C.J. Stein, M. Grütter, L. Freitag, C. Lauzin, M. Reiher and F. Merkt, Phys. Chem. Chem. Phys. **20**, 1072–1081 (2018).
[93] T. Shida and T. Kato, Chem. Phys. Lett. **68**, 106–110 (1979).
[94] J.R. Morton, K.F. Preston and S.J. Strach, J. Phys. Chem. **83**, 533–536 (1979).
[95] R.W. Fessenden, J. Phys. Chem. **71**, 74–83 (1967).
[96] R.W. Fessenden and R.H. Schuler, J. Chem. Phys. **43**, 2704 (1965).
[97] S.A. Perera, L.M. Salemi and R.J. Bartlett, J. Chem. Phys. **106**, 4061 (1997).
[98] T.N. Lan, Y. Kurashige and T. Yanai, J. Chem. Theory Comput. **10**, 1953–1967 (2014).
[99] H.J. McManus, R.W. Fessenden and D.M. Chipman, J. Phys. Chem. **92**, 3781–3784 (1988).
[100] L.B. Knight, M. Winiski, P. Miller, C.A. Arrington and D. Feller, J. Chem. Phys. **91**, 4468 (1989).
[101] E. Kwan and R.Y. Liu, J. Chem. Theory Comput. **11** (11), 5083–5089 (2015).
[102] V. Barone, P. Cimino and E. Stendardo, J. Chem. Theory Comput. **4**, 751–764 (2008).
[103] R.W. Kreilick, J. Chem. Phys. **46**, 4260 (1967).

Supporting Information: Important Components for Accurate Hyperfine Coupling Constants: Electron Correlation, Dynamic Contributions, and Solvation Effects

Sigurd Vogler,[†] Johannes C. B. Dietschreit,[†] Laurens D. M. Peters, Christian Ochsenfeld[*]

Chair of Theoretical Chemistry, Department of Chemistry,

University of Munich (LMU), Butenandtstr. 7, D-81377 München, Germany

[†]Contributed equally to this work

[*]E-Mail: christian.ochsenfeld@uni-muenchen.de

Contents

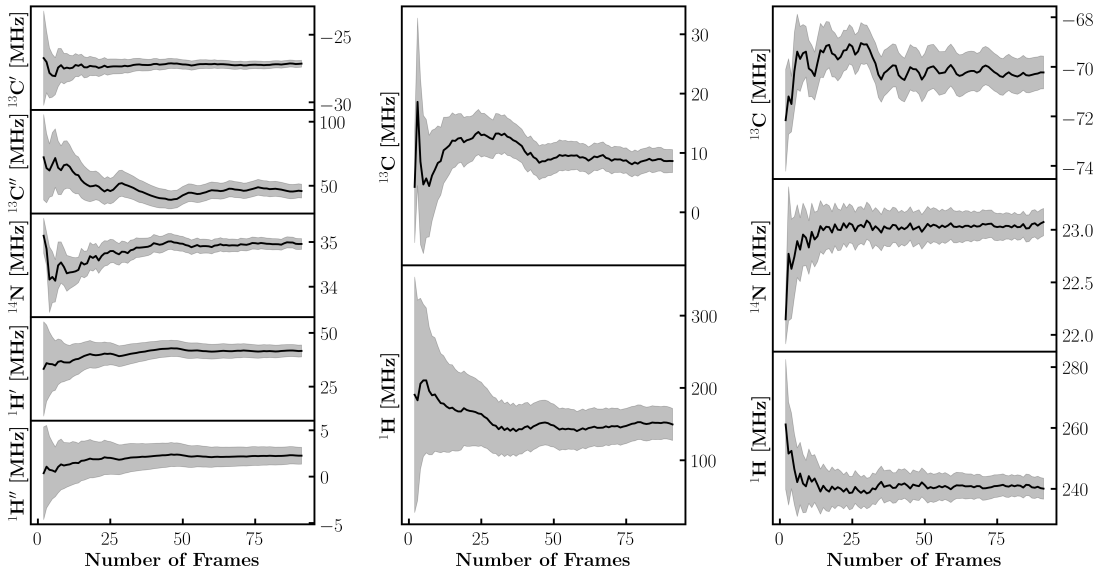
1 Image Processing

All plots were generated using the python-package *matplotlib*[1]. The chemical structures were drawn with *ChemDraw*.

2 Data and Materials Availability

All inputs and trajectories are available upon request. The program package *FermiONs++* [2–4] is not yet available for public usage.

3 Convergence with number of frames



(a) Radical **2** (b) Radical **3** (c) Radical **11**

SFigure 1: Convergence of the mean isotropic HFCCs at the B3LYP level with the number of frames taken from the PBEH3c-AIMD simulation. The frames are sorted chronologically. The standard error of the mean is indicated in grey.

4 Data Analysis

We have performed a small statistical analysis in order to show the independence of the different contributions. The only set of data that is large enough to allow for such a treatment is the one in Table 3. We have computed the standardized covariance ($\widetilde{\mathrm{COV}}(X,Y) = \frac{\mathrm{COV}(X,Y)}{\sigma_X \sigma_Y}$) for those ^{13}C nuclei that do not show a Jahn-Teller distortion (15 values). We looked at the following contributions: Δ_{dyn}(B3LYP), Δ_{dyn}(B2PLYP), Δ_{corr}(opt), and Δ_{corr}(dyn). The first two are the difference between the columns 1 and

2 of the respective method; the latter two, are the difference between the column 1 of each method and column 2 of each method, respectively.

X	Y	$\widetilde{\mathrm{COV}}(X,Y)$
Δ_{dyn}(B3LYP)	Δ_{dyn}(B2PLYP)	0.997
Δ_{corr}(opt)	Δ_{corr}(dyn)	0.993
Δ_{dyn}(B3LYP)	Δ_{corr}(opt)	0.275
Δ_{dyn}(B3LYP)	Δ_{corr}(dyn)	0.284
Δ_{dyn}(B2PLYP)	Δ_{corr}(opt)	0.258
Δ_{dyn}(B2PLYP)	Δ_{corr}(dyn)	0.275

The results of this analysis is twofold: Firstly, the effects of the conformational ensemble (Δ_{dyn}) between B3LYP and B2PLYP calculations as well as the e-correlation effect (Δ_{corr}) between optimized structure and dynamics are highly correlated. This means that calculating one of those is as good as the other. Secondly, different types of contributions show little covariance. Therefore, the level, at which the conformational ensemble and the correlation effect are computed, are largely independent of one another.

These features justify to choose a set of calculations with as little computational effort as possible.

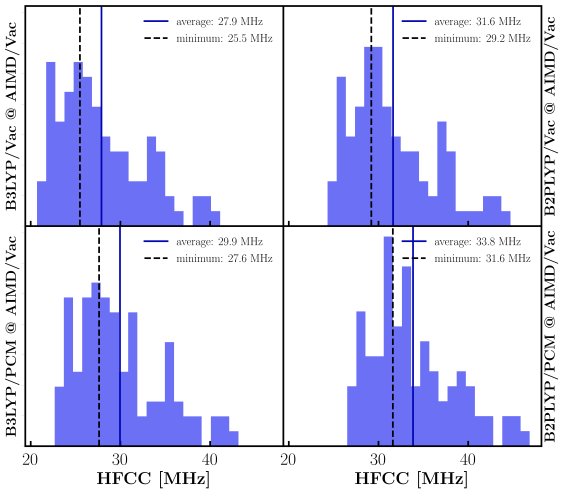
5 Solvation

STable 1: This table contains the values from the optimized structures only. Notably, the size of the implicit (Δ_{solv}(structure)) and explicit (Δ_{solv}(mol. prop.)) solvation contributions are independent of the method used for the HFCC calculation. However, they depend on the level of theory used for structure generation. The same is true for the difference between hybrid and double-hybrid functional (here noted with Δ_{corr}). The largest influence comes from Δ_{method}.

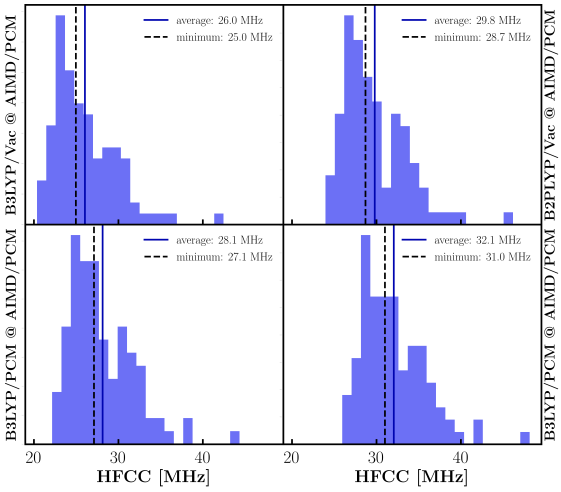
Struct. Gen. / e^--density	HF3c		PBEH3c		Δ_{solv}(structure)		Δ_{method}(opt)	
	@vac	@PCM	@vac	@PCM	HF3c	PBEH3c	vac	PCM
B3LYP@vac	25.5	25.0	33.7	32.7	-0.5	-0.9	+8.2	+7.8
B3LYP@PCM	27.6	27.1	36.6	35.7	-0.5	-0.9	+9.0	+8.6
B2PLYP@vac	29.2	28.7	38.0	37.1	-0.5	-0.9	+8.8	+8.4
B2PLYP@PCM	31.6	31.0	40.9	40.0	-0.5	-0.9	+9.3	+9.0
Δ_{solv}(mol. prop.)								
B3LYP	+2.1	+2.1	+2.9	+3.0				
B2PLYP	+2.3	+2.3	+2.8	+2.9				
Δ_{corr}								
@vac	+3.7	+3.7	+4.4	+4.4				
@PCM	+3.9	+3.9	+4.3	+4.3				

STable 2: All results presented here in the right hand top corner are averages over AIMD simulations. All dynamics were performed at the HF3c level of theory. As in STable 1, we separate off Δ_{solv}(structure) and Δ_{solv}(mol. prop.), and Δ_{corr}. As we used only HF3c, there is no Δ_{method}. By subtracting the corresponding values from the geometry optimizations, we could calculate Δ_{dyn} for all combinations except those involving QM/MM, as the global minimum is not defined.

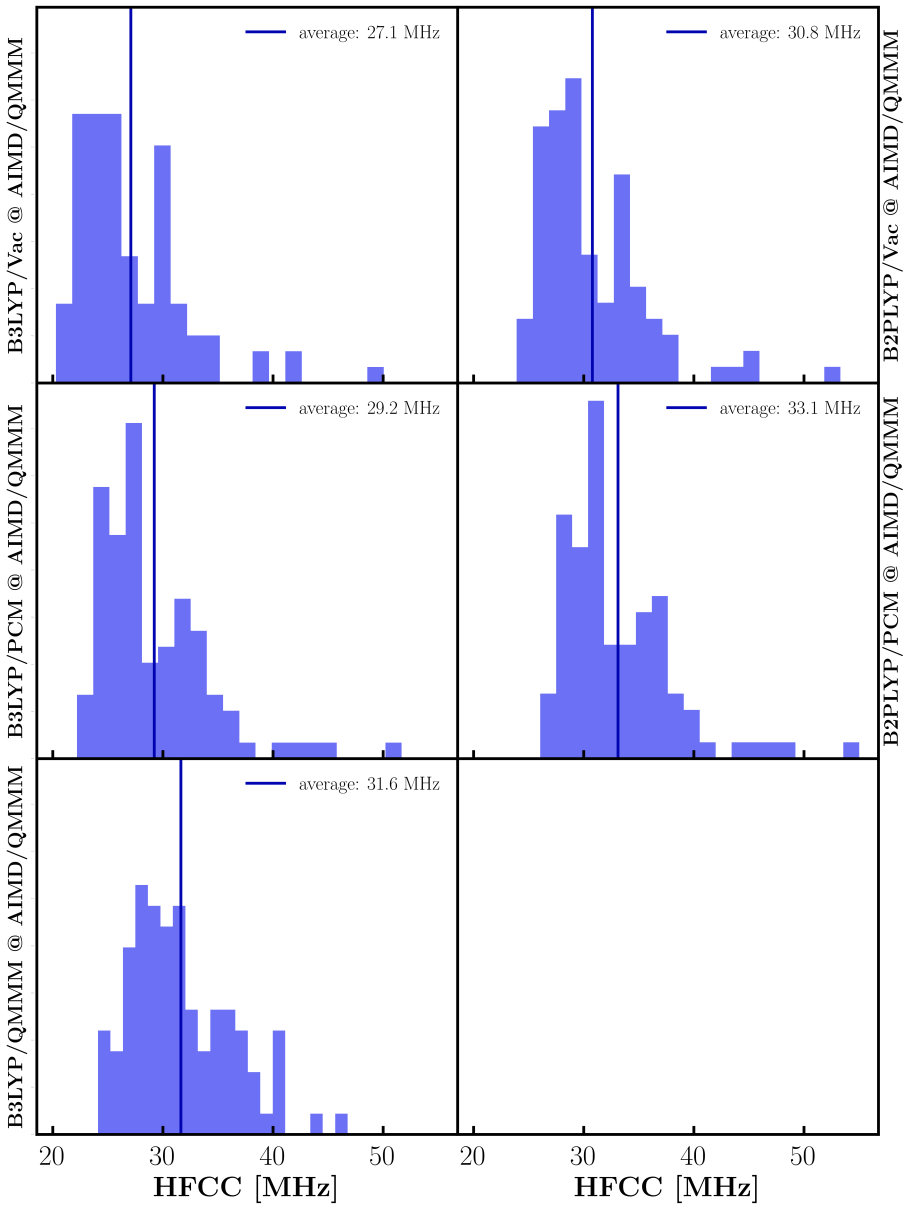
Struct. Gen. / e^--density	HF3c			Δ_{solv}(structure)		Δ_{dyn}	
	@vac	@PCM	@QM/MM	PCM - vac	QM/MM - vac	vac	PCM
B3LYP@vac	27.9	26.0	27.1	-1.9	-0.8	+2.4	+1.0
B3LYP@PCM	29.9	28.1	29.2	-1.9	-0.8	+2.4	+1.0
B3LYP@QM/MM	–	–	31.6	–	–	–	–
B2PLYP@vac	31.6	29.8	30.8	-1.8	-0.8	+2.4	+1.1
B2PLYP@PCM	33.8	32.1	33.1	-1.7	-0.7	+2.2	+1.1
Δ_{solv}(mol. prop.)							
B3LYP	+2.0	+2.1	+2.1/4.5				
B2PLYP	+2.2	+2.3	+2.3/–				
Δ_{corr}							
@vac	+3.7	+3.8	+3.7				
@PCM	+3.8	+4.0	+3.9				



SFigure 2: Distribution of the HFCCs calculated for one hundred frames from an AIMD in vacuum. Indicated are the average value over the one hundred frames (dark blue) and the value calculated for the optimized minimum energy geometry (black). The left column contains all HFCCs computed with B3LYP, the right column all results from B2PLYP.



SFigure 3: Distribution of the HFCCs calculated for one hundred frames from an AIMD in PCM. Indicated are the average value over the one hundred frames (dark blue) and the value calculated for the optimized minimum energy geometry (black). The left column contains all HFCCs computed with B3LYP, the right column all results from B2PLYP.



SFigure 4: Distribution of the HFCCs calculated for one hundred frames from a HF3c QM/MM-MD which included 66 water molecules in the QM-sphere. Indicated is the average value over the one hundred frames (dark blue). The left column contains all HFCCs computed with B3LYP, the right column all results from B2PLYP. The top two rows only use the geometries of hydroxylated TEMPO, only the lowest includes all QM/MM atoms.

References

[1] J. D. Hunter, *Computing In Science & Engineering* **2007**, *9*, 90–95.

[2] J. Kussmann, C. Ochsenfeld, *J. Chem. Phys.* **2013**, *138*, 134114.

[3] J. Kussmann, C. Ochsenfeld, *J. Chem. Theory Comput.* **2015**, *11*, 918–922.

[4] J. Kussmann, C. Ochsenfeld, *J. Chem. Theory Comput.* **2017**, *13*, 3153–3159.

Chapter 4

Conclusion and Outlook

The influence of different molecular configurations is the recurrent theme in the six studies presented in this work. Without MD simulations none of these projects would have been possible or would have reached the accuracy they did. The exploration of configuration space and its inclusion in the description of a system, whether it concerns energetics or observables, is crucial. Without sampling at finite temperatures, effects like the vibrations of bonds or exchange between local minimum energy configurations cannot be properly described.

The first two publications have introduced and showcased the method DSI, which offers not only a way to analyze, but also to localize free energy differences. The application of DSI is not limited to standard free energy problems such as the binding of an inhibitor to a protein. As was shown by analyzing glucose, free energy hot-spots can be used to identify all atoms involved in the anomeric effect, an intramolecular stereoelectronic effect. Hence, DSI can be used to analyze molecular systems in small detail down to single atoms.

Building on the presented results, DSI is now ready to be used in future studies. The VDoS can be used to rate the strength of hydrogen bonds, which is especially important in biomolecular systems, as hydrogen bonds often play a crucial role, e.g., in protein residue interactions. Future work should also aim to improve the numerical stability of the algorithm. The method's dependence on slow modes has to be decreased, such that relatively short trajectories are sufficient. The number of data points along a trajectory has to be reduced as saving velocities every femtosecond accumulates large amounts of data quickly. In this context, the influence of SHAKE [149] and RATTLE [150] have to be analyzed. These two methods enable freezing vibrations within a system. Most likely, RATTLE can be used in conjunction with DSI, as it not only modifies the relative position of atoms but also corrects their velocities, which will create less or no artefacts in the VDoS compared to SHAKE. The method from Ref. [74], which foregoes computing the VDoS and directly calculates the free energy from the vibrational autocorrelation

function, avoids using the weighting function, which changes rapidly for frequencies near zero, and thus may most likely increase the numerical accuracy of vibrational free energies.

The machine learning study on reaction barriers demonstrated how strongly minimal energy paths depend on the starting configurations of the system. The potential energy surface of biomolecular systems is so rough that a single or a few minimal energy paths are meaningless when trying to estimate an effective reaction barrier height from them. Our results agree with those of Ryde [29] that millions of frames are needed, many more than the 150 frames we could afford, when one aims to estimate the free energy barrier as an exponentially weighted average from those minimal energy paths. However, machine learning helps to reduce the number of frames necessary and to locate those regions within configuration space which allow a reaction to occur.

After machine learning, we went one step further and actually sampled the system with QM/MM-MD to calculate the free energy profile along the first reaction step of Sirt5. However, employing QM/MM makes an MD time step very costly. Usually, in the context of QM/MM one aims for QM size converged results, meaning using a QM region so large that the results do not change when increasing the QM region further. Working with QM size converged regions is not possible yet when performing QM/MM-MDs of biomolecular systems, as it would include all the thousands atoms within a sphere of several Ångstroms around the reaction centre. Thus, we chose a QM region that included a few atoms more than the bare minimum of necessary atoms. Another option to reduce cost is to choose a lower level approximation when solving the Schrödinger equation. Hence, we chose HF-3c/minix as a low level but still *ab initio* approach. However, HF-3c seems to be unable to describe stretched bonds accurately, which always occur during chemical reactions. This leads to a systematic overestimation of the reaction barrier. We proved this as we compared the value of the HF-3c energy barrier to higher level approximations with larger basis sets in the machine learning study. Therefore, it was planned to use reweighting the HF-3c configurational ensemble to a more accurate DFT method, but this appears to be less straightforward than expected. The potential energy surface away from low energy configurations is so different that the reweighting entropy loss is significant. 20 ps long MD simulations appear to be too short to perform any kind of reweighting from HF-3c to DFT.

Future work will have to focus on combining the speed of seminumerical-DFT or other fast methods developed within FermiONs++ [151–153] with MD. It will then be possible to perform the sampling needed for free energies on a higher level of theory than HF-3c, given that reactions which include the breaking or formation of bonds have to be described.

The study on HFCCs has shown how crucial sampling in the form of MD is when trying to compute accurate observables, i.e., computational protocols which focus solely on minimum energy configurations should be avoided if possible. The MD sampling can be performed on many different levels of theory ranging from MD over HF-3x/minix to DFT. The level of theory used for the MD has a non-negligible influence on the value of the observable, which luckily can be corrected without having to perform the entire MD on a higher level of theory. Such corrections that reduce computational time without reducing the accuracy of the theoretical description are important when tackling systems beyond a few atoms and comparison to experiment is sought.

Additional studies will have to prove that the findings for HFCCs are indeed transferable to other observables, such as for example, to NMR chemical shifts, and whether even cheaper sampling such as MM-MD can be corrected in the same manner as was possible for HF-3c to PBEh-3c in the case of small radicals. As this is very likely the case, large (bio-)molecular systems can be tackled on a regular basis, as we have already shown in the study on ^{19}F NMR shifts. As a test case, the Tpx-CFT system could be revisited and it has to be tested by how much the accuracy, which was already high, can be improved further when systematic corrections are used.

To conclude, this thesis has laid the groundwork for the analysis tool DSI and has successfully improved observable calculations by applying statistical thermodynamics, opening the doors to exciting new projects.

Chapter 5

Bibliography

[1] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller, *J. Chem. Phys.* **1953**, *21*, 1087–1092.

[2] E. Fermi, J. Pasta, S. Ulam, *Los Alamos report LA-1940* **1955**.

[3] B. J. Alder, T. E. Wainwright, *J. Chem. Phys.* **1959**, *31*, 459.

[4] A. Rahman, *Phys. Rev. A* **1964**, *136*, 405–411.

[5] L. Verlet, *Phys. Rev.* **1967**, *159*, 98–103.

[6] K. Lindorff-Larsen, S. Piana, R. O. Dror, D. E. Shaw, *Science* **2011**, *334*, 517–520.

[7] A. Warshel, M. Levitt, *J. Mol. Biol.* **1976**, *103*, 227–249.

[8] The Noble Prize in Chemistry 2013, Press Release, **2013**.

[9] M. Born, R. Oppenheimer, *Ann. Phys.* **1927**, *389*, 457–484.

[10] E. Schrödinger, *Phys. Rev.* **1926**, *28*, 1049–1070.

[11] W. K. Hastings, *Biometrika* **1970**, *57*, 97–109.

[12] G. M. Torrie, J. P. Valleau, *J. Comp. Phys.* **1977**, *23*, 187–199.

[13] P. H. Berens, D. H. J. Mackay, G. M. White, K. R. Wilson, *J. Chem. Phys.* **1983**, *79*, 2375–2389.

[14] R. W. Zwanzig, *J. Chem. Phys.* **1954**, *22*, 1420–1426.

[15] C. H. Bennett, *J. Comp. Phys.* **1976**, *22*, 245–268.

[16] P. Filippakopoulos, S. Picaud, M. Mangos, T. Keates, J.-P. Lambert, D. Barsyte-Lovejoy, I. Felletar, R. Volkmer, S. Müller, T. Pawson, A.-C. Gingras, C. H. Arrowsmith, S. Knapp, *Cell* **2012**, *149*, 214–231.

[17] A. C. Runcie, M. Zengerle, K.-H. Chan, A. Testa, L. van Beurden, M. G. J. Baud, O. Epemolu, L. C. J. Ellis, K. D. Read, V. Coulthard, A. Brien, A. Ciulli, *Chem. Sci.* **2018**, *9*, 2452–2468.

[18] E. Juaristi, G. Cuevas, *Tetrahedron* **1992**, *48*, 5019–5087.

[19] Y. Mo, *Nat. Chem.* **2010**, *2*, 666–671.

[20] E. J. Cocinero, P. Carcabal, T. D. Vaden, J. P. Simons, B. G. Davis, *Nature* **2011**, *469*, 76–80.

[21] C. Wang, F. Ying, W. Wu, Y. Mo, *J. Org. Chem.* **2014**, *79*, 1571–1581.

[22] C. M. Filloux, *Angew. Chemie - Int. Ed.* **2015**, *54*, 8880–8894.

[23] J. Schemies, U. Uciechowska, W. Sippl, M. Jung, *Med. Res. Rev.* **2009**, *30*, 861–889.

[24] P. Parihar, I. Solanki, M. L. Mansuri, M. S. Parihar, *Exp. Gerontol.* **2015**, *61*, 130–141.

[25] R. Sure, S. Grimme, *J. Comput. Chem.* **2013**, *34*, 1672-1685.

[26] S. Hur, T. C. Bruice, *Proc. Natl. Acad. Sci.* **2002**, *99*, 1176–1181.

[27] S. Hur, T. C. Bruice, *Proc. Natl. Acad. Sci.* **2003**, *100*, 12015–12020.

[28] H. Guo, Q. Cui, W. N. Lipscomb, M. Karplus, *Angew. Chemie - Int. Ed.* **2003**, *42*, 1508–1511.

[29] U. Ryde, *J. Chem. Theory Comput.* **2017**, *13*, 5745–5752.

[30] M. R. Shirts, J. D. Chodera, *J. Chem. Phys.* **2008**, *129*, 1–10.

[31] P. Hu, S. Wang, Y. Zhang, *J. Am. Chem. Soc.* **2008**, *130*, 16721–16728.

[32] Z. Liang, T. Shi, S. Ouyang, H. Li, K. Yu, W. Zhu, C. Luo, H. Jiang, *J. Phys. Chem. B* **2010**, *114*, 11927–11933.

[33] S. Vosko, L. Wilk, M. Nusair, *Can. J. Phys.* **1980**, *58*, 1200–1211.

[34] C. Lee, W. Yang, R. G. Parr, *Phys. Rev. B* **1988**, *37*, 785–789.

[35] A. D. Becke, *J. Chem. Phys.* **1993**, *98*, 5648–5652.

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

[37] F. Weigend, R. Ahlrichs, *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–305.

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

[39] M. A. Comini, R. L. Krauth-Siegel, L. Floh, *Biochem. J.* **2007**, *402*, 43–49.

[40] A. Wagner, E. Diehl, R. L. Krauth-Siegel, U. A. Hellmich, *Biomol. NMR Assignments* **2017**, *11*, 193–196.

[41] A. Wagner, T. A. Le, M. Brennich, P. Klein, N. Bader, E. Diehl, D. Paszek, A. K. Weickhmann, N. Dirdjaja, R. L. Krauth-Siegel, B. Engels, T. Opatz, H. Schindelin, U. A. Hellmich, *Angew. Chemie - Int. Ed.* **2019**, *58*, 3640–3644.

[42] S. Vogler, M. Ludwig, M. Maurer, C. Ochsenfeld, *J. Chem. Phys.* **2017**, *147*, 024101.

[43] S. Vogler, G. Savasci, M. Ludwig, C. Ochsenfeld, *J. Chem. Theory Comput.* **2018**, *14*, 3014–3024.

[44] T. L. Hill, *An introduction to statistical thermodynamics*, Addison-Wesley Pub. Co, Reading, Mass., **1960**.

[45] D. A. McQuarrie, *Statistical mechanics*, Harper & Row, New York, **1976**.

[46] D. Chandler, *Introduction to modern statistical mechanics*, Oxford University Press, New York, **1987**.

[47] C. Jarzynski, *Phys. Rev. Lett.* **1997**, *78*, 2690–2693.

[48] C. Jarzynski, *Phys. Rev. E* **1997**, *56*, 5018–5035.

[49] G. Crooks, *J. Stat. Phys.* **1998**, *90*, 1481–1487.

[50] G. Crooks, *Phys. Rev. E* **1999**, *60*, 2721–2726.

[51] F. Feil, S. Naumov, J. Michaelis, R. Valiullin, D. Enke, J. Kärger, C. Bräuchle, *Angew. Chemie - Int. Ed.* **Jan. 2012**, *124*, 1152–1155.

[52] E. M. Pearson, T. Halicioglu, W. A. Tiller, *Phys. Rev. A* **1988**, *32*, 3030–3039.

[53] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, F. Noe, *J. Chem. Phys.* **2011**, *134*, 174105.

[54] J.-H. Prinz, B. Keller, F. Noé, *Phys. Chem. Chem. Phys.* **2011**, *13*, 16912–16927.

[55] G. R. Bowman, V. S. Pande, F. Noe, *An introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*, Springer, **2014**.

[56] F. Vitalini, A. S. J. S. Mey, F. Noé, B. G. Keller, *J. Chem. Phys.* **2015**, *142*, 084101.

[57] F. Vitalini, F. Noé, B. G. Keller, *J. Chem. Theory Comput.* **2015**, *11*, 3992–4004.

[58] S. Lloyd, *IEEE Trans. Inf. Theory* **1982**, *28*, 129–137.

[59] M. Ester, H.-P. Kriegel, J. Sander, X. Xu in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, AAAI Press, Portland, Oregon, **1996**, pp. 226–231.

[60] O. Lemke, B. G. Keller, *Algorithms* **2018**, *11*, 19–21.

[61] *Free Energy Calculations*, (Eds.: C. Chipot, A. Pohorille), Springer-Verlag, Berlin Heidelberg, **2007**.

[62] S.-T. Lin, M. Blanco, W. A. Goddard, *J. Chem. Phys.* **2003**, *119*, 11792–11805.

[63] S.-T. Lin, P. K. Maiti, W. A. Goddard, *J. Phys. Chem. B* **2010**, *114*, 8191–8198.

[64] J. M. Dickey, A. Paskin, *Phys. Rev.* **1969**, *188*, 1407–1418.

[65] M. Hesse, H. Meier, B. Zeeh, *Spektroskopische Methoden in der Organischen Chemie*, 7th ed., Thieme, **2005**.

[66] B. Brooks, M. Karplus, *Proc. Natl. Acad. Sci. USA* **1983**, *80*, 6571–6575.

[67] B. Tidor, M. Karplus, *J. Mol. Biol.* **1994**, *238*, 405–414.

[68] M. Karplus, J. N. Kushick, *Macromolecules* **1981**, *14*, 325–332.

[69] R. M. Levy, A. R. Srinivasan, W. K. Olson, *Biopolymers* **1984**, *23*, 1099–1112.

[70] P. E. Smith, W. F. van Gunsteren, *J. Phys. Chem.* **1994**, *98*, 13735–13740.

[71] S. Boresch, G. Archontis, M. Karplus, *Proteins* **1994**, *20*, 25–33.

[72] S. Boresch, M. Karplus, *J. Mol. Biol.* **1995**, *254*, 801–807.

[73] B. W. J. Irwin, D. J. Huggins, *J. Chem. Theory Comput.* **2018**, 3218–3227.

[74] D. Berta, D. Ferenc, I. Bakó, Á. Madarász, *J. Chem. Theory Comput.* **2020**, *16*, 3316–3334.

[75] P. Kollman, *Chem. Rev.* **1993**, *93*, 2395–2417.

[76] N. Lu, D. A. Kofke, T. B. Woolf, *J. Phys. Chem. B* **2003**, *107*, 5598–5611.

[77] T. A. Pascal, S.-T. Lin, W. A. Goddard III, *Phys. Chem. Chem. Phys.* **2011**, *13*, 169–181.

[78] R. A. Persson, V. Pattni, A. Singh, S. M. Kast, M. Heyden, *J. Chem. Theory Comput.* **2017**, *13*, 4467–4481.

[79] V. Pattni, T. Vasilevskaya, W. Thiel, M. Heyden, *J. Phys. Chem. B* **2017**, *121*, 7431–7442.

[80] S. Belsare, V. Pattni, M. Heyden, T. Head-Gordon, *J. Phys. Chem. B* **2018**, *122*, 5300–5307.

[81] M. Born, *Z. Phys.* **1920**, *1*, 45–48.

[82] J. G. Kirkwood, *J. Chem. Phys.* **1935**, *3*, 300–313.

[83] W. L. Jorgensen, C. Ravimohan, *J. Chem. Phys.* **1985**, *83*, 3050–3054.

[84] T. Simonson, G. Archontis, M. Karplus, *Acc. Chem. Res.* **2002**, *35*, 430–437.

[85] S. Bruckner, S. Boresch, *J. Comput. Chem.* **2010**, *32*, 1320–1333.

[86] In *The Concise Encyclopedia of Statistics*, (Ed.: Y. Dodge), Springer New York, New York, NY, **2008**, pp. 66–68.

[87] A. Amadei, M. E. F. Apol, H. J. C. Berendsen, *J. Chem. Phys.* **1996**, *104*, 1560–1574.

[88] G. Hummer, L. Pratt, A. E. Garcia, *J. Am. Chem. Soc.* **1997**, *119*, 8523–8527.

[89] S. T. Bramwell, K. Christensen, J. Y. Fortin, P. C. W. Holdsworth, H. J. Jensen, S. Lise, J. M. López, M. Nicodemi, J. F. Pinton, M. Sellitto, *Phys. Rev. Lett.* **2000**, *84*, 3744–3747.

[90] H. Nanda, N. Lu, D. A. Kofke, *J. Chem. Phys.* **2005**, *122*, 134110:1–8.

[91] J. P. Valleau, D. N. Card, *J. Chem. Phys.* **1972**, *57*, 5457–5462.

[92] C. Y. Lee, H. L. Scott, *J. Chem. Phys.* **1980**, *73*, 4591–4596.

[93] M. R. Shirts, E. Bair, G. Hooker, V. S. Pande, *Phys. Rev. Lett* **2003**, *91*, 140601.

[94] N. D. Lu, J. K. Singh, D. A. Kofke, *J. Chem. Phys.* **2003**, *118*, 2977–2984.

[95] M. R. Shirts, V. S. Pande, *J. Chem. Phys.* **2005**, *122*, 144107.

[96] R. Srinivasan, *Importance Sampling - Applications in Communications and Detection*, 1st ed., Springer-Verlag Berlin Heidelberg, **2002**.

[97] T. Huber, A. E. Torda, W. E. van Gunsteren, *J. Comput. Aided Mol. Des.* **1994**, *8*, 8–695.

[98] A. Laio, M. Parrinello, *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 12562–12566.

[99] J. Kästner, H. M. Senn, S. Thiel, N. Otte, W. Thiel, J. Ka, *J. Chem. Theory Comput.* **2006**, *2*, 452–461.

[100] J. Kästner, W. Thiel, *J. Chem. Phys.* **2006**, *124*.

[101] J. Kästner, *Wiley Interdiscip. Rev.-Comput. Mol. Sci.* **2011**, *1*, 932–942.

[102] G. N. Patey, J. P. Valleau, *J. Chem. Phys.* **1975**, *63*, 2334–2339.

[103] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, P. A. Kollman, *J. Comput. Chem.* **1992**, *13*, 1011–1021.

[104] J. Kästner, W. Thiel, *J. Chem. Phys.* **2005**, *123*.

[105] E. Rosta, G. Hummer, *J. Chem. Theory Comput.* **2015**, *11*, 276–285.

[106] L. S. Stelzl, A. Kells, E. Rosta, G. Hummer, *J. Chem. Theory Comput.* **2017**, *13*, 6328–6342.

[107] H. Wu, F. Paul, C. Wehmeyer, F. Noé, *Proc. Natl. Acad. Sci. USA* **2016**, *113*, E3221–E3230.

[108] B. Roux, *Comput. Phys. Commun.* **1995**, *91*, 275–282.

[109] Z. Tan, *J. Am. Stat. Assoc.* **2004**, *99*, 1027–1036.

[110] L. Kantorovich, *Uspekhi Mat. Nauk* **1948**, *3*, 89–185.

[111] C. G. Broyden, *IMA J. Appl. Math.* **1970**, *6*, 76–90.

[112] R. Fletcher, *Comput. J.* **1970**, *13*, 317–322.

[113] D. Goldfarb, *Math. Comput.* **1970**, *24*, 23–26.

[114] D. F. Shanno, *Math. Comput.* **1970**, *24*, 647–656.

[115] F. Zhu, G. Hummer, *J. Comput. Chem.* **2012**, *33*, 453–465.

[116] P. Pulay, *Chem. Phys. Lett.* **1980**, *73*, 393–398.

[117] C. Zhang, C.-L. Lai, B. M. Pettitt, *Mol. Simul.* **2016**, *42*, 1079–1089.

[118] P. Li, X. Jia, X. Pan, Y. Shao, Y. Mei, *J. Chem. Theory Comput.* **2018**, *14*, 5583–5596.

[119] M. Campbell, A. Hoane, F.-h. Hsu, *Artif. Intell.* **2002**, *134*, 57–83.

[120] M. H. S. Segler, M. P. Waller, *Chem. Eur. J.* **2017**, *23*, 5966–5971.

[121] J. C. Cole, C. R. Groom, M. G. Read, I. Giangreco, P. McCabe, A. M. Reilly, G. P. Shields, *Acta Crystallogr. B* **2016**, *72*, 530–541.

[122] O. A. von Lilienfeld, *Angew. Chem. - Int. Ed.* **2018**, *57*, 4164–4169.

[123] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, *Nature* **2018**, *559*, 547–555.

[124] K. Pearson, *Philos. Mag.* **1901**, *2*, 559–572.

[125] M. Rupp, A. Tkatchenko, K. R. Müller, O. A. von Lilienfeld, *Phys. Rev. Lett.* **Jan. 2012**, *108*, 058301.

[126] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko, K. R. Müller, *J. Chem. Theory Comput.* **Aug. 2013**, *9*, 3404–3419.

[127] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K. R. Müller, A. Tkatchenko, *J. Phys. Chem. Lett.* **June 2015**, *6*, 2326–2331.

[128] D. Weininger, *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

[129] W. Hamilton, *Discussions on philosophy and literature, education and university reform*, Brown, Green and Longmans, **1852**.

[130] H. Zou, T. Hastie, *J. Royal Stat. Soc. B* **2005**, *67*, 301–320.

[131] M. Rupp, *Int. J. Quantum Chem.* **2015**, *115*, 1058–1073.

[132] F. Galton, *J. R. Anthropol. Inst.* **1886**, *15*, 246–263.

[133] R. Tibshirani, *J. Royal Stat. Soc. B* **1994**, *58*, 267–288.

[134] A. E. Hoerl, R. W. Kennard, *Technometrics* **1970**, *12*, 55–67.

[135] P. Pyykkö in *Calculation of NMR and EPR Parameters*, John Wiley & Sons, Ltd, **2004**, Chapter 2, pp. 7–19.

[136] F. Neese, M. L. Munzarová in *Calculation of NMR and EPR Parameters*, John Wiley & Sons, Ltd, **2004**, Chapter 3, pp. 21–32.

[137] J. Gauss, J. F. Stanton in *Calculation of NMR and EPR Parameters*, John Wiley & Sons, Ltd, **2004**, Chapter 8, pp. 123–139.

[138] J. Gauss in *Modern Methods and Algorithms of Quantum Chemistry*, John von Neumann Institute for Computing, Jülich, **2000**, pp. 541–592.

[139] T. Helgaker, M. Jaszuński, K. Ruud, *Chem. Rev.* **1999**, *99*, 293–352.

[140] K. Ruud, P.-O. Åstrand, P. R. Taylor, *J. Chem. Phys.* **2000**, *112*, 2669–2683.

[141] A. Y. Adam, A. Yachmenev, S. N. Yurchenko, P. Jensen, *J. Chem. Phys.* **2015**, *143*, 244306.

[142] S. Grimme, C. Bannwarth, S. Dohm, A. Hansen, J. Pisarek, P. Pracht, J. Seibert, F. Neese, *Angew. Chemie - Int. Ed.* **2017**, *56*, 14763–14769.

[143] E. Stendardo, A. Pedone, P. Cimino, M. C. Menziani, O. Crescenzi, V. Barone, *Phys. Chem. Chem. Phys.* **2010**, *12*, 11697–11709.

[144] M. Dračínský, H. M. Möller, T. E. Exner, *J. Chem. Theory Comput.* **2013**, *9*, 3806–3815.

[145] E. E. Kwan, R. Y. Liu, *J. Chem. Theory Comput.* **2015**, *11*, 5083–5089.

[146] V. Barone, P. Cimino, A. Pedone, *Magn. Reson. Chem.* **2010**, *48*, S11–S22.

[147] T. Zhu, J. Z. H. Zhang, X. He, *J. Chem. Theory Comput.* **2013**, *9*, 2104–2114.

[148] T. Giovannini, P. Lafiosca, B. Chandramouli, V. Barone, C. Capelli, *J. Chem. Phys.* **2019**, *150*, 124102.

[149] J. P. Ryckaert, G. Ciccotti, H. J. Berendsen, *J. Comput. Phys.* **1977**, *23*, 327–341.

[150] H. C. Andersen, *J. Comput. Phys.* **1983**, *52*, 24–34.

[151] J. Kussmann, C. Ochsenfeld, *J. Chem. Phys.* **2013**, *138*, 134114.

[152] J. Kussmann, C. Ochsenfeld, *J. Chem. Theory Comput.* **2015**, *11*, 918–922.

[153] J. Kussmann, C. Ochsenfeld, *J. Chem. Theory Comput.* **2017**, *13*, 3153–3159.