universität wien

# MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Endstate corrections from molecular mechanics to quantum machine learning potential“

verfasst von / submitted by

Sara Tkaczyk, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2022 / Vienna, 2022

| | |
|---|---|
| Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet: | UA 066 862 |
| Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet: | Masterstudium Chemie |
| Betreut von / Supervisor: | Univ.-Prof. Mag. Dr. Stefan Boresch |

# Acknowledgements

First of all, I would like to thank Univ.-Prof. Mag. Dr. Stefan Boresch for taking me on as one of his Master students and for supervising this work. I am very grateful for your advice, explanations and discussions. Thank you for all the time and knowledge you share with your students. Secondly, special thanks are given to Univ.-Prof. Mag. Dr. Thierry Langer for giving me the opportunity to work in his group and providing me with the resources needed for this project. Also, sincere thanks are given to Dr. Marcus Wieder. Thank you for your immense support, your patience and the time spent on this project. Thank you for introducing me to many concepts and workflows. I am very grateful for your guidance and everything I could learn from you. I would also like to thank Johannes Karwounopoulos, B.Sc. M.Sc. for the collaboration and discussions on this project. Furthermore, I want to thank all members of the Langer research group for a warm welcome and (even if partially remote) very nice working atmosphere. Lastly, I want to thank my family and loved ones for their unconditional support and company in good and challenging times. Thank you, God, for all the wisdom, goodness and love of all people who were involved and accompanied me though this project.

# Contents

# 1. Introduction

## 1.1. Free energy calculations in drug discovery

Free energies are central for determining the behavior and stability of a system, which is at or close to equilibrium [30]. In a (bio)chemical context, the free energy difference determines which side of the reaction is favored and whether the reaction occurs voluntarily. Reactions with a negative free energy difference between reactants and products will take place voluntarily and are called exergonic. Endergonic reactions, on the other hand, are characterized by a free energy increase and are not favored. Thus, leaving kinetic effects aside, the calculation of free energy differences between reactants and products allows to predict equilibria of biologically relevant processes. The free energy can be calculated with a broad range of methods typically referred to as free energy calculations. Free energy calculations are increasing in popularity in computational chemistry and drug discovery and are, thus, gradually integrated into a variety of molecular dynamics (MD) programs. Additionally, with the increasing speed and efficiency of MD simulations, free energy calculations also gain interest in the industrial sector [38, 42].

In drug design free energy calculations can aid the development process by addressing a number of relevant research questions [39]: For instance, solvation free energies are calculated to predict bioavailability and to understand solvation mechanisms [6, 29]. By computing the transfer free energy of a small molecule between two phases, membrane-water partition coefficients can be determined [16]. Moreover, the calculation of ligand/receptor binding free energies (i.e. affinity) is central to lead optimization applications in pharmaceutical drug development. When the receptor, the binding site and the binding mode are defined, it is of interest to maximize the potency (i.e. affinity) of the ligand by modifying the molecular scaffold. Here, the computational assessment of favored binders can greatly reduce the cost of synthesizing multiple variations of a compound [8]. Simulating such a binding process can be computationally demanding, if the direct (i.e. physical) binding path is considered (cf. $\Delta G_A$ and $\Delta G_B$ in figure 1.1). The reason for the high computational cost is that the timescale of those calculations is dominated by the dissociation constant of the ligand, which can reach up to microseconds or seconds for a typical drug [27]. There are also alchemical absolute free energy calculations, which do not target the direct simulation of the binding event (i.e. alchemical absolute free energy calculations), but these are typically

harder to converge and show comparably larger errors in their free energy estimate.

Even with GPU (graphics processing unit) support and modern molecular dynamics software packages, the timescales of the physical binding event (association or dissociation of the ligand from the protein binding pocket) are inaccessible for pharmaceutical drug discovery. Using enhanced sampling methods are a potential approach to help with sampling these rare events. This, however, often requires prior knowledge of a high probable reaction coordinate.

Thus, it is more practical to approach this problem by calculating relative binding energies and making use of a thermodynamic cycle, as illustrated in figure 1.1. Instead of considering two binding events, two alchemical processes can be simulated, where ligand $A$ is transformed into ligand $B$, in solution as well as in complex with the protein. This unphysical (or alchemical) path might involve the modification of one of the ligands residues. As the free energy is a state function and, thus, path independent, the binding free energy difference can then be obtained analogously, as if both binding processes were simulated directly:

$$\Delta\Delta G_{\text{bind,AB}} = \Delta G_B - \Delta G_A = \Delta G_2 - \Delta G_1. \tag{1.1}$$

In principle, these alchemical binding free energies could be obtained instead of carrying out experimental measurements. However, there are a number of factors, which limit the accuracy of free energy calculations [31]. In general, computer resources are limited and so is the exploration of the conformational space. This may lead to a sampling problem, where important parts of the phase space are missed [45]. Another challenge lies in the system setup, where questions such as correct assignment of protonation states or the choice of tautomers may influence the final result [26, 40]. Finally, the energetic description of the molecular system plays a role. Classical force fields, which were developed for biological systems often struggle with the diversity of small organic molecules. In molecular mechanics (MM) force fields atoms are treated as spheres with fixed point charges; information about the location of electrons is not encoded. An additional challenge for classical force fields is the correct modeling of torsion angles [36], which stems from omitting important energetic contributions and naturally has an effect on accuracy. A correct energetic description of a small molecule is only possible with quantum mechanics (QM). However, even if only the ligand has to be evaluated, sampling at the QM level of theory is computationally not feasible. Apart from improved force field parameterization, this problem can be circumvented by using quantum machine learning (QML) potentials, which are capable of reproducing QM energetics at speeds comparable to highly optimized MM force fields (cf. section 1.2).

The best way to utilize a QML potential for free energy calculations is still an open question, but an obvious way to include different levels of theory in a free energy calculation is the use of so called endstate corrections (also known as bookending or indirect free energy calculations [19, 17]). In this collection of methods, the high level description is exclusively applied to the physical endstates of the free energy calculations (applicable to each physical node in the thermodynamic

cycle), as visualized in figure 1.1. The free energy difference between two levels of theory can be obtained by a variety of methods, which will be introduced in sections 1.1.2-1.1.4 below.

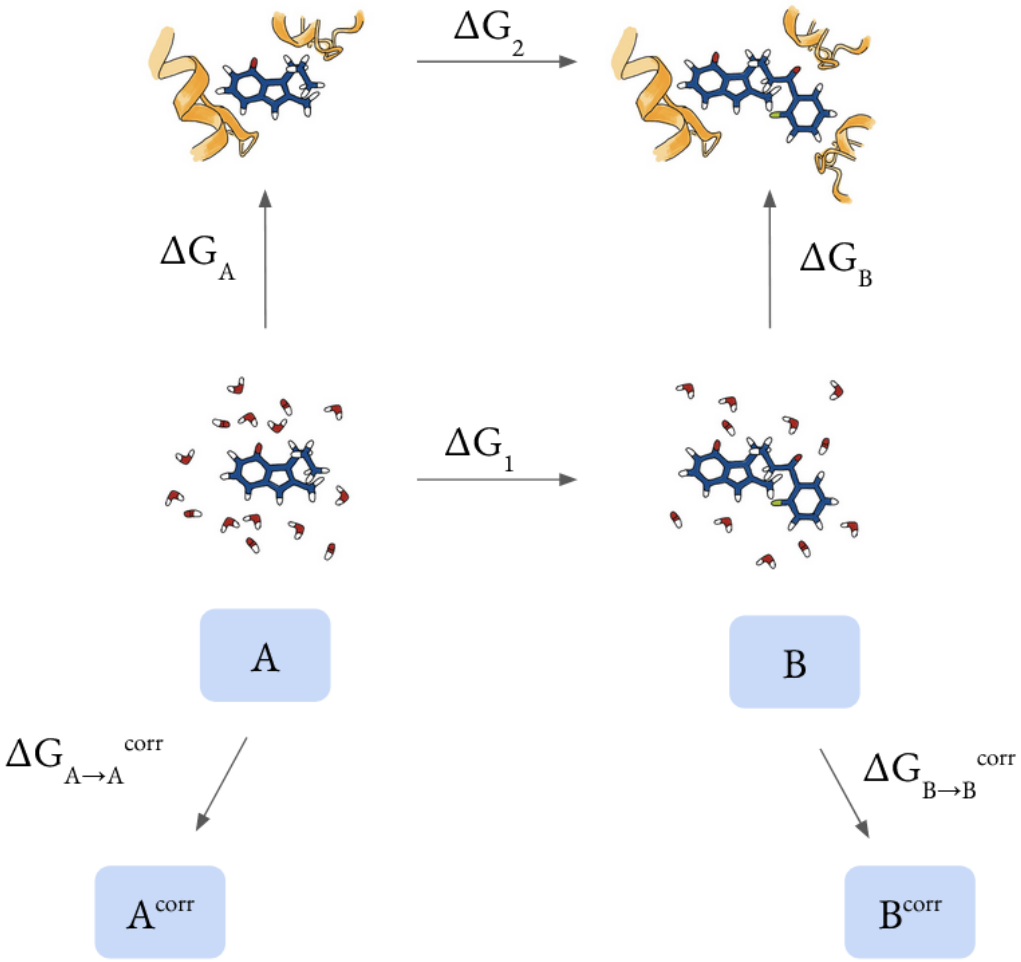


Figure 1.1.: Thermodynamic cycle for the relative binding free energy (see eqn. 1.1) of two ligands $A$ and $B$ to a receptor. Binding free energies calculated from the direct binding path (solvated ligand to ligand in complex) are denoted as $\Delta G_A$ and $\Delta G_B$, those along the alchemical paths (transformation of ligand $A$ into ligand $B$) are labeled as $\Delta G_1$ and $\Delta G_2$. $\Delta G_{\mathrm{A}\to\mathrm{A}^{\mathrm{corr}}}$ and $\Delta G_{\mathrm{B}\to\mathrm{B}^{\mathrm{corr}}}$ are the corrected free energies, which are used to correct the original binding free energy estimate. These corrections are applied to the physical endstates of the thermodynamic cycle, $A$ and $B$, respectively, to obtain a free energy estimate with the higher level of theory $\Delta G_{\mathrm{A}^{\mathrm{corr}}}$ and $\Delta G_{\mathrm{B}^{\mathrm{corr}}}$.

### 1.1.1. Motivation

The goal of this work was to find efficient methods for the correction of free energy differences from a low level to a higher level of theory and, in future work, to apply this methodology to binding and solvation free energies of biologically relevant molecular systems. Thus, considering the thermodynamic cycle illustrated in figure 1.1 the focus of the study was on one of the correction "legs" (e.g. $\Delta G_{\mathrm{A}\rightarrow\mathrm{A}^{\mathrm{corr}}}$). The correction was performed from a classical mechanical potential (CHARMM General Force Field (CGenFF) [37] — referred to as $U_{\mathrm{MM}}$, or $\lambda = 0$) to a QML potential (ANI-2x — referred to as $U_{\mathrm{QML}}$, $\lambda = 1$). Given that the correction only concerned the ligand, each system was studied in vacuum to reduce the complexity. Nevertheless, the methodology should be easily transferable to molecules in solution or as part of a complex (either using a hybrid QML/MM scheme or treating all atoms with a QML potential [15]).

In this work we were interested to ascertain which level of algorithmic complexity and computational effort is needed to obtain accurate free energy endstate corrections for a set of molecules (c.f. section 1.3). We first describe the most accurate, but also computationally most demanding step-wise equilibrium approach (see section 1.1.2). Additionally, the accuracy of two computationally less costly methods (FEP and NEQ; introduced in section 1.1.3 and 1.1.4) is tested. Especially for NEQ switching, we investigate the impact of switching length of the protocol on the obtained free energy estimates.

Moreover, the impact of switching length on the conformations obtained at the endpoints of the switching trajectories is investigated. For this purpose we focus on the dihedral angles of rotatable bonds. This decision was motivated by previous observations that torsion profiles are sensitive to different levels of theory and might cause convergence failures and inaccurate free energy estimates [24].

### 1.1.2. Equilibrium free energy calculations

Even though computationally demanding, equilibrium free energy calculations remain the method of choice whenever feasible. In an NVT ensemble, i.e. for a system with a constant number of particles, at constant temperature and volume, the Helmholtz free energy $A$ is minimized and given as:

$$A = -k_B T \ln Q \tag{1.2}$$

where $k_B$ is the Boltzmann's constant, $T$ the temperature and $Q$ the partition function of the system. The Helmholtz free energy is related to the Gibbs free energy with (and reduces to it in vacuum):

$$G = A + p\Delta V. \tag{1.3}$$

Typically, we are interested in the difference of free energy values (e.g. between two ligands in the same binding site). Such differences are directly related to the

ratio of the partition function of two states $A$ and $B$ [34]:

$$\Delta A_{AB} = -k_B T \ln \frac{Q_B}{Q_A} = -k_B T \ln \frac{\int_{V_B} exp(-\frac{U_B(\vec{x})}{k_B T}) d\vec{x}}{\int_{V_A} exp(-\frac{U_A(\vec{x})}{k_B T}) d\vec{x}} \tag{1.4}$$

where $U_A$ and $U_B$ are the potential energies, which are functions of the coordinates $\vec{x}$ (a $N \times 3$-dimensional vector with $N$ the number of atoms), and $V_A$ and $V_B$ are the conformational-space volumes of $\vec{x}$ (i.e. the total set of coordinates where the system shows a probability $\neq 0$ of being found). Free energy differences are calculated as follows. First, using molecular simulations, independent samples are generated from the equilibrium distribution of the molecular system. The estimate of the free energy difference is then obtained by using statistical methods, such as the Multistate Bennett Acceptance Ratio (MBAR) estimator [33]. Since the sampling procedure and the subsequent analysis are statistical processes, the obtained free energy differences are not exact, but rather estimated values with some uncertainty. MBAR is known as the free energy estimator with the lowest variance, provided there is sufficient phase-space overlap between the states of interest [34, 23]. If there is no phase-space overlap between the two physical states of interest, it is common to define several intermediate states, traditionally referred to as $\lambda$ or bridging states, which do have good overlap (cf. figure 1.2).

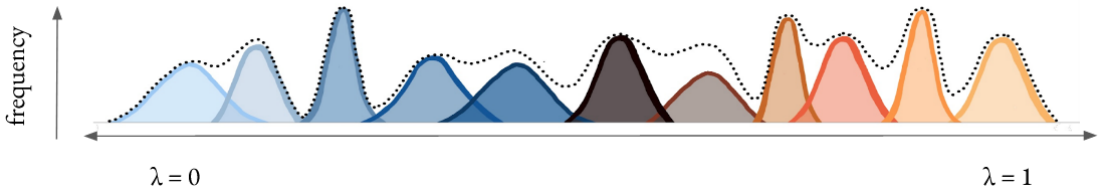


Figure 1.2.: Visual representation of 11 overlapping $\lambda$ states as a linear interpolation of the potential energy function $U(\vec{x}, \lambda)$ between two endstates denoted as $\lambda = 0$ and $\lambda = 1$. Each probability density function (PDF) represents an ensemble of samples $\pi(\vec{x}, \lambda)$, requiring a separate molecular simulation.

In practice this means that samples have to be drawn from both endstates ($\lambda = 0$ and $\lambda = 1$), and each of the intermediate states. Therefore, a sizeable number of molecular simulations is required to compute a single free energy difference of interest. The potential energy function describing the system is now a function of both the coordinates $\vec{x}$ and the coupling parameter $\lambda$ [27]. The simplest way to use the coupling parameter is to linearly interpolate between two states $A$ and $B$:

$$U(\vec{x}; \lambda) = (1 - \lambda) U_A(\vec{x}) + \lambda U_B(\vec{x}). \tag{1.5}$$

The final free energy difference is then estimated by using energy differences between configurations at all $\lambda$ states. Typically, equilibrium free energy calculations (given sufficient sampling and a correct system setup) are highly capable methods and lead to free energies with high accuracy; however, they are also expensive to perform. Specifically, equilibrium sampling is computationally demanding and has to be performed for each $\lambda$ state (usually 11 states).

Thus, in the following sections 1.1.3 and 1.1.4 we will introduce alternative algorithms to obtain the endstate free energy correction with lower computational cost. Both of these methods can be applied either uni- or bidirectionally. For bidirectional methods equilibrium samples have to be available at both endstates; however, no intermediate states have to be evaluated as with the equilibrium sampling procedure just described. In general, the limiting factor for the accuracy of the final free energy estimate is the phase-space overlap between both endstates. Unidirectional methods require an equilibrium ensemble of samples at only a single endstate. Again, no intermediate states have to be sampled, which further reduces the computational cost. Nevertheless, unidirectional methods have one major drawback: As opposed to bidirectional methods they do not have the capability to compensate for tailed distributions and tend to overestimate the weight of these isolated samples in the final free energy estimate. Further details of these procedures will be discussed in the following sections.

### 1.1.3. Free energy perturbation

One of the earliest methods for calculating free energy differences from simulations is the Zwanzig relationship, also known as free energy perturbation (FEP), exponential averaging (EXP) or importance sampling [46]. First developed for Monte Carlo simulations this sampling method is used to infer a target distribution based on samples generated from an initial distribution, which approximates the targeted distribution (and has common support) [28]. Weights are then assigned to compensate for the approximation:

$$w^{(i)} = \frac{f(x^{(i)})}{g(x^{(i)})} \tag{1.6}$$

where $w^{(i)}$ are the importance weights for the $i^{th}$ sample, $f(x^{(i)})$ defines the target distribution, and $g(x^{(i)})$ the initial distribution (from which sampling is possible). The expectation $a(x)$ of a quantity, such as the free energy, can then be estimated with respect to the distribution, which is defined by $f(x^{(i)})$ as follows:

$$\overline{a} = \frac{\sum_{i=1}^{N} w^{(i)} a(x^{(i)})}{\sum_{i=1}^{N} w^{(i)}}. \tag{1.7}$$

Such a procedure comes in handy if sampling directly from the target distribution is computationally demanding or not possible. This approach can be rephrased in the

context of endstate corrections: Here, the "initial" distribution would correspond to the equilibrium ensemble $\pi(\vec{x})$ generated with the MM force field ($U_A(\vec{x})$ when keeping the generalized notation). To obtain the free energy correction, the energy difference between $U_A(\vec{x})$ and $U_B(\vec{x})$ (i.e. the QML potential) has to be calculated, which corresponds to evaluating the samples drawn from $\pi(\vec{x})$ with the target potential (to obtain the target distribution). To make an analogy to importance sampling, this calculated energy difference corresponds to the assigned weights. The average of the negative of this energy difference is then used to obtain the free energy difference [24]:

$$\Delta G_{\mathrm{A}\to\mathrm{B}} = -k_B T \ln\left\langle \exp\left(\frac{-\Delta U_{\mathrm{A}\to\mathrm{B}}}{k_B T}\right)\right\rangle_{\mathrm{A}} \tag{1.8}$$

where the angle brackets denote that the average is taken over the ensemble generated at the low level of theory (MM; unidirectional protocol). As the switch between potential $A$ and $B$ is carried out without any intermediate steps, this method is also known as instantaneous switching; however, in this work it will be referred to as FEP. The advantage of this method is that the samples (usually generated at the low level of theory) can be evaluated with the target potential in a post-processing step, which is very cost-efficient. On the downside, the convergence of the resulting free energy estimate depends on the variance of the importance weights (i.e. the energy differences). If after several repetitions of an FEP protocol the resulting $\Delta E$ distribution is broad (i.e. has high variance), the free energy estimate will be based on only few values [18]. The reason for the estimate being dominated by "extreme" values can be seen from eqn.1.8; the energy difference is in the exponent. Therefore, even if not frequent, the most negative energy differences will contribute most significantly to the average, leading to convergence problems [30]. For FEP to work well, the distribution defined by the low-level potential has to be a good approximation to the target distribution. When this target distribution is complex or multimodal, FEP protocols are usually challenging. For molecular systems, disparities between low and high level of theory might be caused by differences in the conformation space obtained with both potentials [24].

Improved free energy estimates (more than an order of magnitude more precise) can be obtained if $\Delta U_{\mathrm{A}\to\mathrm{B}}$ and the energy differences in the backward direction ($\Delta U_{\mathrm{B}\to\mathrm{A}}$) are accessible [33, 24]:

$$\Delta G_{\mathrm{A}\to\mathrm{B}} = k_B T \ln\left(\frac{\left\langle f(U_{\mathrm{A}} - U_{\mathrm{B}} + C)\right\rangle_{\mathrm{B}}}{\left\langle f(U_{\mathrm{B}} - U_{\mathrm{A}} - C)\right\rangle_{\mathrm{A}}}\right) + C \tag{1.9}$$

with $f(x) = (1+\exp(x/k_B T))^{-1}$ and $C = k_B T \ln(Q_{\mathrm{A}} N_B / Q_B N_{\mathrm{A}})$. $Q$ is the partition function and $N$ denotes the number of samples from each state. This relation is known as Bennett Acceptance Ratio (BAR) [2] and can be seen as a special case of MBAR, where only two states are considered. Like with MBAR, phase-space overlap is required to obtain a reasonable result. This bidirectional method

requires samples from the equilibrium distribution in state $A$ and state $B$. The increased accuracy and higher resilience against single outliers comes with a higher computational cost.

### 1.1.4. Non-equilibrium free energy methods

Methods referred to as non-equilibrium (NEQ) switching or annealed importance sampling make it possible to compute equilibrium free energies from NEQ work [10, 43].

The maximum work theorem states that the free energy difference $\Delta G$ between the equilibrium endstates is less than or equal the amount of work performed on or extracted out of a system during a non-equilibrium (NEQ) process. The equal sign only holds if the transformation is realized reversibly with the system being at equilibrium at all times. Taking the average over many realizations of the NEQ process, denoted by the angle brackets, is pivotal if microscopic systems are the object of study, i.e.

$$\langle W \rangle \geq \Delta G. \tag{1.10}$$

To clarify, consider a gas filled cylinder with a piston and a biomolecule within an atomic force microscope as examples of a macroscopic and microscopic system, respectively. Assuming, that the measuring set-up and protocol for the macroscopic experiment can be designed in a way such that each compression or extension of the gas within the cylinder can be repeated in exactly the same way (which is not feasible in reality), each measurement would yield the same amount of work $W$. In such a case, the average in eqn. 1.10 is not necessary. In a microscopic system, however, thermal fluctuations come into play and many repetitions of the same experiment, e.g. stretching a biomolecule with an atomic force microscope, usually lead to a distribution of $W$. Moreover, it might even happen that the obtained work values are smaller than the free energy difference. Thus, the average in eqn. 1.10 is essential. Even if the described process drives the microscopic system out of equilibrium, it was shown that it is possible to extract equilibrium properties out of non-equilibrium trajectories. In 1997 Christopher Jarzynski formulated a relation, often referred to as the Jarzynski equation, which transforms the maximum work theorem into an equality by relating the equilibrium free energy difference to NEQ work [22] according to

$$\langle e^{-\frac{W}{k_B T}} \rangle = e^{-\frac{\Delta G}{k_B T}}. \tag{1.11}$$

In this relation, averaging over the work exponential can be viewed as a compensation for driving the system out of equilibrium. In other words, it corresponds to removing the bias which is introduced during the NEQ process. This theorem has found applications in experimental settings, such as the interpretation of single-molecule pulling experiments, and also in the computational field, where the development of more efficient free energy calculation algorithms is of interest

[20, 21, 43]. One of the major advantages of NEQ switching protocols is that multiple trajectories can be processed in parallel, making it computationally efficient provided a sufficient number of compute cores is available.

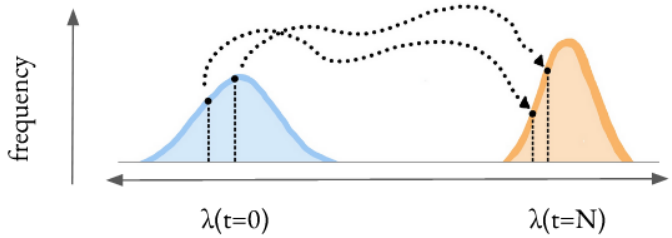


Figure 1.3.: Visual representation of the NEQ switching protocol. The displayed PDFs illustrate both endstate ensembles $\pi(\vec{x}, \lambda = 0)$ and $\pi(\vec{x}, \lambda = 1)$ (as in figure 1.2). In the context of endstate corrections, this figure shows NEQ switching from $U_{\mathrm{MM}}$ to $U_{\mathrm{QML}}$. The black points correspond to selecting individual samples and moving along a NEQ trajectory (dotted curve) to the target PDF. The parameter $\lambda$ is changed with time and controls the interpolation between both endstate potentials according to eqn. 1.12.

The prerequisite for running a NEQ switching process is an equilibrium ensemble of samples, generated as described in section 1.1.2. Samples can be either drawn from the $\pi(\vec{x}, \lambda = 0)$, as visualized in figure 1.3, or the $\pi(\vec{x}, \lambda = 1)$ equilibrium distribution (unidirectional protocol). Starting with an initial sample at state $\vec{x}(t = 0)$, a typical non-equilibrium protocol is defined as an iterative and alternating series of propagation steps (sampling the given equilibrium distribution) and perturbation steps (changing the potential using the coupling parameter $\lambda$). When propagating the coordinates of the system, state $\vec{x}(t + \Delta t)$ is computed using a MD step while keeping the coupling parameter at $\lambda(t)$. The perturbation step involves modifying the alchemical coupling parameter $\lambda(t)$ to $\lambda(t + \Delta t)$ without changing the state $\vec{x}(t + \Delta t)$. While the $\lambda$ parameter is altered, the corresponding potential $U_t$ is changed according to:

$$U_t = \lambda(t)U_A + (1 - \lambda(t))U_B. \tag{1.12}$$

The irreversible process of altering $\lambda$ at a finite rate over a time interval $\tau$ drives the system out of equilibrium. Thus, the perturbation step is the part of the protocol, where work is carried out on the system. Performing such a protocol, the work accumulated along a particular trajectory can be written as [10]:

$$W(t + \Delta t) = W(t) + U(\vec{x}_{t+\Delta t}, \lambda_{t+\Delta t}) - U(\vec{x}_{t+\Delta t}, \lambda_t). \tag{1.13}$$

Repeating this protocol multiple times produces the $W$ distribution. To obtain the free energy difference, those work values are evaluated with eqn.1.11, which can be

formulated as [22]:

$$\Delta G_{\mathrm{A}\to\mathrm{B}} = -k_B T \ln\left\langle \exp\left(\frac{-W_{\mathrm{A}\to\mathrm{B}}}{k_B T}\right)\right\rangle_{\mathrm{A}}. \tag{1.14}$$

If the NEQ protocol is initialized from both endstates and switching is performed in both directions ($A \to B$ and $B \to A$), the switching procedure is referred to as bidirectional. The resulting distributions $W_{\mathrm{A}\to\mathrm{B}}$ and $W_{\mathrm{B}\to\mathrm{A}}$ are evaluated with the Crooks' equation [9]:

$$\Delta G_{\mathrm{A}\to\mathrm{B}} = k_B T \ln\left(\frac{\left\langle f(W_{\mathrm{B}\to\mathrm{A}} + C)\right\rangle_{\mathrm{B}}}{\left\langle f(W_{\mathrm{A}\to\mathrm{B}} + C)\right\rangle_{\mathrm{A}}}\right) + C. \tag{1.15}$$

Visually, the free energy difference is the intersection point of the resulting forward and backward work distributions (only strictly true if the work distributions are normal distributions) [18]. Initiating a bidirectional NEQ switching protocol requires sampling at both levels of theory ($A$ and $B$), which is computationally more demanding, but yields more accurate free energy estimates (because exponential averaging can be avoided, c.f. uni- and bidirectional methods in section 1.1.3).

NEQ switching protocols can be either fast, where the number of time steps needed to switch from potential $U_A$ to $U_B$ (or *vice versa*) is small, or slow, where the system has more time to adapt to the changing potential. In the limiting case, where the switch from $U_A$ to $U_B$ is carried out instantaneously without any integration step in between, the protocol reduces to the FEP method already introduced in section 1.1.3.

## 1.2. Quantum machine learning potentials

As already mentioned in section 1.1, QML potentials address a range of issues which lead to inaccurate results when using classical force fields. Being trained on QM data, electronic information is encoded in the model, which aids in modeling complex interatomic interactions and is essential when chemical bonds are formed or broken [31]. These qualities make QML potentials applicable for simulations that would benefit from a more accurate energetic description but also enable the study of e.g. tautomerization, which is highly relevant in drug discovery [41]. Another major advantage is their ability to combine the strengths of both MM force fields (computational efficiency) and QM calculations (chemical accuracy), without making large compromises. QML potentials are fully differentiable, which means that results obtained in a first iteration can be used to retrain the network, if needed. In general, QML potentials can be broadly classified in two types [35]: (1) bespoken QML potentials are designed for a very specific or only small class of systems. Despite high accuracy, they have a limited range of use and come with high costs, as QM calculations are necessary prior to any application [44]. (2)

transferable QML potentials, on the other hand, are developed for predicting the potential energy of large classes of molecules, without retraining the model. For transferable QML potential the predicted energies are still on the order of chemical accuracy (1 kcal/mol).

ANI-2x (Accurate NeurAl networK engINe for Molecular Energies), presented in section 1.2.1, is one example of a transferable QML potential. Despite many advantages, QML potentials also come with several limitations: First, even though orders of magnitude faster than QM, QML potentials are still distinctively slower than a GPU-accelerated MM simulation, which might lead to sampling problems [15]. Most QML potentials are trained on density functional theory (DFT) data, which contains electronic information; however, DFT methods are known for sporadic failures and their inability to model exchange interactions correctly [7]. QML potentials are mostly black-box functions, which means that it is difficult to identify potential point of failures due to their obfuscated functional form. This might lead to artefacts (which might also be hard to identify) if those potentials are applied outside their training domain [41]. Finally, it is still an open question how an alchemical protocol (alchemically mutating from one ligand to the other) would look like using a QML potential (with the exception of the most trivial case [41]).

### 1.2.1. ANI potential

ANI-2x is a transferable deep learning potential for organic molecules [35, 11]. The neural network was trained to QM energies and forces of equilibrium and off-equilibrium conformations (in total $\approx$ 26M conformations from 70K molecules) of molecules containing up to seven different elements (C, H, N, O, F, Cl, and S). With those elements approximately 90% of drug-like molecules can be covered. However, highly abundant chemical species, such as ions and cofactors, or phosphorus-containing compounds cannot be described. The training QM energies were calculated at the DFT level of theory using the $\omega$B97X functional and the 6-31G* basis set. ANI-2x was shown to predict molecular energies and forces within chemical accuracy and a $10^6$ speedup compared to the DFT level of theory. Additionally, ANI-2x was refined to better predict molecular torsion profiles. The training data was derived from a variety of sources, such as the GDB-11 [14, 13] and the CheMBL [3] database, as well as the s66x8 [5] benchmark.

ANI-2x is implemented as an element-specific multilayer perceptron, as displayed in figure 1.4. For the molecular representation the neuronal network uses a modified version of Behler and Parrinello symmetry functions [1]. These symmetry functions are used to build element-specific single-atom environment vectors, taking into account the radial and angular chemical environment. The local atomic environment is modeled with cutoff functions, which brings along the substantial limitation of missing long-range interactions [15]. Each atom of a molecule is evaluated in a separate neuronal network. As shown in figure 1.4, the total energy of a molecule

is calculated as a sum of all atomic energies. Forces, which are required for the MD simulation are computed as the gradient of the predicted energy values.

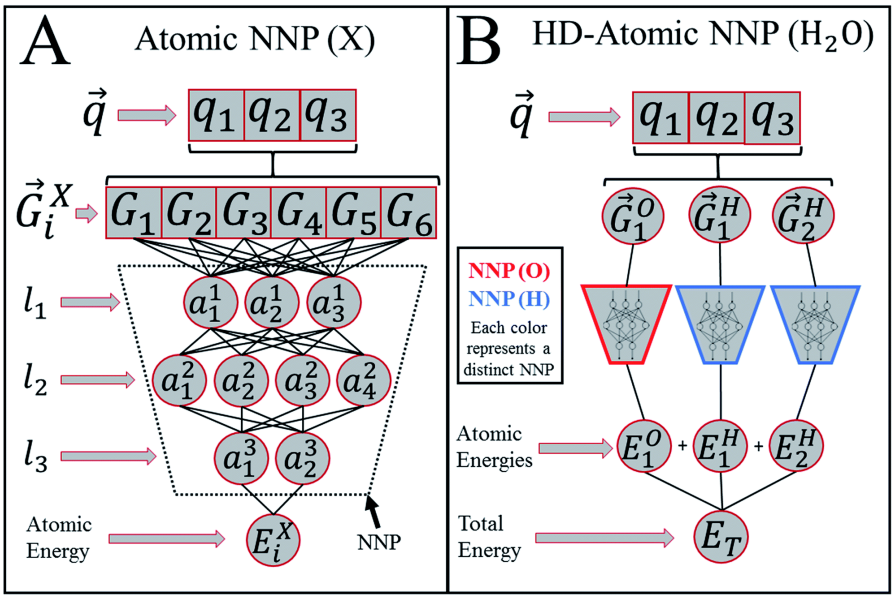


Figure 1.4.: The structure of the high-dimensional QML potential ANI-2x. (A) The coordinates $x$ are used as input to generate an atomic environment vector $\vec{G}_i^X$, which is forwarded to a neuronal network. The output is the atomic energy $E_i^X$. (B) An example model for a water molecule. Each atom is treated with a separate neuronal network; the atomic energies are summed to the total energy $E_T$. For more details see [35].

## 1.3. Dataset

Figure 1.5 shows the molecules which were selected for this work. The dataset contains 19 small molecules, with the majority having one or more aromatic groups. The compounds in the dataset contain nitrogen, sulfur, chlorine and oxygen as heteroatoms. The elements contained in these molecules are also present in the training set of the QML potential ANI-2x, which was used alongside with CGenFF for the potential energy description of the molecules.

In the course of the study, for several molecules torsion profiles were evaluated. These are also indicated in figure 1.5. Torsion profiles were generated for the endstate samples from the equilibrium simulations ($\pi(\vec{x}, \lambda = 0)$ and $\pi(\vec{x}, \lambda = 1)$) and from the NEQ trajectories (samples taken at t=$\tau$, with $\tau$ = 5 ps, 10 ps, 20 ps and 50 ps).

The compounds investigated in this work were taken from the so-called `HiPen`

dataset [24], which in turn, is a small subset of the `Maybridge` collection. The name `HiPen` refers to the fact that the CGenFF parameters for these particular molecules all had high penalties assigned by `ParamChem`. Since parameters are assigned using similarity criteria (i.e. parameter lookup in a database, if no match is found, parameters are inferred using similarity criteria) these parameters have a high uncertainty associated. From previous work ([24]) this selection of molecules have shown convergence failures for endstate corrections from molecular mechanics force fields to semi-empircal quantum mechanics. These characteristics (high penalty criterion in combination with the chemical diversity and convergence problems) make the `HiPen` data set challenging for accurate free energy estimates.

In the original work the molecules were grouped into three categories ("good", "bad", and "ugly") based on their behavior in various endstate reweighting schemes (uni-, and bidirectional FEP and NEQ switching — for more details on these methods see section 2). The "good" label was assigned to molecules for which the free energy could be estimated using a unidirectional NEQ switching approach. If this criterion was not met, neither for the forward, nor for the backward switching protocol, the compound was labeled "bad". Molecules for which it was not possible to obtain a converged free energy estimate (regardless of the used approach) were labeled "ugly". These annotations are included in table 3.1; the corresponding compound number alongside with the ID used in the original HiPen work is shown in figure 1.5.

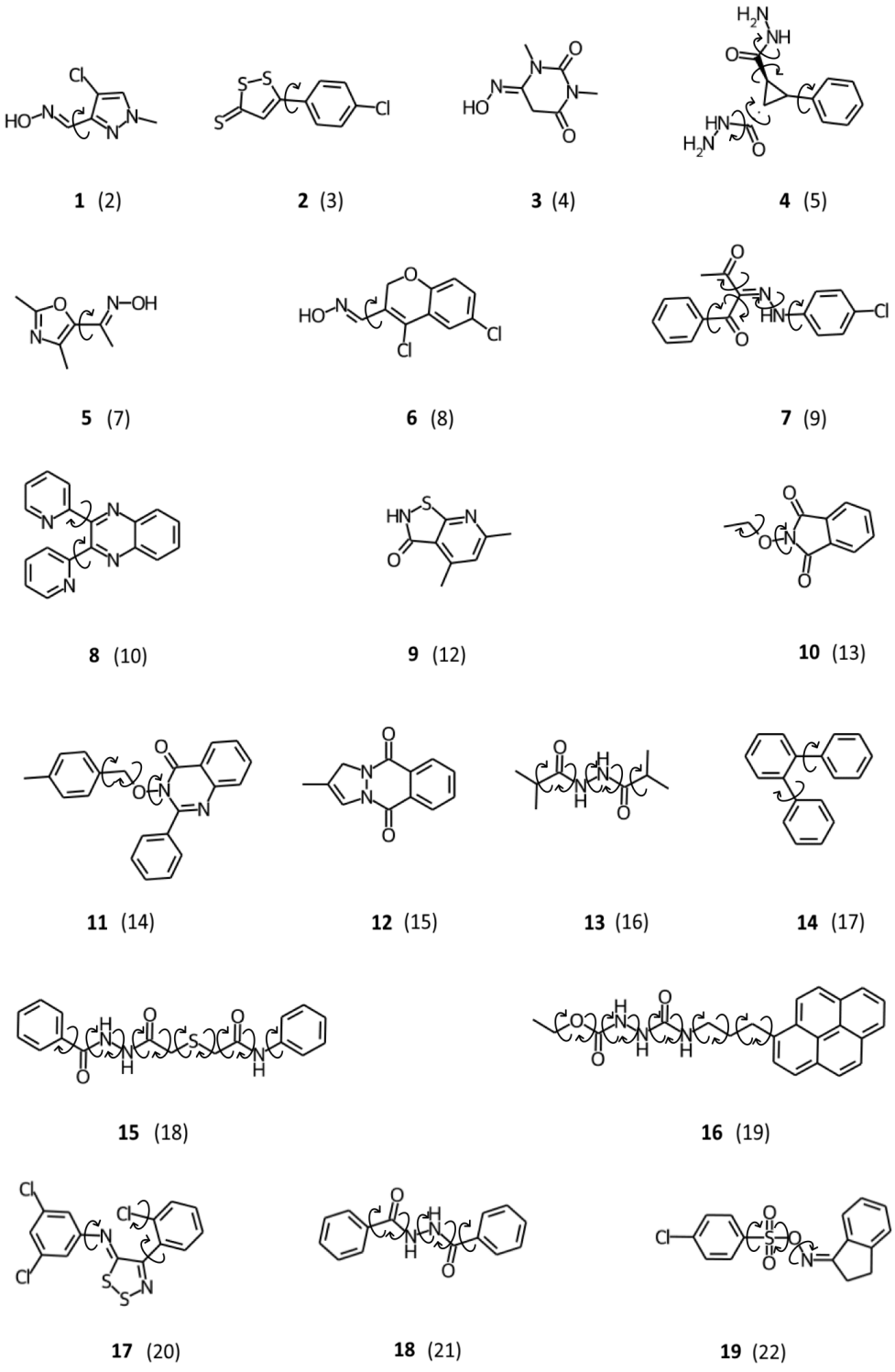


Figure 1.5.: Dataset studied in this work. Compound IDs used in this work are shown in bold print; original `HiPen` numbering is indicated in parentheses. The corresponding ZINC Database IDs can be found in table 3.1. The heavy atom torsions investigated in detail are indicated for each molecule.

# 2. Methods

## 2.1. Dataset, system setup and simulation details

For this work 19 molecules were selected from the `HiPen` dataset (cf. Section 1.3 and [24]); their structures are shown in Figure 1.5. The compound IDs were adjusted to ensure continuous numbering. Parameter and topology files were taken from the data repository associated with [24]) (see doi:10.5281/zenodo.2328952). Molecules were parameterized using the CHARMM Generalized Force Field (CGenFF 3.0.1) and `ParamChem`. Initial 3D coordinate files were obtained from the $\text{ZINC}_{12}$ Structural Database.

All simulations were performed with the `openMM` simulation package [12]. Each molecule was set up in vacuum with no cutoff to the non-bonded interactions. Langevin dynamics was performed with an integration step of 1 fs and a collision rate of 1 $\text{ps}^{-1}$. Random velocities were drawn from a Maxwell-Boltzmann distribution at the simulation temperature of 300 K. Simulations were carried out on a Nvidia `GeForce GTX 1080 Ti` graphics processing unit (GPU) using the `CUDA toolkit 10.3`.

The energetic description of the molecular system at the molecular mechanics (MM) endstate ($U_{\text{MM}}$) was defined using the CGenFF parameters and topology as described above, for the quantum machine learning (QML) description at the QML endstate ($U_{\text{QML}}$) the ANI-2x neural network potential was used (as implemented in `torchani`[1]). The potential energy of the intermediate $\lambda$ states was defined as a linear combination of both endstate potential energy functions ($U_{\text{MM}}$ and $U_{\text{QML}}$) and controlled by the alchemical coupling parameter $\lambda \in [0, 1]$ (as described in Section 1.1.2). To perform the interpolation between the potential energy functions, the `openmm-ml`[2] package was used.

## 2.2. Endstate corrections

### 2.2.1. Equilibrium free energy calculations

Equilibrium free energy calculations were performed for each of the investigated systems using 11 alchemical $\lambda$ states, $\lambda = 0.0, 0.1, \ldots, 1.0$. For each of the systems the free energy calculation was independently repeated three times. Each of the

[1] `https://github.com/aiqm/torchani`
[2] `https://github.com/openmm/openmm-ml`

repeats was started from a random conformation, which was generated with the chemoinformatics package `rdkit`[3].

For each of the alchemical $\lambda$ states the sampling procedure followed the same general approach: the system was minimized using the `L-BFGS` optimizer with a maximum iteration number of 100. Subsequently, for each $\lambda$ state simulations were performed for 5 ns and samples were saved every 1 ps. To ensure equilibration, the first 1 ns of each trajectory was discarded (i.e. the first 20% of the samples). The $11 \times 4,000$ samples were further pruned and only every $10^{th}$ sample was used for the final free energy calculation. The free energy was calculated using the Multistate Bennett Acceptance Ratio (MBAR) estimator, as implemented in the `pymbar`[4] Python package.

Uncertainty of the free energy estimate as well as the overlap of the probability density functions of neighboring $\lambda$ states were visually checked. Additionally, the overlap between the forward and backward work ($W$) and energy ($\Delta E$) distributions was computed using `pymbar`. Finally, the atom indices of all rotatable bonds forming heavy atom dihedral angels were recorded and the corresponding torsion profiles at both endstates were computed with the Python package `MDTraj`[5].

### 2.2.2. Non-equilibrium switching protocol

The non-equilibrium (NEQ) switching protocols were initialized by randomly selecting conformations saved during the equilibrium sampling calculations, either drawing from the MM ensemble, $\pi(\vec{x}, \lambda = 0)$, or from the QML equilibrium distribution, $\pi(\vec{x}, \lambda = 1)$. The ensembles $\pi(\vec{x}, \lambda = 0)$ and $\pi(\vec{x}, \lambda = 1)$ contained samples from three independently repeated simulations, resulting in a pool of $3 \times 4,000$ conformations.

A single NEQ protocol consists of slowly interpolating the potential energy function between endstates, using the coupling parameter $\lambda$. For a given time step $t$ of a switch of length $\tau$, $\lambda$ was calculated according to $\lambda = \frac{t}{\tau}$. Each perturbation of the control parameter $\lambda$ resulted in a change of the alchemical system potential energy, according to eqn. 1.12. Each perturbation step was followed by a propagation step with a step size of 1 fs at a fixed $\lambda$, using otherwise identical conditions as those of the equilibrium MD protocol described above. This pattern of propagation and perturbation was repeated until the simulation length $\tau$ was reached, and the control parameter $\lambda$ was either 0 or 1. NEQ switching protocol lengths $\tau$ of 5, 10, 20 and 50 ps were used.

Both unidirectional (samples drawn from state $A$ and switching from state $A$ to state $B$) as well as bidirectional protocols (samples drawn from state $A$ and switching from state $A$ to state $B$ and *vice versa*) were performed.

[3] `https://github.com/rdkit/rdkit`
[4] `https://github.com/choderalab/pymbar`
[5] `https://github.com/mdtraj/mdtraj`

Uni- and bidirectional switches were performed 500 times, resulting in 500 dimensionless work values $W$. Unidirectional work values were evaluated using the Jarzynski equation, bidirectional work values the Crooks' equation (see eqn. 1.14 and 1.15, respectively) to obtain free energy estimates $\Delta G$. Those estimates were generated by implementing the `EXP` and `BAR` routines in `pymbar`. The cumulative standard deviation of the $W$ values was recorded for the 500 switches to monitor its convergence. To investigate the difference in the systems' torsion profiles, the final coordinates at the end of a switch were saved as well.

### 2.2.3. Free energy perturbation

We also performed uni- and bidirectional free energy perturbation (FEP) for each of the systems. Samples were randomly drawn from the three independent simulations at each of the endstates ($3 \times 4,000$) and 5,000 $\Delta E$ values were calculated. Using only the unidirectional energy differences, the Zwanzig equation (see eq. 1.8, as implemented in the `EXP` routine in `pymbar`) was used to calculate the free energy estimate. To estimate the free energy difference using bidirectional energy differences the Bennett Acceptance Ratio (BAR) estimator (eq. 1.9, as implemented in the `pymbar` Python package) was utilized.

# 3. Results and Discussion

This chapter is structured as follows: I start by presenting the results obtained from equilibrium free energy calculations for the molecules of the `HiPen` dataset. Next, an overview of uni- and bidirectional FEP, as well as NEQ switching results will be given. This is followed by a comparison of the three methods. Additionally, the results will be discussed in detail using the variance of the accumulated work (for NEQ switching) and the potential energy difference (for FEP). To analyze the conformational sampling capability of the NEQ switching approach we investigate torsion profiles obtained from the switching trajectories. Finally, the impact of longer switching lengths on the behavior of the NEQ protocols will be assessed.

Due to a high number of investigated molecules and features the majority of the figures were moved to the Supporting Materials section. Only tables and figures relevant for the following discussion are shown in this chapter.

## 3.1. Equilibrium free energy estimates

We start by visually inspecting the equilibrium trajectories of the physical endstate to ensure reasonable behavior. With the exception of molecule **19** (c.f. figure 1.5), none of the investigated molecules showed any signs of instability, pathological or non-physical behavior. Simulations performed with the QML potential were noticeably more "fluid" than simulations performed with the MM force field. This observation is, maybe, an effect of the harmonic functional of bonds/angle terms of the MM potential.

Molecule **19** showed instability during simulations with the QML potential and heavy atom bonds were broken early in the simulation. The bond that showed this behavior is located at the sulfonate ester moiety connecting the sulfonate group and oxygen atom, which does not reflect a suitable leaving group. Other sulfur-containing molecules did not show similar behavior and molecule **19** was the only sulfonate compound in the studied dataset. Further investigation to understand this behavior is needed, but outside the scope of this work, and, therefore, the molecule was excluded from further investigation.

Information about all compounds and the corresponding equilibrium free energy estimates are summarized in table 3.1. For each molecule the ZINC database identifier, the original `HiPen` numbering scheme, and the corresponding `HiPen` group affiliation is shown. The equilibrium free energy differences between the MM and QML level of theory were obtained by repeating the free energy calculation

Table 3.1.: Equilibrium free energy results for all studied molecules. The corresponding ZINC database identifiers as well as the `HiPen` label are listed. $\Delta G_{\mathrm{eq}}$ is the free energy difference between the MM and QML level of theory and is given as the mean of three sampling runs (in $k_B T$).

| Cmp. Nr. | `HiPen` Nr. | ZINC ID | `HiPen` Group | $\Delta G_{\mathrm{eq}}$ | |
|---|---|---|---|---|---|
| 1 | 2 | 00077329 | Good | -940,543.68 | ± 0.08 |
| 2 | 3 | 00079729 | Good | -2,105,811.38 | ± 0.09 |
| 3 | 4 | 00086442 | Good | -656,584.59 | ± 0.11 |
| 4 | 5 | 00087557 | Bad | -839,346.75 | ± 0.15 |
| 5 | 7 | 00107550 | Good | -560,516.43 | ± 0.09 |
| 6 | 8 | 00107778 | Ugly | -1,590,089.70 | ± 0.07 |
| 7 | 9 | 00123162 | Ugly | -1,407,890.86 | ± 0.14 |
| 8 | 10 | 00133435 | Good | -960,187.53 | ± 0.12 |
| 9 | 12 | 00140610 | Good | -939,376.93 | ± 0.10 |
| 10 | 13 | 00164361 | Good | -701,793.46 | ± 0.14 |
| 11 | 14 | 00167648 | Good | -1,167,290.21 | ± 0.12 |
| 12 | 15 | 00169358 | Good | -761,146.86 | ± 0.09 |
| 13 | 16 | 01755198 | Bad | -604,674.34 | ± 0.11 |
| 14 | 17 | 01867000 | Good | -730,760.50 | ± 0.09 |
| 15 | 18 | 03127671 | Bad | -1,521,955.63 | ± 0.11 |
| 16 | 19 | 04344392 | Bad | -1,347,288.72 | ± 0.14 |
| 17 | 20 | 04363792 | Ugly | -2,972,461.52 | ± 0.11 |
| 18 | 21 | 06568023 | Bad | -842,682.65 | ± 0.08 |
| 19 | 22 | 33381936 | Bad | - | - |

three times and reporting $\Delta G_{\mathrm{eq}}$ as the average value including the corresponding standard deviation (sd). In this work, all free energy values are given in $k_B T$. The individual $\Delta G_{\mathrm{eq}}$ values for each run can be found in the Supporting Materials section. To assess whether the equilibrium result delivered by the MBAR estimator is a reliable reference, the overlap between neighboring $\lambda$ states was monitored. For each studied molecule good overlap was found between consecutive states. A more detailed description and an example of an overlap plot can be seen in figure 3.1 in section 3.4.1.

## 3.2. Non-equilibrium switching

I start by discussing the results obtained for the 5 ps switching length shown in table 3.2. Longer switching lengths will be discussed in section 3.5. Estimates derived from bidirectional protocols with the Crooks fluctuation theorem (c.f. eqn. 1.15) are denoted as $\Delta G_{\mathrm{NEQ,bi}}$. For unidirectional results, the switching direction

is indicated in the superscript as $\Delta G^{\mathrm{MM,QML}}_{\mathrm{NEQ,uni}}$ and $\Delta G^{\mathrm{QML,MM}}_{\mathrm{NEQ,uni}}$, respectively. Those values are obtained with the Jarzynski equality, eq. 1.14. To make discussing the results easier filtering criteria were defined that help to interpret the accuracy and precision of the free energy estimates. Thus, it will also be easier to make statements about the efficiency of the various methods and protocols compared in this work.

The bidirectional results $\Delta G_{\mathrm{NEQ,bi}}$ depend on the overlap of the underlying work distributions, the corresponding values are listed in table 3.2. The overlap is calculated as the percentage of overlap between the $W$(MM→QML) and $W$(QML→MM) distributions and ranges from 20% for molecule **7** up to 96% for compound **2**. The accuracy was determined by comparing the obtained free energy estimates to the equilibrium free energy estimate. It was found that for every molecule studied the estimate obtained with the bidirectional method agreed excellently with this "gold standard". Even for the NEQ bidirectional free energy estimate with the lowest overlap of the work values (20.37 % for **7**), the Crooks equation delivered a reliable result.

By contrast, the unidirectional protocol (for which free energies were estimated using the Jarzynski equality) showed some deviations from the equilibrium free energy estimate. To detect and quantify this, two criteria are included in Table 3.2. First, molecules for which the deviation from the equilibrium value $\Delta G_{eq}$ is more than 1 $k_BT$ are marked. For the switching length of 5 ps, this was the case for three molecules (compounds **1**, **5**, and **6**).

Second, the "broadness" of the work (or, for FEP, of the energy difference) distribution (i.e. the variance) also may influence the accuracy of the free energy estimate. The variance was monitored by calculating the cumulative standard deviation of the work/energy values. An example of such a plot is shown in section 3.4.1. If the standard deviation exceeds the limit of 2 $k_BT$ (a generally accepted rule-of-thumb [10, 30]), this indicates possible pathologies in the resulting free energy estimate and marks results that should be interpreted carefully. Such variance assessments were made for each NEQ switching direction (i.e. MM → QML and QML → MM). Most of the obtained work distributions have final cumulative standard deviation below 2 $k_BT$. The 5 ps switching protocol yielded five work distributions with high variance (compound **4**, **7**, **13**, **15**, and **16**). Interestingly, there seems to be no correlation between high variance in the work distributions and incorrect unidirectional free energy estimates, except for one case (molecule **16**), which will be discussed in section 3.5.1 and section 3.5.2. All features presented above, as well as their evolution with respect to the switching length of the protocol will be also discussed in greater detail in section 3.5.

Table 3.2.: NEQ switching results for a switching length of 5 ps. Bidirectional results $\Delta G_{\mathrm{NEQ,bi}}$ were calculated with the Crooks' equation; the unidirectional ones $\Delta G^{\mathrm{MM,QML}}_{\mathrm{NEQ,uni}}$ and $\Delta G^{\mathrm{QML,MM}}_{\mathrm{NEQ,uni}}$ with the Jarzynski equation. Overlap between forward and backward work distributions is given in percent. If the free energy estimate obtained with Jarzynski ($\Delta G^{\mathrm{MM,QML}}_{\mathrm{NEQ,uni}}$) deviates more than 1 $k_BT$ from the equilibrium value $\Delta G_{\mathrm{eq}}$, the corresponding molecule is marked with an x. Finally, if the cumulative standard deviation (cum. sd.) of the work distribution (either in one or the other switching direction) exceeds 2 $k_BT$, the compound is also marked with an x.

| Cmp. | $\Delta G_{\mathrm{NEQ,bi}}$ | | $\Delta G^{\mathrm{MM,QML}}_{\mathrm{NEQ,uni}}$ | | $\Delta G^{\mathrm{QML,MM}}_{\mathrm{NEQ,uni}}$ | | overlap $W$ (%) | $\Delta G^{\mathrm{MM,QML}}_{\mathrm{NEQ,uni}}$ diff. from $\Delta G_{\mathrm{eq}}$ | cum. sd. $W$ (MM,QML) $>2\ k_BT$ | cum. sd. $W$ (QML,MM) $>2\ k_BT$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | -940,543.84 | ± 0.03 | -940,545.64 | ± 0.18 | 940,543.80 | ± 0.03 | 84.08 | **x** | - | - |
| 2 | -2,105,811.41 | ± 0.01 | -2,105,811.45 | ± 0.02 | 2,105,811.38 | ± 0.02 | 95.81 | - | - | - |
| 3 | -656,584.61 | ± 0.02 | -656,584.62 | ± 0.03 | 656,584.62 | ± 0.03 | 92.46 | - | - | - |
| 4 | -839,346.73 | ± 0.05 | -839,346.71 | ± 0.14 | 839,346.84 | ± 0.12 | 58.22 | - | - | **x** |
| 5 | -560,516.72 | ± 0.03 | -560,517.55 | ± 0.09 | 560,516.52 | ± 0.04 | 83.14 | **x** | - | - |
| 6 | -1,590,089.69 | ± 0.02 | -1,590,094.74 | ± 0.54 | 1,590,089.69 | ± 0.03 | 91.01 | **x** | - | - |
| 7 | -1,407,890.85 | ± 0.12 | -1,407,890.42 | ± 0.25 | 1,407,891.77 | ± 0.44 | 20.37 | - | **x** | - |
| 8 | -960,187.64 | ± 0.03 | -960,187.78 | ± 0.09 | 960,187.63 | ± 0.05 | 83.76 | - | - | - |
| 9 | -939,376.84 | ± 0.02 | -939,376.88 | ± 0.02 | 939,376.81 | ± 0.02 | 93.89 | - | - | - |
| 10 | -701,793.56 | ± 0.03 | -701,793.56 | ± 0.04 | 701,793.53 | ± 0.06 | 81.19 | - | - | - |
| 11 | -1,167,290.24 | ± 0.05 | -1,167,290.37 | ± 0.12 | 1,167,290.13 | ± 0.11 | 64.70 | - | - | - |
| 12 | -761,146.92 | ± 0.02 | -761,146.91 | ± 0.03 | 761,146.91 | ± 0.03 | 92.50 | - | - | - |
| 13 | -604,674.30 | ± 0.08 | -604,673.89 | ± 0.23 | 604,674.59 | ± 0.24 | 37.33 | - | - | **x** |
| 14 | -730,760.46 | ± 0.01 | -730,760.45 | ± 0.02 | 730,760.47 | ± 0.02 | 95.38 | - | - | - |
| 15 | -1,521,955.61 | ± 0.06 | -1,521,955.32 | ± 0.13 | 1,521,955.11 | ± 0.26 | 48.66 | - | - | **x** |
| 16 | -1,347,288.12 | ± 0.09 | -1,347,288.52 | ± 0.25 | 1,347,286.86 | ± 0.38 | 32.32 | - | **x** | - |
| 17 | -2,972,461.53 | ± 0.02 | -2,972,461.56 | ± 0.03 | 2,972,461.51 | ± 0.03 | 92.87 | - | - | - |
| 18 | -842,682.64 | ± 0.02 | -842,682.61 | ± 0.04 | 842,682.66 | ± 0.03 | 88.90 | - | - | - |
| 19 | - | - | - | - | - | - | - | - | - | - |

Table 3.3.: NEQ switching results for a switching length of 10 ps. See table 3.2 caption for further details.

| Cmp. | $\Delta G_{\mathrm{NEQ,bi}}$ | $\Delta G^{\mathrm{MM,QML}}_{\mathrm{NEQ,uni}}$ | $\Delta G^{\mathrm{QML,MM}}_{\mathrm{NEQ,uni}}$ | overlap $W$ (%) | $\Delta G^{\mathrm{MM,QML}}_{\mathrm{NEQ,uni}}$ diff. from $\Delta G_{\mathrm{eq}}$ | cum. sd. $W$ (MM,QML) $>2\ k_BT$ | cum. sd. $W$ (QML,MM) $>2\ k_BT$ |
|---|---|---|---|---|---|---|---|
| 1 | -940,543.77 ± 0.02 | -940,545.55 ± 0.18 | 940,543.70 ± 0.03 | 87.63 | **x** | - | - |
| 2 | -2,105,811.44 ± 0.01 | -2,105,811.43 ± 0.02 | 2,105,811.44 ± 0.01 | 97.39 | - | - | - |
| 3 | -656,584.60 ± 0.01 | -656,584.60 ± 0.02 | 656,584.60 ± 0.02 | 95.41 | - | - | - |
| 4 | -839,346.66 ± 0.04 | -839,346.61 ± 0.08 | 839,346.71 ± 0.07 | 71.32 | - | - | - |
| 5 | -560,516.80 ± 0.03 | -560,517.77 ± 0.08 | 560,516.60 ± 0.03 | 82.61 | **x** | - | - |
| 6 | -1,590,089.68 ± 0.02 | -1,590,089.66 ± 0.02 | 1,590,089.67 ± 0.02 | 94.25 | - | - | - |
| 7 | -1,407,890.65 ± 0.10 | -1,407,890.58 ± 0.23 | 1,407,890.36 ± 0.60 | 27.62 | - | **x** | - |
| 8 | -960,187.64 ± 0.02 | -960,187.64 ± 0.04 | 960,187.67 ± 0.03 | 90.28 | - | - | - |
| 9 | -939,376.89 ± 0.01 | -939,376.90 ± 0.02 | 939,376.86 ± 0.02 | 96.24 | - | - | - |
| 10 | -701,793.55 ± 0.03 | -701,793.56 ± 0.03 | 701,793.55 ± 0.05 | 85.71 | - | - | - |
| 11 | -1,167,290.28 ± 0.04 | -1,167,290.34 ± 0.08 | 1,167,290.28 ± 0.07 | 71.78 | - | - | - |
| 12 | -761,146.91 ± 0.01 | -761,146.92 ± 0.02 | 761,146.90 ± 0.02 | 96.10 | - | - | - |
| 13 | -604,674.28 ± 0.05 | -604,674.09 ± 0.12 | 604,674.28 ± 0.12 | 56.92 | - | - | **-** |
| 14 | -730,760.47 ± 0.01 | -730,760.48 ± 0.02 | 730,760.48 ± 0.01 | 97.51 | - | - | - |
| 15 | -1,521,955.51 ± 0.06 | -1,521,955.16 ± 0.09 | 1,521,955.40 ± 0.13 | 55.70 | - | - | **x** |
| 16 | -1,347,288.23 ± 0.07 | -1,347,288.08 ± 0.14 | 1,347,285.85 ± 0.88 | 44.36 | - | **x** | - |
| 17 | -2,972,461.53 ± 0.01 | -2,972,461.54 ± 0.02 | 2,972,461.53 ± 0.02 | 96.04 | - | - | - |
| 18 | -842,682.60 ± 0.02 | -842,682.60 ± 0.03 | 842,682.59 ± 0.03 | 93.22 | - | - | - |
| 19 | - | - | - | - | - | - | - |

Table 3.4.: NEQ switching results for a switching length of 20 ps. See table 3.2 caption for further details.

| Cmp. | $\Delta G_{\mathrm{NEQ,bi}}$ | $\Delta G^{\mathrm{MM,QML}}_{\mathrm{NEQ,uni}}$ | $\Delta G^{\mathrm{QML,MM}}_{\mathrm{NEQ,uni}}$ | overlap $W$ (%) | $\Delta G^{\mathrm{MM,QML}}_{\mathrm{NEQ,uni}}$ diff. from $\Delta G_{\mathrm{eq}}$ | cum. sd. $W$ (MM,QML) $>2\ k_BT$ | cum. sd. $W$ (QML,MM) $>2\ k_BT$ |
|---|---|---|---|---|---|---|---|
| 1 | -940,543.80 ± 0.02 | -940,545.55 ± 0.17 | 940,543.73 ± 0.02 | 88.29 | **x** | - | - |
| 2 | -2,105,811.44 ± 0.01 | -2,105,811.43 ± 0.01 | 2,105,811.44 ± 0.01 | 98.58 | - | - | - |
| 3 | -656,584.60 ± 0.01 | -656,584.59 ± 0.01 | 656,584.60 ± 0.01 | 97.87 | - | - | - |
| 4 | -839,346.68 ± 0.03 | -839,346.73 ± 0.07 | 839,346.76 ± 0.05 | 79.15 | - | - | - |
| 5 | -560,516.78 ± 0.02 | -560,517.81 ± 0.07 | 560,516.60 ± 0.02 | 85.21 | **x** | - | - |
| 6 | -1,590,089.68 ± 0.01 | -1,590,092.99 ± 0.96 | 1,590,089.68 ± 0.02 | 96.23 | **x** | - | - |
| 7 | -1,407,890.87 ± 0.07 | -1,407,891.19 ± 0.18 | 1,407,891.36 ± 0.18 | 42.79 | - | **x** | - |
| 8 | -960,187.65 ± 0.02 | -960,187.63 ± 0.02 | 960,187.68 ± 0.02 | 94.54 | - | - | - |
| 9 | -939,376.86 ± 0.01 | -939,376.85 ± 0.01 | 939,376.86 ± 0.01 | 98.06 | - | - | - |
| 10 | -701,793.54 ± 0.02 | -701,793.60 ± 0.03 | 701,793.46 ± 0.04 | 91.62 | - | - | - |
| 11 | -1,167,290.18 ± 0.03 | -1,167,290.32 ± 0.06 | 1,167,290.09 ± 0.05 | 80.43 | - | - | - |
| 12 | -761,146.91 ± 0.01 | -761,146.93 ± 0.01 | 761,146.90 ± 0.01 | 97.64 | - | - | - |
| 13 | -604,674.24 ± 0.04 | -604,674.24 ± 0.07 | 604,674.25 ± 0.06 | 72.85 | - | - | - |
| 14 | -730,760.48 ± 0.01 | -730,760.48 ± 0.01 | 730,760.48 ± 0.01 | 98.74 | - | - | - |
| 15 | -1,521,955.49 ± 0.05 | -1,521,955.38 ± 0.11 | 1,521,955.23 ± 0.12 | 63.40 | - | - | - |
| 16 | -1,347,288.18 ± 0.06 | -1,347,288.32 ± 0.13 | 1,347,287.47 ± 0.27 | 52.72 | - | **x** | - |
| 17 | -2,972,461.55 ± 0.01 | -2,972,461.55 ± 0.01 | 2,972,461.56 ± 0.01 | 97.83 | - | - | - |
| 18 | -842,682.61 ± 0.01 | -842,682.62 ± 0.02 | 842,682.61 ± 0.02 | 96.47 | - | - | - |
| 19 | - | - | - | - | - | - | - |

Table 3.5.: NEQ switching results for a switching length of 50 ps. See table 3.2 caption for further details.

| Cmp. | $\Delta G_{\mathrm{NEQ,bi}}$ | $\Delta G^{\mathrm{MM,QML}}_{\mathrm{NEQ,uni}}$ | $\Delta G^{\mathrm{QML,MM}}_{\mathrm{NEQ,uni}}$ | overlap $W$ (%) | $\Delta G^{\mathrm{MM,QML}}_{\mathrm{NEQ,uni}}$ diff. from $\Delta G_{\mathrm{eq}}$ | cum. sd. $W$ (MM,QML) $>2\ k_BT$ | cum. sd. $W$ (QML,MM) $>2\ k_BT$ |
|---|---|---|---|---|---|---|---|
| 1 | -940,543.77 ± 0.02 | -940,545.57 ± 0.16 | 940,543.69 ± 0.02 | 90.24 | **x** | - | - |
| 2 | -2,105,811.43 ± 0.00 | -2,105,811.43 ± 0.01 | 2,105,811.42 ± 0.01 | 99.46 | - | - | - |
| 3 | -656,584.61 ± 0.01 | -656,584.62 ± 0.01 | 656,584.61 ± 0.01 | 99.07 | - | - | - |
| 4 | -839,346.68 ± 0.02 | -839,346.81 ± 0.16 | 839,346.73 ± 0.03 | 87.14 | - | - | - |
| 5 | -560,516.70 ± 0.02 | -560,517.61 ± 0.07 | 560,516.54 ± 0.02 | 88.83 | **x** | - | - |
| 6 | -1,590,089.69 ± 0.01 | -1,590,089.66 ± 0.01 | 1,590,089.69 ± 0.01 | 97.18 | **-** | - | - |
| 7 | -1,407,890.76 ± 0.06 | -1,407,890.85 ± 0.10 | 1,407,890.63 ± 0.20 | 56.79 | - | **-** | - |
| 8 | -960,187.65 ± 0.01 | -960,187.66 ± 0.02 | 960,187.65 ± 0.01 | 97.56 | - | - | - |
| 9 | -939,376.85 ± 0.01 | -939,376.86 ± 0.01 | 939,376.85 ± 0.01 | 99.17 | - | - | - |
| 10 | -701,793.59 ± 0.01 | -701,793.62 ± 0.02 | 701,793.55 ± 0.02 | 96.25 | - | - | - |
| 11 | -1,167,290.20 ± 0.02 | -1,167,290.30 ± 0.05 | 1,167,290.18 ± 0.03 | 88.45 | - | - | - |
| 12 | -761,146.91 ± 0.01 | -761,146.91 ± 0.01 | 761,146.90 ± 0.01 | 98.85 | - | - | - |
| 13 | -604,674.30 ± 0.02 | -604,674.28 ± 0.04 | 604,674.28 ± 0.04 | 86.88 | - | - | **-** |
| 14 | -730,760.47 ± 0.00 | -730,760.48 ± 0.01 | 730,760.47 ± 0.01 | 99.45 | - | - | - |
| 15 | -1,521,955.41 ± 0.04 | -1,521,955.35 ± 0.11 | 1,521,955.40 ± 0.08 | 69.94 | - | **-** | **-** |
| 16 | -1,347,288.90 ± 0.05 | -1,347,296.88 ± 0.61 | 1,347,287.33 ± 0.42 | 56.27 | **x** | **x** | - |
| 17 | -2,972,461.55 ± 0.01 | -2,972,461.55 ± 0.01 | 2,972,461.55 ± 0.01 | 99.15 | - | - | - |
| 18 | -842,682.63 ± 0.01 | -842,682.64 ± 0.01 | 842,682.61 ± 0.01 | 98.58 | - | - | - |
| 19 | - | - | - | - | - | - | - |

## 3.3. Free energy perturbation

Table 3.6 shows the results obtained with FEP. Bidirectional free energy estimates are denoted as $\Delta G_{\mathrm{FEP,bi}}$ and unidirectional estimates as $\Delta G^{\mathrm{MM,QML}}_{\mathrm{FEP,uni}}$ and $\Delta G^{\mathrm{QML,MM}}_{\mathrm{FEP,uni}}$ (depending on the target distribution). Free energy estimates are reported as the average of three independent repetitions (for details see the Methods section).

FEP results deviate for more systems from the equilibrium free energy estimate than in the case of NEQ simulations. Notably, in contrast to the NEQ results, in some cases also the bidirectional protocol results in free energy estimates that clearly deviate from the equilibrium result.

Five compounds were identified for which the bidirectional FEP protocol deviates by more than 1 $k_BT$ from the equilibrium free energy estimate (compounds **4**, **7**, **10**, **11**, and **16**). Free energy estimates obtained with the unidirectional FEP protocol were not considered in table 3.6 due to the large number of inaccurate results.

The overlap between the energy difference distributions shown in table 3.6 indicate a link between the percentage of overlap and the accuracy of the free energy estimate. Low overlap ($\leq 0.02\%$) for the $\Delta$E distributions were obtained for compounds **4**, **7**, **10**, **11**, and **16**, i.e., the same molecules for which also the free energy estimate deviates by more than 1 $k_BT$ from $\Delta G_{\mathrm{eq}}$.

In most cases, the overlap between the forward and backward energy distribution is quite low. It is surprising that the BAR estimator is able to obtain an accurate result even with energy distributions having less than 1% overlap; yet, the results indicate exactly that. In general, energy distributions are much broader than the respective work distributions, and for each system the FEP cumulative standard deviation converged to a value of $\approx 5\,k_BT$ or higher. Boresch and Woodcock [4] noted that $4\,k_BT$ should be considered an upper threshold for reliable results, which would indicate that for many of these systems the resulting free energy estimates should be considered uncertain.

Table 3.6.: FEP results. Bidirectional results $\Delta G_{\mathrm{FEP,bi}}$ were derived from the BAR estimator; the unidirectional ones $\Delta G_{\mathrm{FEP,uni}}^{\mathrm{MM,QML}}$ and $\Delta G_{\mathrm{FEP,uni}}^{\mathrm{QML,MM}}$ with the exponential averaging method. Overlap between forward and backward energy distributions is given in percent. If the free energy estimate $\Delta G_{\mathrm{FEP,bi}}$ deviates more than 1 $k_B T$ from the equilibrium value $\Delta G_{\mathrm{eq}}$, the corresponding molecule is marked with an x. Free energy values and the overlap are given as the mean of three FEP protocols.

| Cmp. | $\Delta G_{\mathrm{FEP,bi}}$ | $\Delta G_{\mathrm{FEP,uni}}^{\mathrm{MM,QML}}$ | $\Delta G_{\mathrm{FEP,uni}}^{\mathrm{QML,MM}}$ | overlap $\Delta E$ (%) | $\Delta G_{\mathrm{FEP,bi}}$ diff. from $\Delta G_{\mathrm{eq}}$ |
|---|---|---|---|---|---|
| 1 | -940,544.16 ± 0.13 | -940,543.41 ± 0.44 | 940,546.73 ± 0.59 | 2.09 | **-** |
| 2 | -2,105,811.45 ± 0.22 | -2,105,805.87 ± 0.13 | 2,105,815.59 ± 1.38 | 0.43 | - |
| 3 | -656,584.72 ± 0.60 | -656,576.65 ± 0.80 | 656,592.91 ± 0.77 | 0.06 | - |
| 4 | -839,345.15 ± 1.19 | -839,325.12 ± 1.90 | 839,365.18 ± 0.48 | 0.00 | **x** |
| 5 | -560,517.00 ± 0.36 | -560,515.23 ± 2.63 | 560,521.85 ± 0.27 | 0.42 | **-** |
| 6 | -1,590,089.83 ± 0.13 | -1,590,089.68 ± 0.69 | 1,590,091.35 ± 0.20 | 2.97 | **-** |
| 7 | -1,407,892.15 ± 0.94 | -1,407,876.80 ± 0.94 | 1,407,907.51 ± 1.17 | 0.00 | **x** |
| 8 | -960,187.89 ± 0.26 | -960,179.89 ± 0.28 | 960,195.91 ± 0.14 | 0.05 | - |
| 9 | -939,376.62 ± 0.17 | -939,370.80 ± 0.78 | 939,381.15 ± 0.93 | 0.24 | - |
| 10 | -701,791.89 ± 2.03 | -701,778.14 ± 0.96 | 701,805.63 ± 4.22 | 0.00 | **x** |
| 11 | -1,167,292.37 ± 0.83 | -1,167,283.25 ± 0.60 | 1,167,301.73 ± 1.59 | 0.02 | **x** |
| 12 | -761,147.16 ± 0.14 | -761,144.04 ± 0.75 | 761,149.85 ± 0.12 | 0.93 | - |
| 13 | -604,674.27 ± 0.78 | -604,668.83 ± 0.88 | 604,678.52 ± 0.30 | 0.16 | - |
| 14 | -730,760.25 ± 0.24 | -730,757.48 ± 1.74 | 730,763.65 ± 1.85 | 0.44 | - |
| 15 | -1,521,955.87 ± 0.71 | -1,521,949.56 ± 1.37 | 1,521,960.70 ± 1.98 | 0.09 | - |
| 16 | -1,347,290.16 ± 4.37 | -1,347,278.12 ± 4.25 | 1,347,302.17 ± 5.72 | 0.01 | **x** |
| 17 | -2,972,461.16 ± 1.14 | -2,972,453.59 ± 1.51 | 2,972,467.96 ± 2.90 | 0.07 | - |
| 18 | -842,682.69 ± 0.19 | -842,685.36 ± 4.93 | 842,683.63 ± 1.40 | 1.54 | - |
| 19 | - | - | - | - | - |

## 3.4. Example: molecule 14 — ZINC0867000

After showing the free energy results for the different protocols I want to discuss one example molecule in depth. For this task molecule **14** (ZINC0867000) was chosen as a toy example due to its simple chemical structure.

### 3.4.1. Free energy estimates

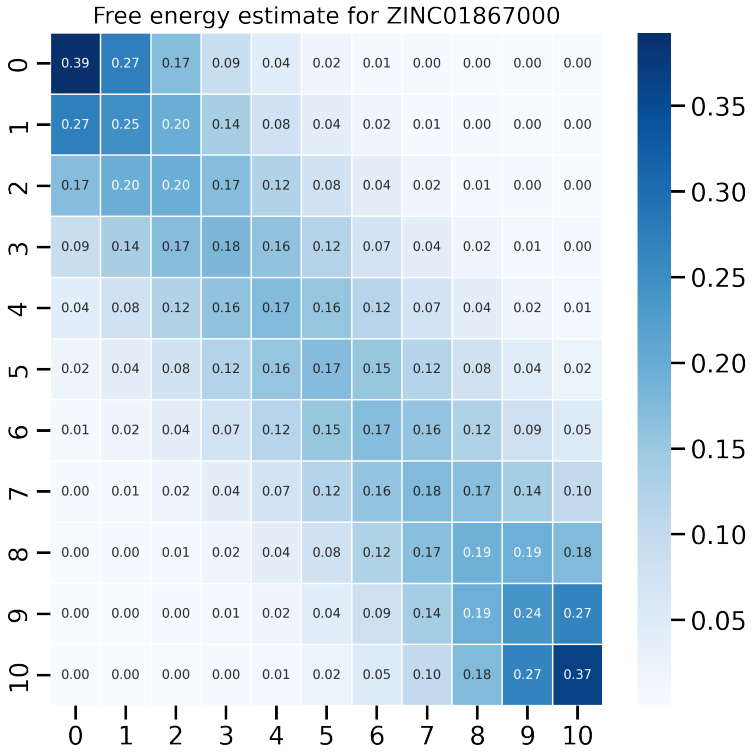


Figure 3.1.: Overlap plot for molecule **14** (ZINC0867000). The rows and columns correspond to the 11 $\lambda$ states ranging from 0 (MM endstate) to 10 (QML endstate). The individual elements of the matrix $O_{i,j}$ denote the probabilities of observing a sample from state i in state j. The sum of each row and column is normalized to one.

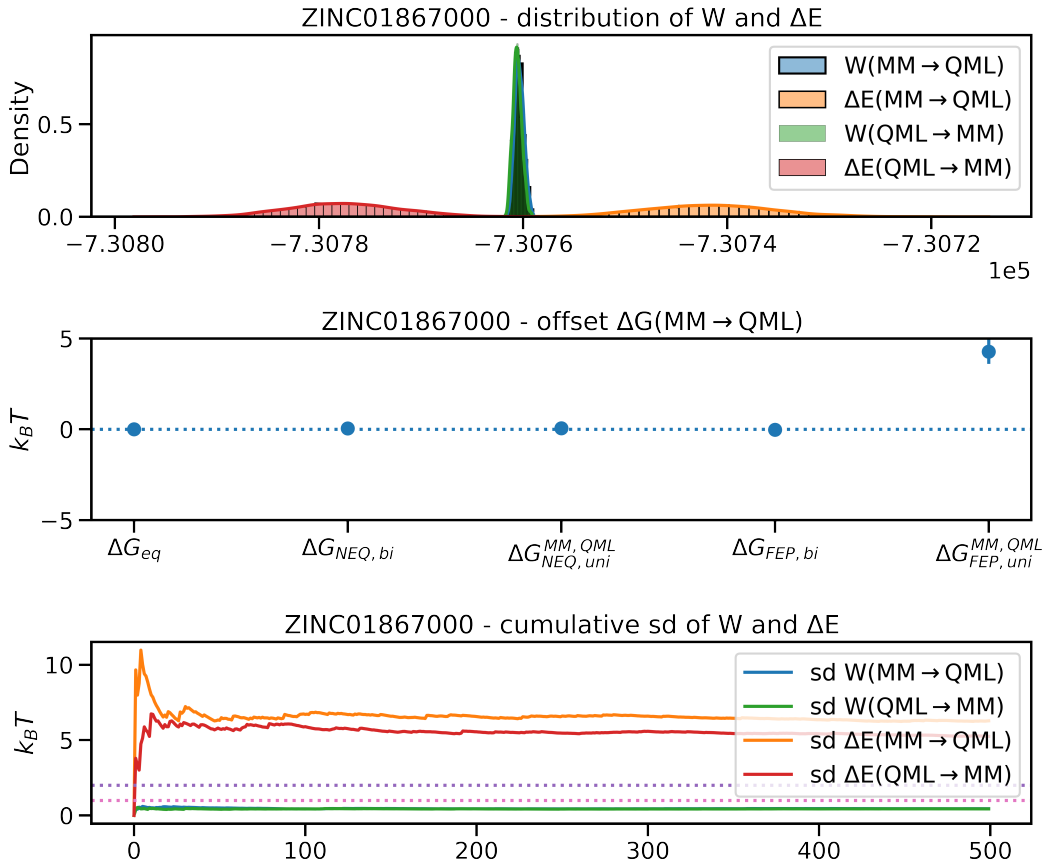


Figure 3.2.: 5 ps NEQ switching and FEP results for molecule **14**. (Top) Density plot of work and energy distributions with frequency on the ordinate and energy values on the x-axis. The MM → QML NEQ work distribution, where initial coordinates were drawn from the MM equilibrium ensemble and switching was performed to the QML endstate potential, is shown in blue. The histogram for the reverse QML → MM switching experiment is shown in green. The same pattern applies to energy distributions derived from FEP experiments, which are depicted in orange (MM → QML) and red (QML → MM), respectively. Both NEQ work distributions are a collection of 500 switching experiments in each direction. FEP experiments were repeated 5,000 times in each direction. (Middle) Comparison of NEQ and FEP free energy values to the equilibrium reference (in $k_BT$). Equilibrium and FEP values are collected from the first (out of three) repeats. NEQ results are the outcome of the 5 ps protocol. The labels are the same as in table 3.2. (Bottom) Cumulative standard deviation (in $k_BT$) of work and energy distributions recorded over 500 switches. The color code is identical to the one used in the top plot. The 1 and 2 $k_BT$ thresholds are shown as pink and purple dotted lines, respectively.

The equilibrium free energy estimate $\Delta G_{\text{eq}}$ (MM $\rightarrow$ QML) for molecule **14** was determined as -730,760.50 $k_BT$ with a standard deviation of 0.09. Since the accuracy of the estimate depends on the overlap between the conformational space of consecutive $\lambda$ states the overlap was monitored by computing the overlap matrix $O$ (an example is shown as a heatmap in figure 3.1). For molecule **14** the overlap plot indicates that 11 $\lambda$ states are sufficient to obtain a free energy estimate with low variance. The $\lambda$ states 4, 5, and 6 even have overlap with both the MM and the QML endstates. Moreover, neighboring states exhibit an overlap of far more than 0.03, which is suggested as the lower limit [25].

Next, the results obtained from bi- and unidirectional NEQ switching and FEP protocols will be discussed. These are summarized in the three plots shown in figure 3.2. The top plot shows the work and energy difference distributions. The distribution shown in blue represents the outcome of the MM $\rightarrow$ QML 5 ps NEQ work protocol, while the distribution shown in green represents the reverse protocol. Analogously, the energy difference distributions derived from FEP experiments are shown in orange and red, respectively. It is not hard to see that work and energy distributions vary in shape and position. While the work distributions are narrow and efficiently overlapping, the energy ones do not seem to have significant overlapping support. The difference in width (i.e. variance) between the work and energy distributions is a result of the different protocols: NEQ protocols slowly drive the system out of equilibrium and allow the molecule to adapt to the changes in interactions. By contrast, instantaneous switches do not allow the molecule to adapt to the new potential and, therefore, result in a much broader distribution of the energy differences. Consequently, the results are typically harder to converge and may deviate significantly from the correct free energy value.

The bottom plot is closely related to the top plot and shows the cumulative standard deviation for each of the switching experiments. The cumulative standard deviation gives insight about the convergence and indicates whether the resulting free energy estimates are trustworthy. Two additional lines provide arbitrary but useful limits: The pink dotted line represents the 1 $k_BT$ threshold, while the purple dotted line indicates the upper safe limit of 2 $k_BT$. The blue and green line describing the cumulative work standard deviation converge to an average standard deviation of around 0.5 $k_BT$. Convergence to this value is reached already after a small number of switches, indicating low variance in the free energy estimate. The standard deviations of the energy distributions (shown in orange and red), by contrast, converge to an average value of $\approx 5\,k_BT$. The resulting free energy estimates, thus, are not expected to have high accuracy.

The middle graph shows the deviations of the 5 ps NEQ and FEP protocols from the "gold standard", the free energy difference $\Delta G_{\text{eq}}$ using the alchemical multi-state equilibrium method. The reference result is plotted as zero, offset by 730,760.50 $k_BT$. Comparing the free energy estimates obtained with the other methods, we can make three observations.

(1) The NEQ results agree in general with the equilibrium value. Not only does the

bidirectional method deliver a satisfying result ($\Delta G_{\mathrm{NEQ,bi}}$: -730,760.46 ± 0.01 $k_BT$), also the unidirectional switching protocol is sufficient to estimate an accurate free energy difference ($\Delta G^{\mathrm{MM,QML}}_{\mathrm{NEQ,uni}}$: -730,760.45 ± 0.02 $k_BT$). Ideally, the forward and reverse NEQ switching protocols should yield the same free energy estimate (with opposite sign). As can be seen in tables 3.2–3.5, this is the case here.

(2) The bidirectional FEP result approaches the equilibrium free energy estimate closely. The average of the free energy estimate of three independently performed FEP protocols $\Delta G_{\mathrm{FEP,bi}}$ was calculated as -730,760.25 ± 0.24 $k_BT$. The value depicted in figure 3.2 is the outcome of the first FEP protocol ($\Delta G_{\mathrm{FEP,bi}}$: -730,760.53 ± 0.31 $k_BT$). Visually (top plot), it is not possible to discern overlap between the energy distributions. But according to table 3.2, in which the $\Delta E$ average overlap is listed, the overlap is 0.44%.

(3) The unidirectional FEP estimate deviates from the equilibrium free energy difference by almost 5 $k_BT$ with $\Delta G^{\mathrm{MM,QML}}_{\mathrm{FEP,uni}}$: -730,756.23 ± 0.67 $k_BT$, c.f. Table 4 in the Supporting Materials). Averaged over three repetitions $\Delta G^{\mathrm{MM,QML}}_{\mathrm{FEP,uni}}$ yields a value of -730,757.48 ± 1.74 $k_BT$. As already mentioned in section 1.1.3, exponential averaging is very sensitive to outliers (or, in general, broad distributions), which may be responsible for the observed deviation.

### 3.4.2. Torsion profiles

For every molecule studied, torsion profiles were generated from equilibrium simulations at the MM and QML endstates, and from the last frame of the 5, 10, 20, and 50 ps NEQ switching trajectories. A maximum in the density plot for a torsion represents a minimum in the potential energy. Figure 3.3 shows the results obtained for molecule **14** (5 ps switch), which is depicted along with the corresponding atom indices. In this symmetric molecule, two rotatable bonds can be identified between the aromatic moieties, respectively. We analyse the rotation around these bonds using the two torsion angles [6,11,12,17] and [2,3,6,11].

Figure 3.3 shows the torsion distributions of the four sample distributions (MM, QML equilibrium samples; samples obtained from NEQ switching from MM to QML and *vice versa*). The dihedral distributions of molecule **14** for torsion angle [6,11,12,17] show two energetic minima (maxima in the torsion profile) for both the MM and QML levels of theory at $\approx \frac{\pi}{4}$ and $\frac{3\pi}{4}$. A similar torsion population can be seen for torsion angle [2,3,6,11], but shifted by $\pi$. It appears as if the MM force field and the QML potential describe this torsion profile similarly.

Given the initial close resemblance, the distribution obtained from the 5 ps switch does not deliver any apparently different results. Nevertheless, subtle differences can be observed. In general, one would expect that the initial torsion profile gets transformed to the target distribution, as soon as the switching procedure has been carried out. In other words, if samples from the MM equilibrium ensemble are used to initialize the protocol, and the potential energy is gradually switched to the

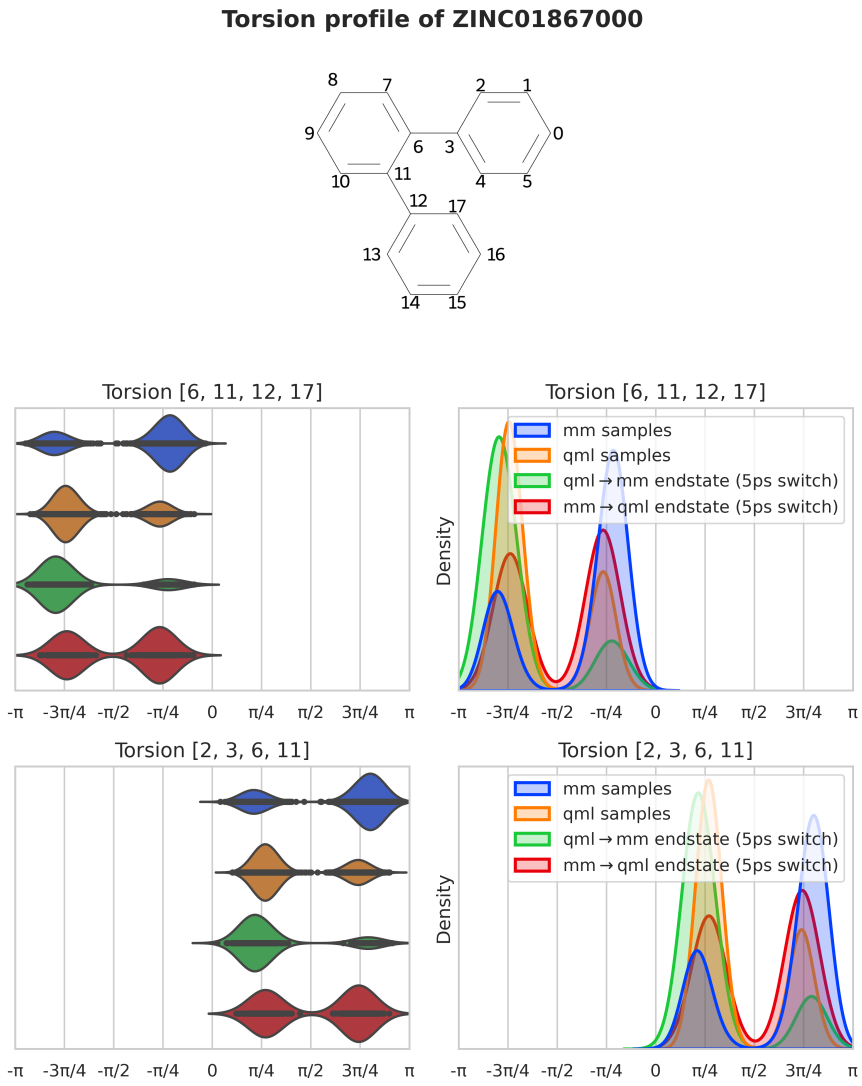


Figure 3.3.: Torsion profiles for two dihedral angles of molecule **14** are shown. The upper panel shows results for torsion [6,11,12,17] and the bottom panel shows results for torsion [2,3,6,11]. The plots on the lefthand side are violin plots of the respective torsion angle. Black ticks correspond to individual samples. The plots on the righthand side are the correponding kernel density estimates. Torsion angles are shown in radiant ranging from $-\pi$ to $+\pi$. Blue and orange data correspond to torsion angles observed during three independent equilibrium simulations ($3 \times 4,000$ samples, respectively, for details see the Methods section) with the MM potential (blue) and the QML potential (orange). Torsion angles plotted in green and red indicate final conformations of the switching trajectories, either from MM to QML (red) or QML to MM (green) potential (500 samples, respectively).

QML description, the energetic minima of the resulting distributions are expected to have a similar location on the angular spectrum as the initial MM torsion profile. Indeed, it can be noted that when switching from MM to QML, both energetic minima have shifted to the position predicted by the QML description. The same trend can be observed when inspecting the switch in the opposite direction.

## 3.5. Effect of longer switching lengths on free energy estimates

The switching length is a property of significant interest for NEQ switching protocols. The question that I wanted to address in this work was if longer switching lengths can improve free energy predictions. I analyze this question using free energy estimates, the corresponding work distributions and torsion profiles as a surrogate observable that can be easily plotted in 1D.

### 3.5.1. Effect on the variance of the work distributions

Intuitively, one expects that longer switching protocols should decrease the variance of the work distributions. From table 3.2 it can be seen that molecules **4**, **7**, **13**, **15**, and **16** have relatively broad work distributions (cumulative standard deviation $> 2\ k_BT$) for a protocol length of 5 ps. From these compounds, molecule **7** (ZINC00123162) was chosen as a representative example to investigate the effect of longer switching lengths.

Figures 3.4–3.7 display NEQ results obtained with switching lengths of 5, 10, 20 and 50 ps for molecule **7** in an analogous way as they were presented earlier for compound **14** (ZINC0867000) in section 3.4. Comparing the plots, one sees that for longer switching protocols the work distributions become indeed narrower. While after the 5 ps switch the cumulative standard deviation of the MM $\rightarrow$ QML work distribution converges to around 5 $k_BT$, this value is reduced to 2 $k_BT$ for the 50 ps protocol. This observation is in accordance with the literature about annealed importance sampling [28]. The trend observed for **7** can, in fact, be observed for several compounds (cf. tables 3.2–3.5 and table 3.7).

In table 3.7 we also list the *range* of the work distributions (i.e. the difference between the highest and lowest work value) for several molecules discussed in this and the following sections as a function of switching length. E.g., for molecule **7**, in line with the observed reduction of the cumulative standard deviation, the range of the W(MM $\rightarrow$ QML) distribution reduces from $\approx 24\ k_BT$ for the 5 ps switching length to $\approx 11\ k_BT$ for the 50 ps protocol. The distributions do not only become narrower, but their overlap increases as well. This can be seen in tables 3.2–3.5: E.g., for molecule **7** the overlap increases from $\approx 20\%$ to $\approx 57\%$ for the longest switching length. Narrower work distributions and increased overlap as the switching length increases can be observed for all molecules, with the exception of

**16** (ZINC04344392). For this molecule, the $W(\mathrm{MM} \rightarrow \mathrm{QML})$ distribution broadens when using the 50 ps protocol; c.f. table 3.5 and figure 128 in the Supporting Materials. Additionally, the work distributions gain a bimodal (or even trimodal) character as the switching length is increased. This is also reflected in the free energy differences: In table 3.5 one sees that $\Delta G^{\mathrm{MM,QML}}_{\mathrm{NEQ,uni}}$ derived from the 50 ps protocol does not match the equilibrium value, which is not the case for the shorter protocols. This aspect will be discussed further in the following section 3.5.2

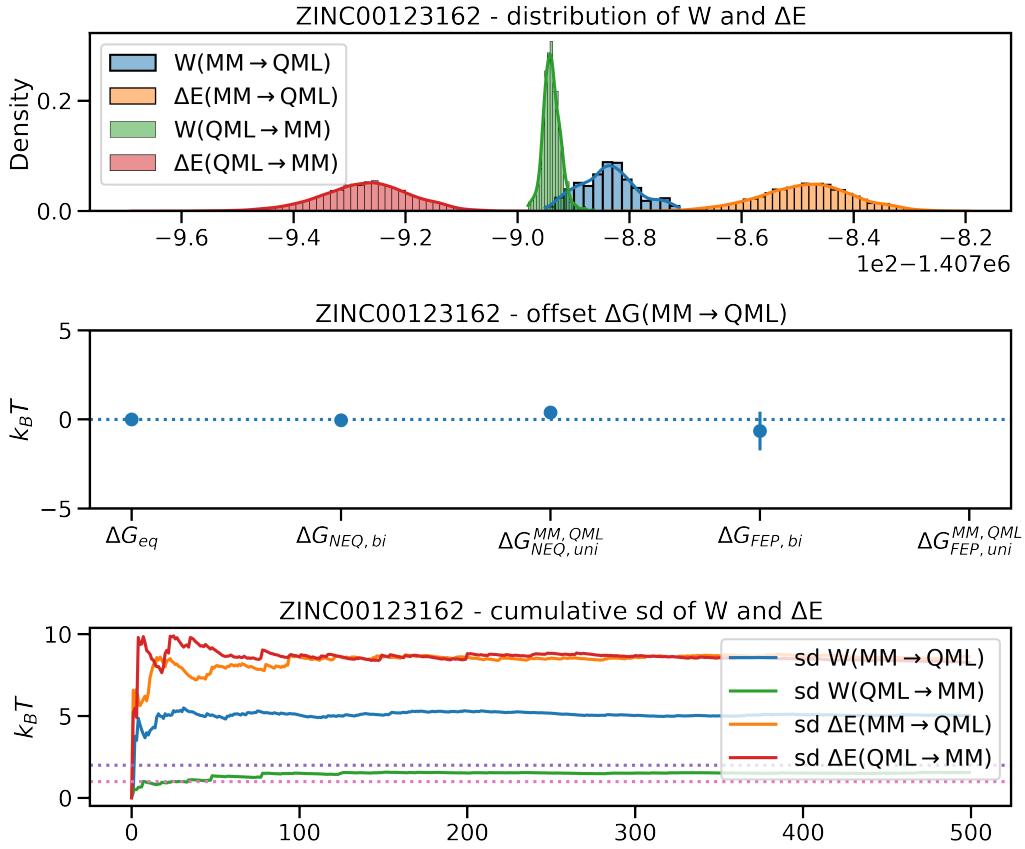


Figure 3.4.: 5 ps NEQ switching and FEP results for molecule **7** (ZINC00123162). (Top) Density plot of work and energy distributions. (Middle) Comparison of NEQ and FEP free energy values to the equilibrium reference from the first protocol run (in $k_BT$). FEP results are calculated each time the plot is generated. Thus, those results vary in the following figures. (Bottom) Cumulative standard deviation (in $k_BT$) of work and energy distributions recorded over 500 switches. See figure 3.2 for additional details.

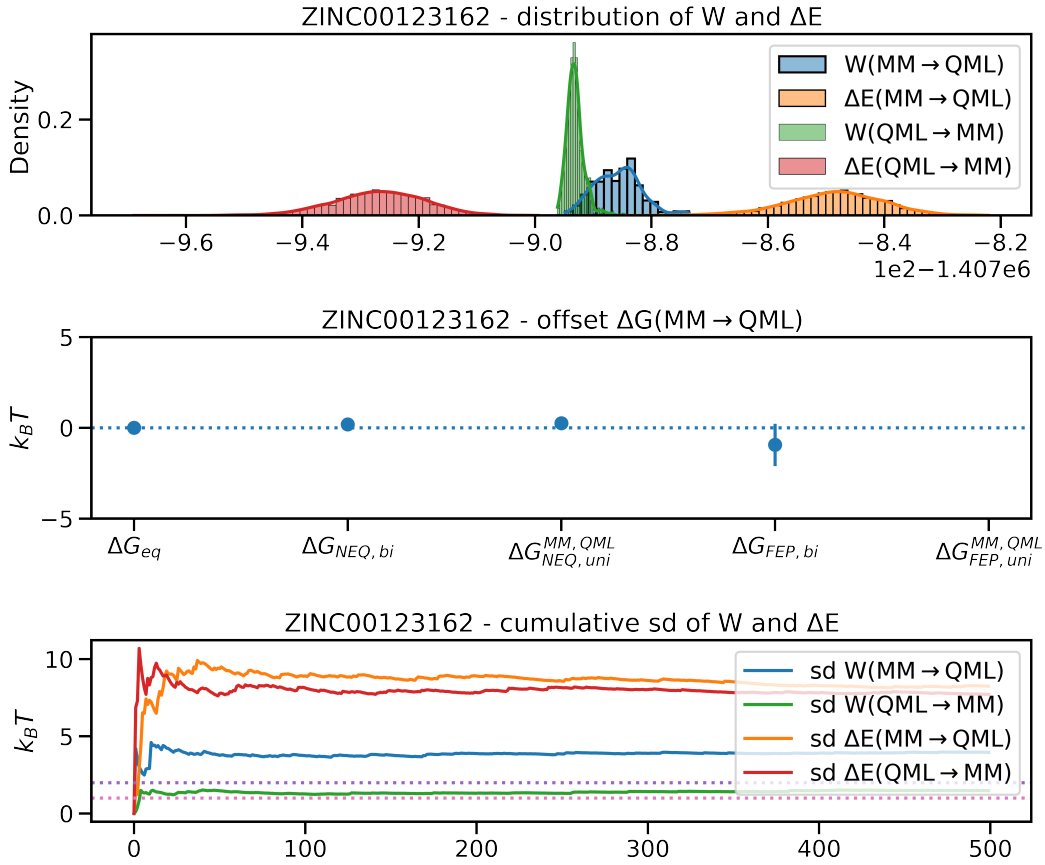


Figure 3.5.: 10 ps NEQ switching and FEP results for molecule **7** (ZINC00123162). See figure 3.2 for additional details.

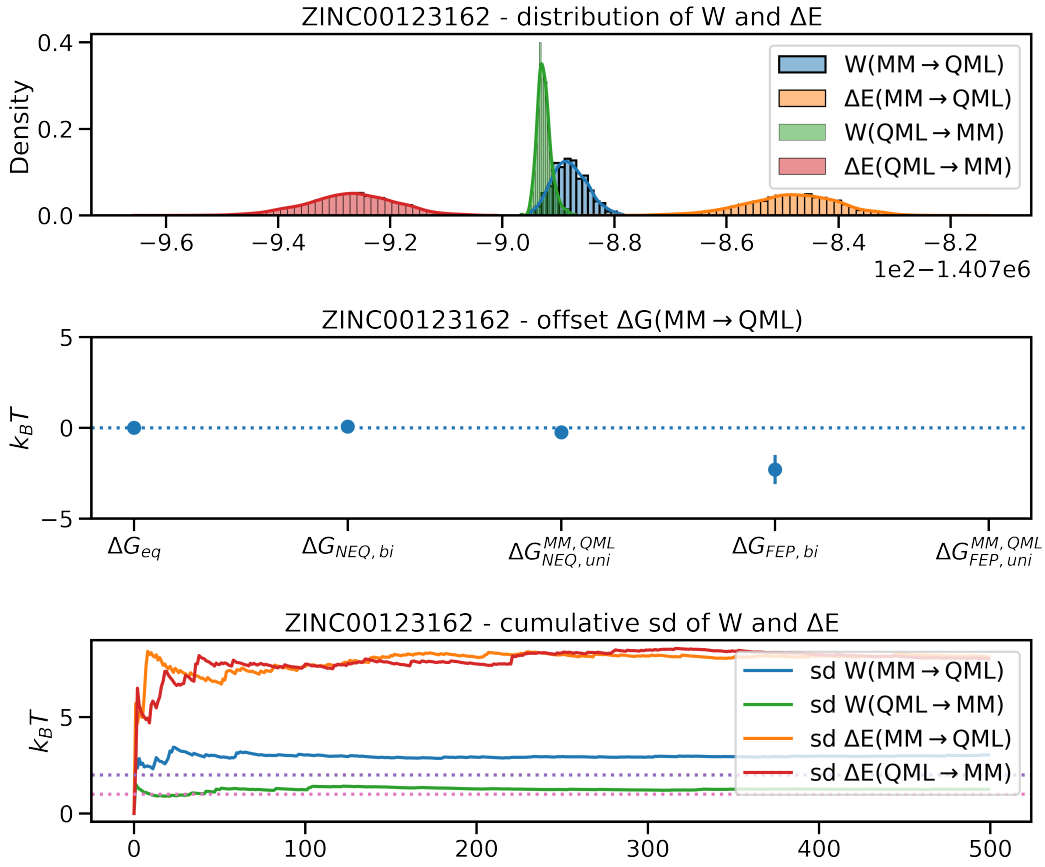


Figure 3.6.: 20 ps NEQ switching and FEP results for molecule **7** (ZINC00123162). See figure 3.2 for additional details.

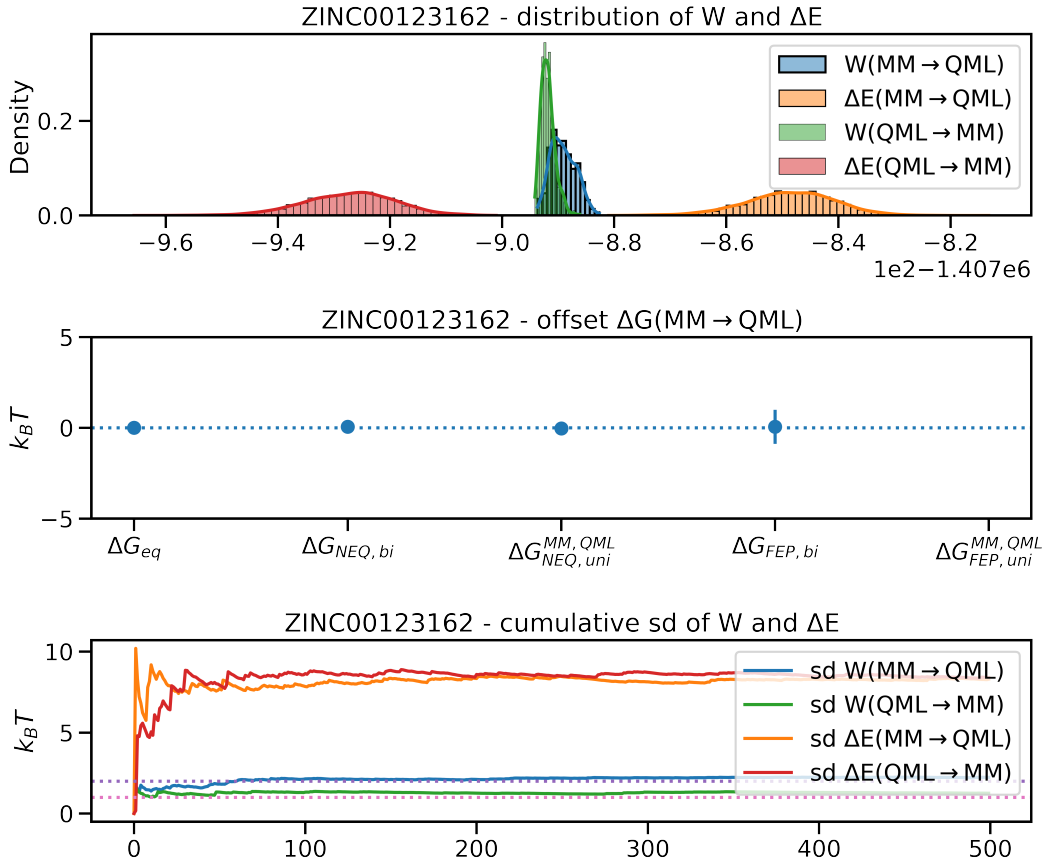


Figure 3.7.: 50 ps NEQ switching and FEP results for molecule **7** (ZINC00123162). See figure 3.2 for additional details.

Table 3.7.: Range of $W$ distributions (in $k_BT$) for different switching lengths $\tau$ for molecules **1**, **5**, **6**, **7**, and **16**. Outliers were only determined for narrow $W$ distributions, which did not exceed a cumulative standard deviation of 2 $k_BT$ (molecules **1**, **5**, **6**). Outliers are defined as values which deviated more than 5 $k_BT$ from the average value and were calculated for the $W$(MM→QML) distribution.

| Cmp. | $\tau$ | range $W$(MM,QML) | range $W$(QML,MM) | outliers $W$(MM,QML) |
|---|---|---|---|---|
| 1 | 5 ps | 8.45 | 7.61 | 4 |
| | 10 ps | 7.05 | 6.93 | 4 |
| | 20 ps | 6.58 | 6.38 | 2 |
| | 50 ps | 5.75 | 5.74 | 0 |
| 5 | 5 ps | 5.91 | 5.75 | 0 |
| | 10 ps | 6.19 | 4.68 | 0 |
| | 20 ps | 4.43 | 4.46 | 0 |
| | 50 ps | 3.70 | 3.67 | 0 |
| 6 | 5 ps | 12.58 | 11.97 | 4 |
| | 10 ps | 2.65 | 11.64 | 0 |
| | 20 ps | 10.28 | 11.06 | 1 |
| | 50 ps | 2.18 | 9.48 | 0 |
| 7 | 5 ps | 23.93 | 11.70 | - |
| | 10 ps | 21.28 | 11.49 | - |
| | 20 ps | 16.57 | 9.59 | - |
| | 50 ps | 11.03 | 7.99 | - |
| 16 | 5 ps | 15.38 | 17.73 | - |
| | 10 ps | 14.10 | 14.49 | - |
| | 20 ps | 13.50 | 11.57 | - |
| | 50 ps | 21.87 | 12.75 | - |

### 3.5.2. Effect on the accuracy of the unidirectional switching protocols

Another factor that was investigated is the influence of the switching length on the free energy estimate. In table 3.2 it can be seen that with a switching length of 5 ps the unidirectional $\Delta G^{\mathrm{MM,QML}}_{\mathrm{NEQ,uni}}$ estimate deviates by more than 1 $k_BT$ from the equilibrium result for molecules **1**, **5** and **6**. When performing longer protocols, however, no obvious improvement can be observed with respect to the free energy estimate. This is highlighted by the results listed in tables 3.2–3.5. Specifically, for molecule **1** (ZINC00077329) we also display this graphically in figures 3.8–3.11. Analogous observations were made for the structurally very similar compound **5** (ZINC00107550); see figures 34-37 in the Supporting Materials. For both systems,

the resulting work distributions (for each switching length) look similar as well.

The results obtained in this work indicate that accuracy of the free energy estimate does not correlate with the width of the underlying work distributions. In table 3.7 it can be seen that the distributions get slightly narrower and the cumulative standard deviation is reduced when applying longer protocols (c.f. tables 3.2–3.5); however, the result does not converge to the reference value. Moreover, for molecule **16**, as already discussed in section 3.5.1, accurate results were obtained in the case of the 5, 10 and 20 ps protocols, although the work distributions are broader than for, e.g., **1**. The 50 ps $W(\mathrm{MM} \rightarrow \mathrm{QML})$ distribution for molecule **16** shows a bimodal (or even trimodal) distribution. It may, thus, be speculated that the inaccurate free energy estimates are a result of the "shape" of the underlying work distributions.

In general, the efficiency of NEQ switching depends on the variance of the normalized importance weights. If NEQ trajectories end up in different modes, however, the situation gets more complicated. The variation in weights due to this will be large if some important modes are found only rarely, and this might not be reflected in the variance. This situation can be observed for molecules **1** and **5** discussed above, for which bimodal work distributions (two isolated modes, cf. figures 3.8–3.11 and figures 34-37 in the Supporting Materials) were obtained for each switching length. For these molecules it was not possible to obtain an accurate unidirectional $\Delta G^{\mathrm{MM,QML}}_{\mathrm{NEQ,uni}}$ estimate, either.

We also inspected whether the work distributions contained outliers, as it was suspected that this might be another factor contributing to an inaccurate free energy estimate. We classify outliers as work values deviating by more than 5 $k_BT$ from the average work value. In table 3.7 outliers can be detected for every switching length except for 50 ps for molecule **1**. For molecule **5**, on the other hand, no outlying work values were identified.

Molecule **6** (ZINC00107778) shows unique behavior with respect to the $\Delta G^{\mathrm{MM,QML}}_{\mathrm{NEQ,uni}}$ estimate. While the 5 and 20 ps switching protocols deliver results which do not match the equilibrium reference, the 10 and 50 ps results seem to converge to the equilibrium free energy value. In this case, the erratic behavior of the free energy results may be caused by the presence of isolated outliers. In the reverse work distribution $W(\mathrm{QML}{\rightarrow}\mathrm{MM})$ one can discern a small "shoulder" towards more negative work values; however, the $W(\mathrm{MM}{\rightarrow}\mathrm{QML})$ distribution (which is used for the $\Delta G^{\mathrm{MM,QML}}_{\mathrm{NEQ,uni}}$ estimation) appears to be unimodal. In general, unimodal distributions (as shown) are not problematic (if they have low variance). In table 3.7 it can be seen that the ranges of the $W(\mathrm{MM}{\rightarrow}\mathrm{QML})$ distributions obtained with the 5 and 20 ps switching protocols are distinctively wider than the ones derived from the 10 and 50 ps experiments; yet, in all cases the cumulative standard deviations remain below the acceptable threshold (c.f. tables 3.2- 3.5). Nevertheless, looking at the plots of the cumulative standard deviations in figures 3.12–3.15, there are signs of outliers. While in the respective bottom plots the blue curve appears to be

a straight line in the 10 and 50 ps cases, the 5 and 20 ps plots show single "steps", where the cumulative standard deviation rises abruptly.

To investigate this further, the $W$(MM→QML) distributions were examined for outliers. Indeed, in table 3.7 one sees that for switching lengths of 5 and 20 ps there are outliers (four and one outliers, respectively), deviating by more than 10 $k_BT$ from the mean value. If those outliers are removed, the free energy estimate approaches the equilibrium free energy value. For the 10 and 50 ps protocols, no such extreme work values were observed (cf. table 3.7). This observation exemplifies again how isolated work values can dominate unidirectional methods.

Thus, one can conclude the following: The accuracy of a unidirectional free energy estimate does not necessarily depend on the variance of the underlying unimodal work distribution. We have seen that the switching length has little effect on the accuracy of unidirectional free energy estimates and the discrepancies found are more likely caused by the shape of the underlying work distributions and/or by the presence of outliers (especially isolated outliers in unimodal distributions). As longer switching lengths do not improve unidirectional results, it may be better to calculate free energy corrections in such "problematic" cases with bidirectional NEQ methods, despite the higher computational cost. As already mentioned in section 3.1 the bidirectional NEQ results are always in accord with the equilibrium value. Longer protocols showed no improvement.

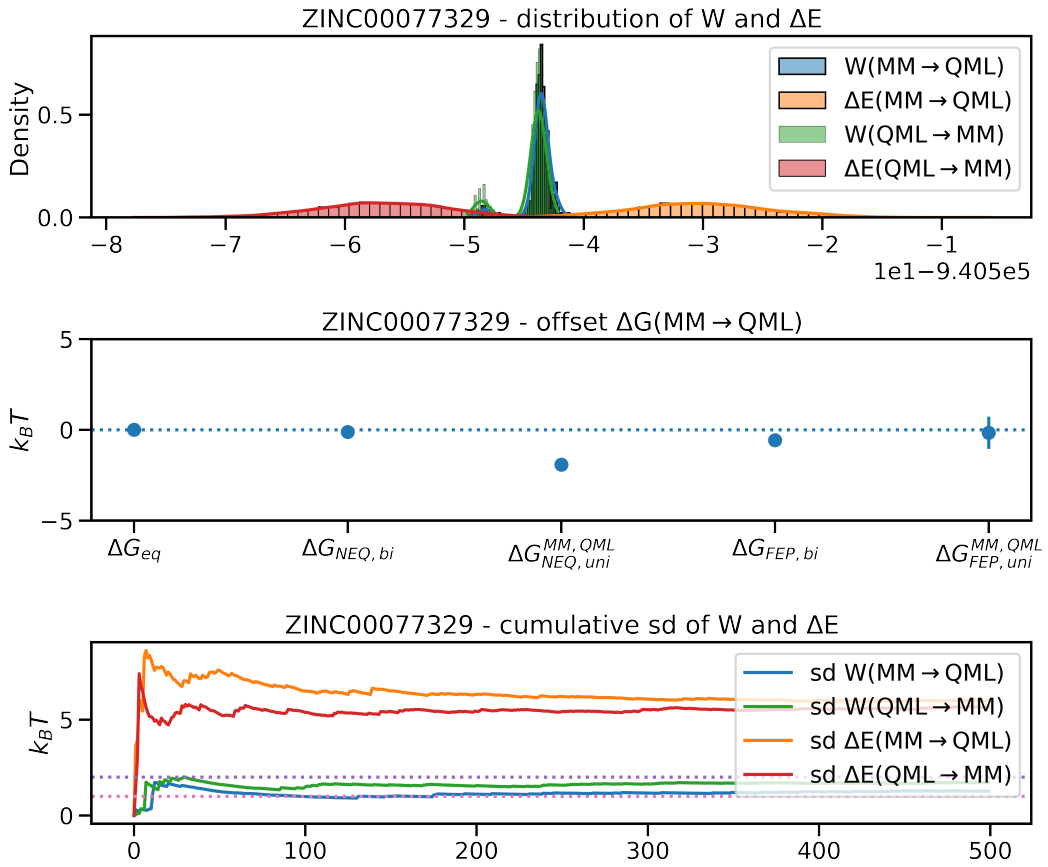


Figure 3.8.: 5 ps NEQ switching and FEP results for molecule **1** (ZINC00077329). (Top) Density plot of work and energy distributions. (Middle) Comparison of NEQ and FEP free energy values to the equilibrium reference from the first protocol run (in $k_BT$). FEP results are calculated each time the plot is generated. Thus, those results vary in the following figures. (Bottom) Cumulative standard deviation (in $k_BT$) of work and energy distributions recorded over 500 switches. For a detailed description of these plots see figure 3.2.

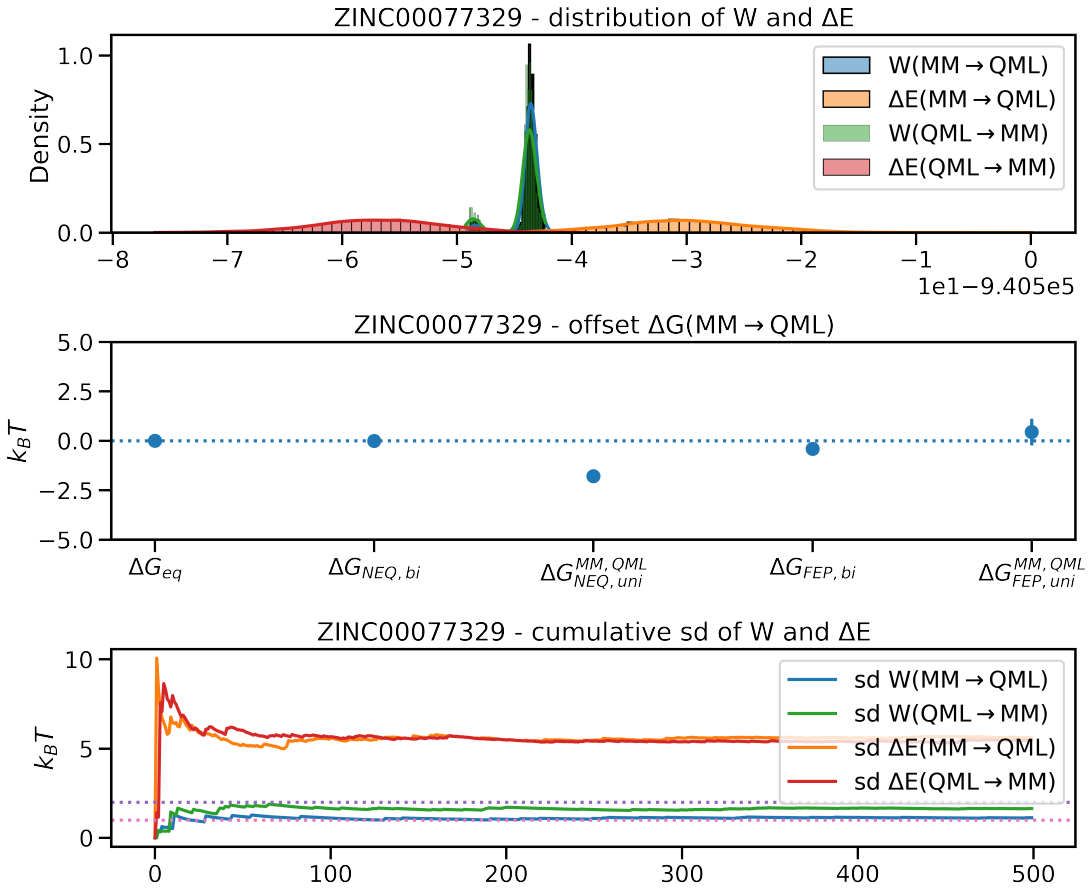


Figure 3.9.: 10 ps NEQ switching and FEP results for molecule **1** (ZINC00077329). See figure 3.2 for additional details.

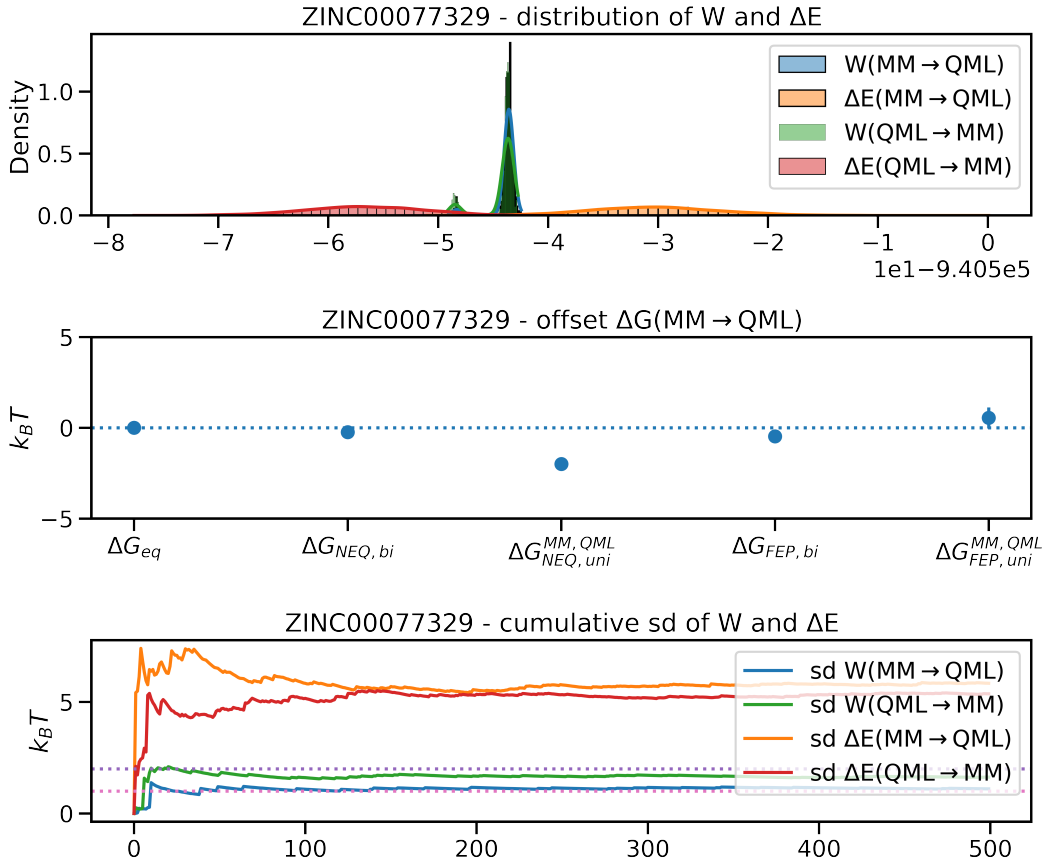


Figure 3.10.: 20 ps NEQ switching and FEP results for molecule **1** (ZINC00077329). See figure 3.2 for additional details.

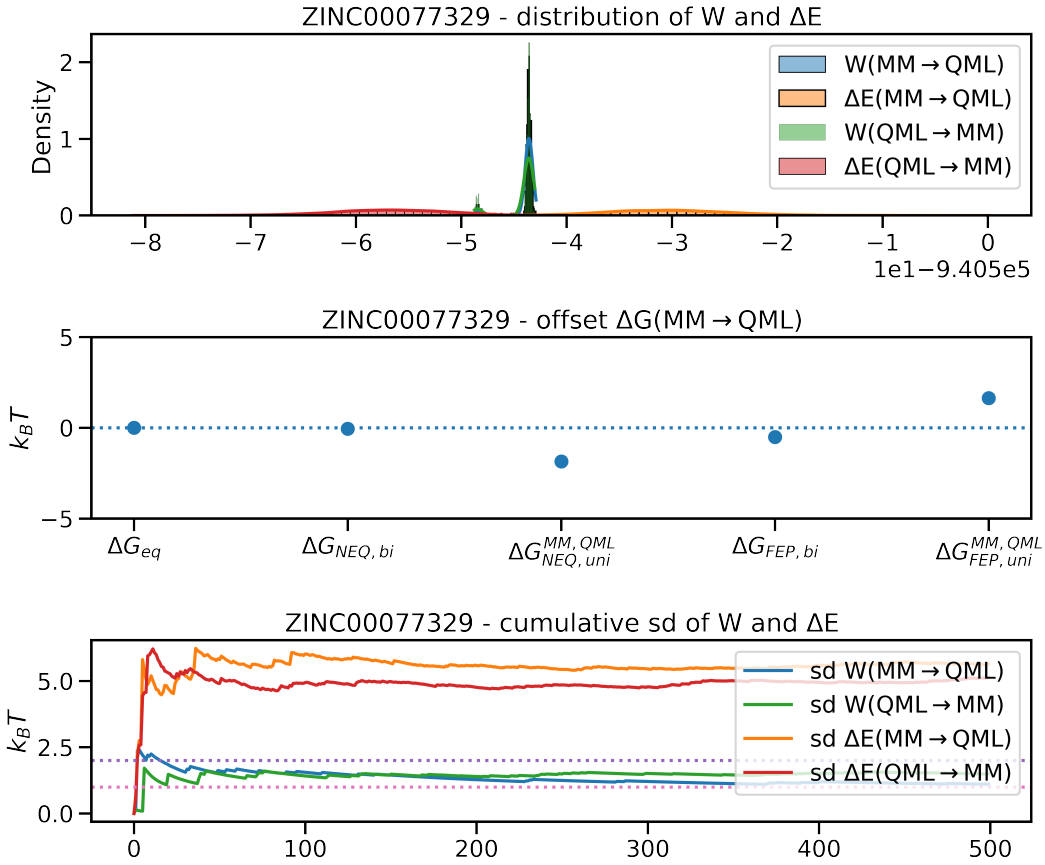


Figure 3.11.: 50 ps NEQ switching and FEP results for molecule **1** (ZINC00077329). See figure 3.2 for additional details.

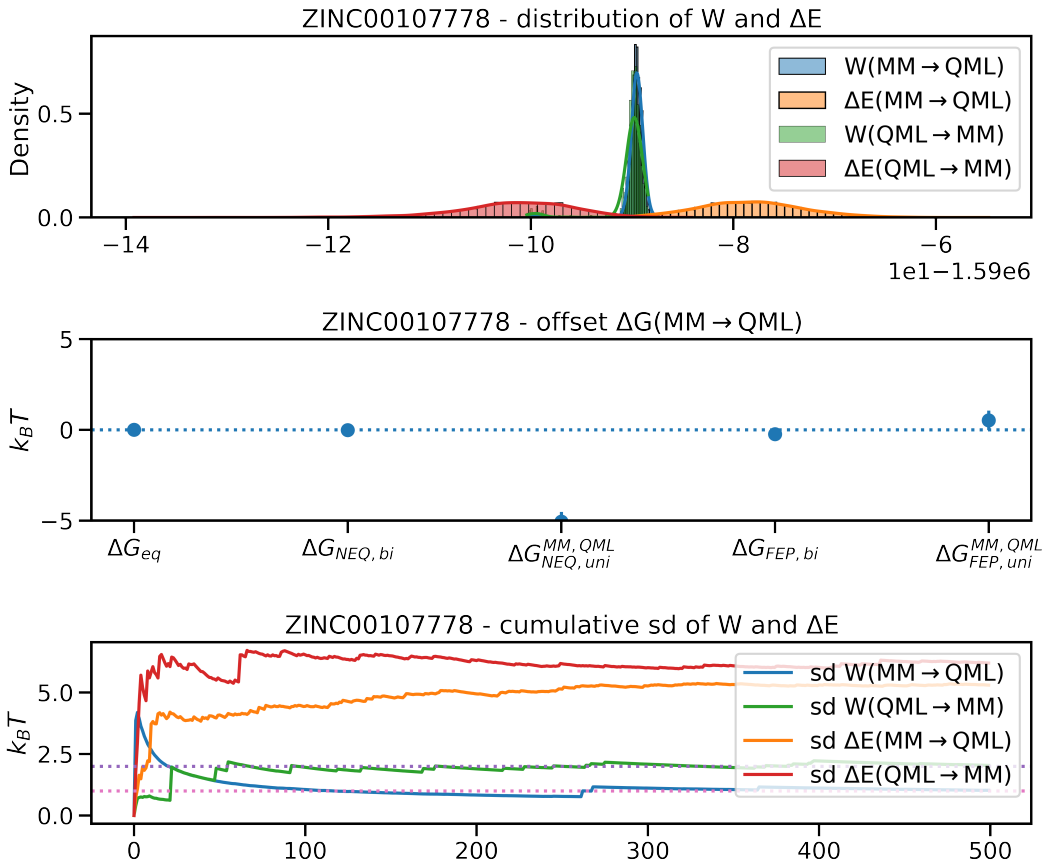


Figure 3.12.: 5 ps NEQ switching and FEP results for molecule **6** (ZINC00107778). (Top) Density plot of work and energy distributions. (Middle) Comparison of NEQ and FEP free energy values to the equilibrium reference from the first protocol run. FEP results are calculated each time the plot is generated. Thus, those results vary in the following figures. (Bottom) Cumulative standard deviation (in $k_BT$) of work and energy distributions recorded over 500 switches. For a detailed description of these plots see figure 3.2

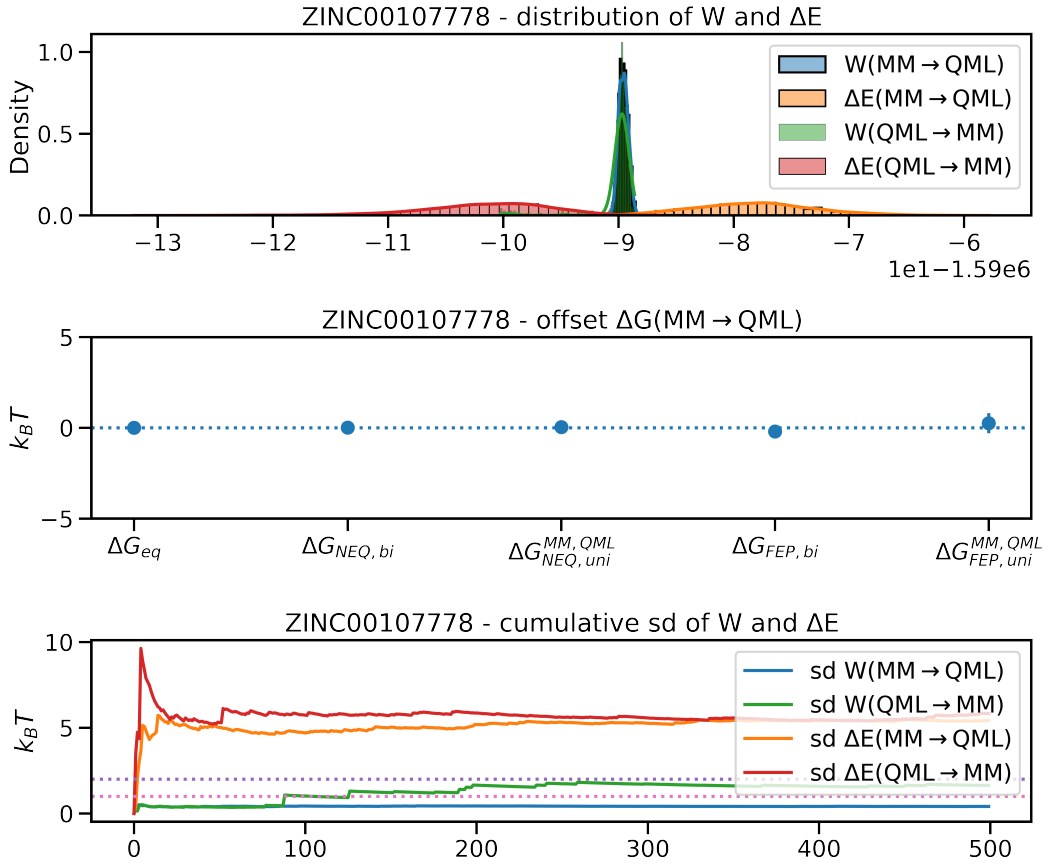


Figure 3.13.: 10 ps NEQ switching and FEP results for molecule **6** (ZINC00107778). See figure 3.2 for additional details.

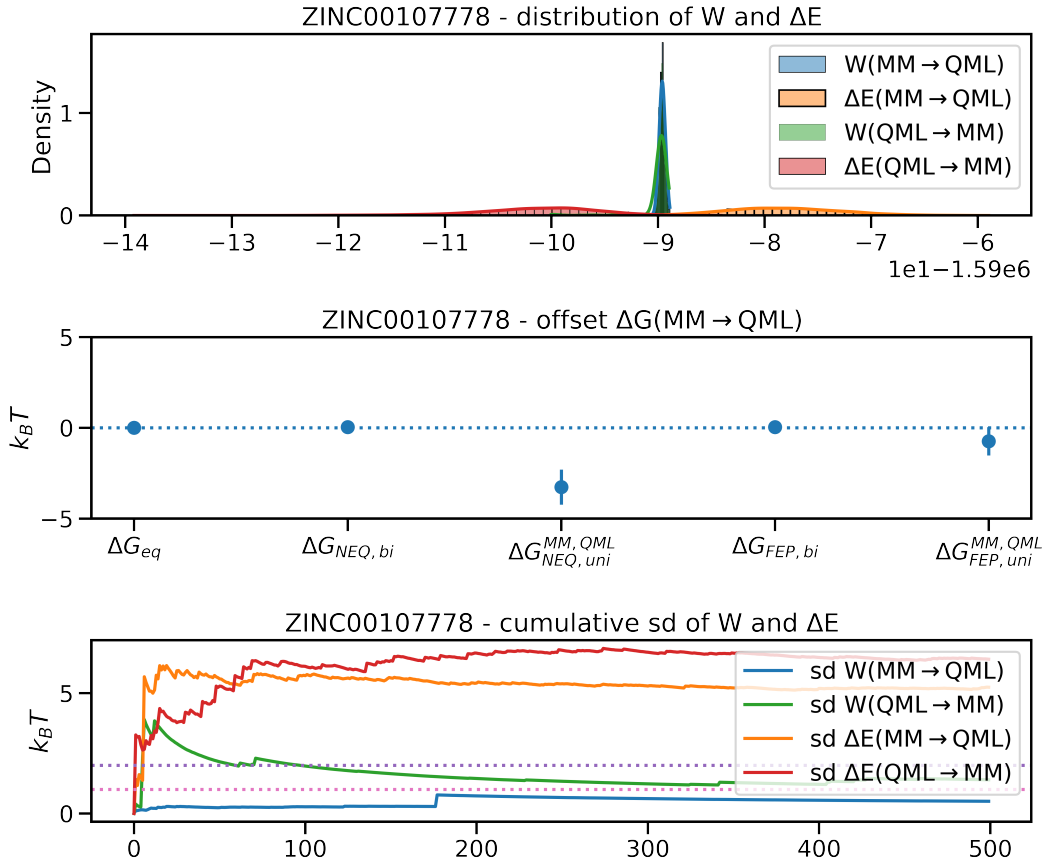


Figure 3.14.: 20 ps NEQ switching and FEP results for molecule **6** (ZINC00107778). See figure 3.2 for additional details.

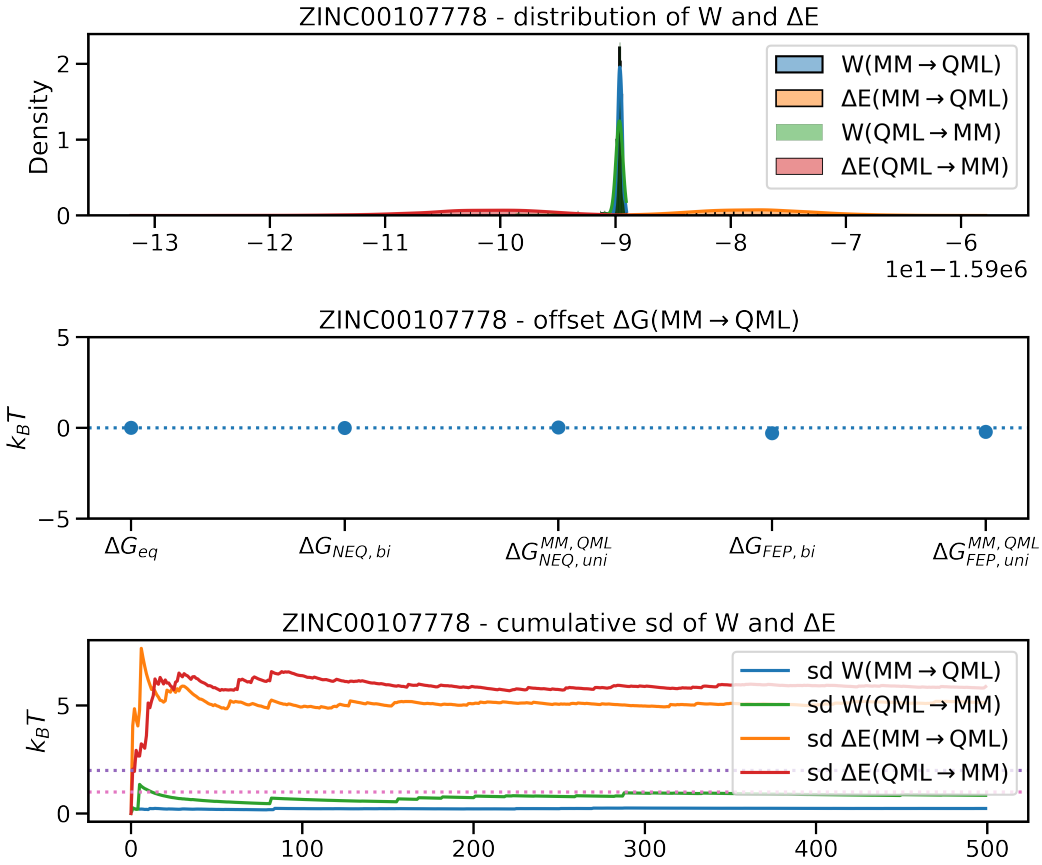


Figure 3.15.: 50 ps NEQ switching and FEP results for molecule **6** (ZINC00107778). See figure 3.2 for additional details.

### 3.5.3. Effect on conformational sampling

Longer switching lengths impact the torsion profiles. This will be illustrated for molecule **7**, which was already analyzed in section 3.5.1. While in this particular molecule five torsion angles can be identified, I will focus on two of these; their torsion plots are shown in figures 3.16–3.19. The remaining torsion profiles can be found in the Supporting Materials section (figures 56–59).

The dihedral angle distribution [1,3,13,14] varies depending on the level of theory used for its generation. While the MM potential results in a rather broad distribution with a clear energetic minimum at around $\frac{\pi}{3}$, the QML description predicts a symmetric distribution with two minima around $\pm\frac{3\pi}{4}$. In this case, the effect of the switching length becomes more evident than in the earlier example (ZINC0867000, discussed in section 3.4.2). This can be best observed when tracking the evolution of the MM $\rightarrow$ QML distribution. Besides the energetic minimum at around $\frac{\pi}{3}$, which is still present after the 5 ps switch, another energetic minimum at $\frac{3\pi}{4}$ emerges (see figure 3.16, red distribution in the top panel). This "new" energetic minimum is in agreement with the preferred angle found in the equilibrium QML simulations. The energetic minimum "on the other side" of the unit circle ($-\frac{3\pi}{4}$), however, remains poorly sampled. This changes for the longer switching lengths. When moving to 10, 20, and finally 50 ps the energetic minimum at $\frac{\pi}{3}$ gradually disappears, and the bulk of the density is shifted to $\frac{3\pi}{4}$. Additionally, even the "mirrored" minimum at -$\frac{3\pi}{4}$ is beginning to be sampled when the switching length is 50 ps (figure 3.19, top panel).

When inspecting the second torsion angle [4,5,6,11] a similar initial situation can be observed. As for the other torsion, the MM and QML level of theory do not agree on the conformational preference. While the QML description yields a dihedral distribution centered around 0, the MM force field leads to two minima at $\pm\frac{\pi}{2}$ (see figure 3.16, blue and orange distributions in the bottom panel). The distributions after the switch, however, reveal new information about the torsion profile. In the MM $\rightarrow$ QML direction (red plot), already for a switching length of 5 ps, one can observe energetic minima at $\approx \pm\pi$, which was neither present in the MM nor in the QML distribution.

It has to be mentioned, however, that the trends presented above are not overly pronounced (cf. figures 3.16–3.19 and torsion profiles presented in the Supporting Materials section).

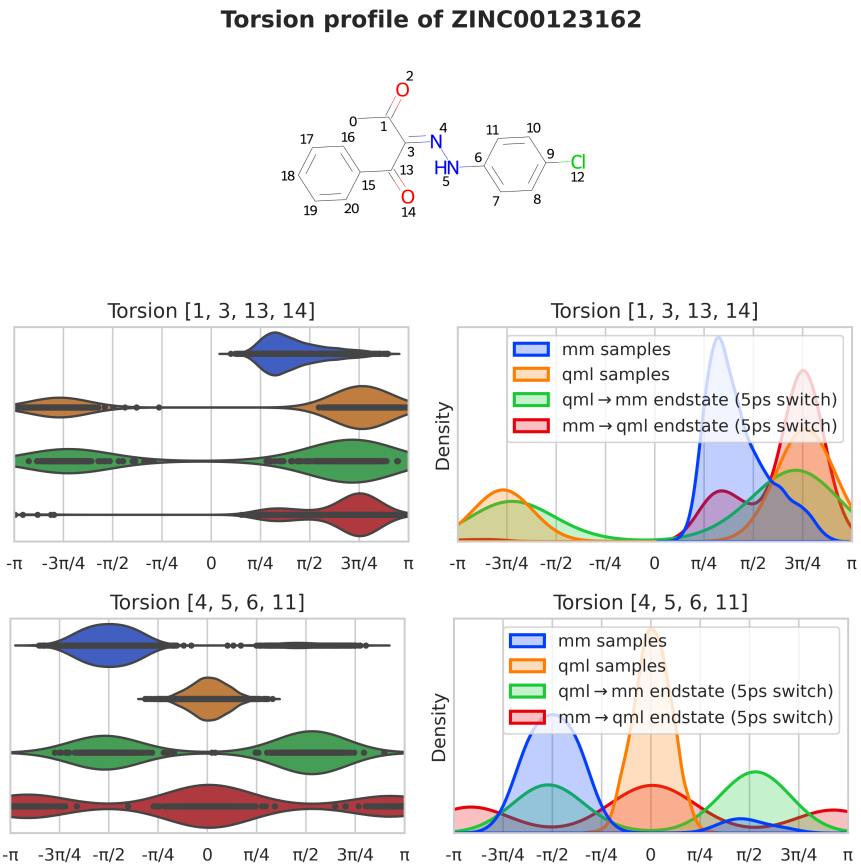


Figure 3.16.: Torsion profiles of molecule **7** before and after the 5 ps. For a detailed description see also figure 3.3.

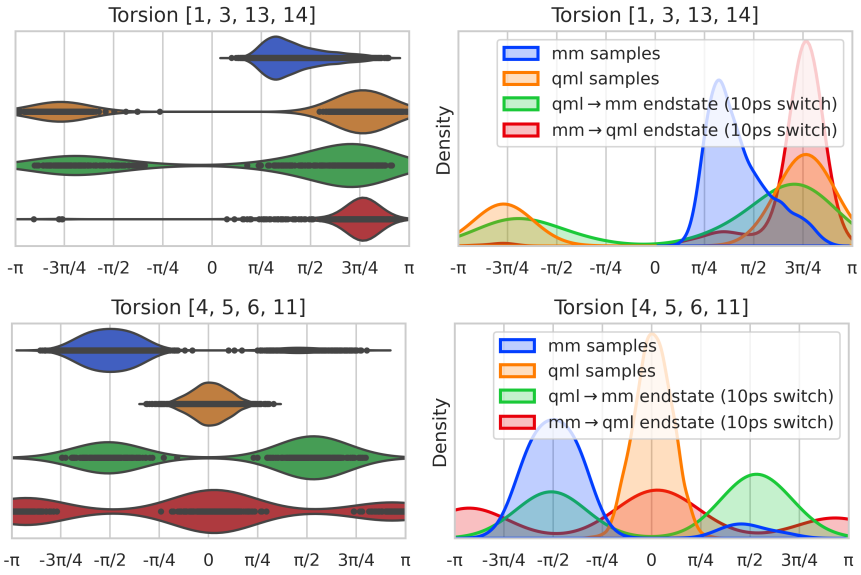


Figure 3.17.: Torsion profiles of molecule **7** before and after the 10 ps. For a detailed description see also figure 3.3.

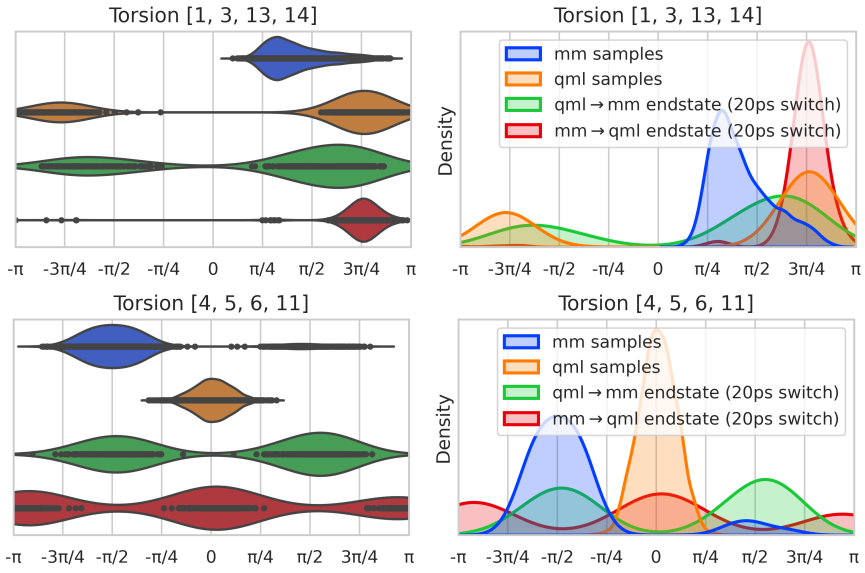


Figure 3.18.: Torsion profiles of molecule **7** before and after the 20 ps switch. For a detailed description see also figure 3.3.

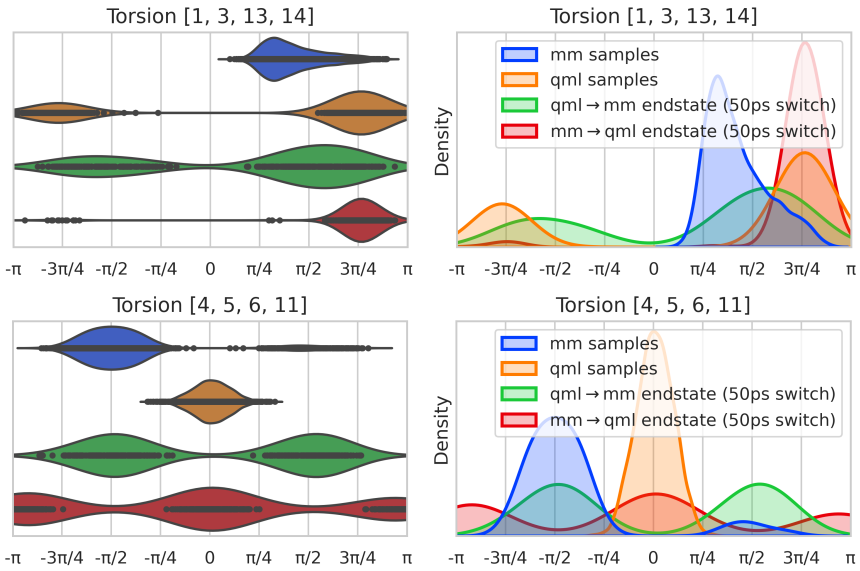


Figure 3.19.: Torsion profiles of molecule **7** before and after the 50 ps switch. For a detailed description see also figure 3.3.

# 4. Conclusion

We computed the free energy endstate correction between an MM (CGenFF) and a QML (ANI-2x) representation for 18 molecules. Three different computational approaches were used: Alchemical multi-state free energy calculations, free energy perturbation and non-equilibrium switching. For the latter two methods uni- and bidirectional free energy calculations were performed. We used alchemical multi-stage free energy calculations as reference calculations, which required 11 $\lambda$ states to model the transition from the CGenFF description of interactions to the ANI-2x potential.

The work (obtained with the NEQ protocol) and energy (obtained with the FEP protocol) distributions were analyzed using a set of criteria (variance, range of work/energy values, overlap and the presence of outliers).

While computationally extremely efficient, the FEP protocols resulted in broad and sparsely overlapping energy distributions, leading to inaccurate free energy estimates. The NEQ results, on the other hand, proved reliable free energy estimates in most cases. The bidirectional protocol led to accurate free energy estimates for each of the investigated systems.

Unidirectional NEQ protocols performed well; however, a few warning signals which could point to inaccurate estimates were identified: Firstly, bimodal work distributions are a potential warning sign and should lead to a closer investigation. It was found that especially bimodal distributions, where the second energetic minimum is clearly separated from the first one, might cause problems. Even if the bimodal distribution is narrow, a search for potential outliers is necessary. Distinctively multimodal and/or broad work distributions were also identified as sources of error.

The reason for the emergence of multimodal distributions and the presence of extreme work values, as well as their correlation with the resulting free energy estimates, is the topic of an ongoing investigation and will be addressed in a separate study. Most unimodal distributions did not cause difficulties. Nevertheless, it is necessary to inspect the obtained work values thoroughly, as isolated outliers, i.e., significantly more negative values compared to the average work, which in turn might cause inaccurate results, may not be discernible in density plots.

The impact of varying switching lengths (5, 10, 20 and 50 ps) on free energy estimates and work/energy distributions was inconclusive and no clear improvements for longer switching lengths could be observed. Overall, a switching length of 5 ps seems sufficient [32]. The narrowing of work distributions, which can be observed for longer switching lengths, has no discernible effect on the quality of the estimate.

It is much more the shape of the distribution which influences the result. In some cases, a longer switching procedure delivers torsion profiles, which are more in line with the target distribution or even shows conformations, which are not accessible with the equilibrium sampling procedure.

The results suggest that in most of the cases the correction of solvation and binding free energies can be efficiently performed with a unidirectional 5 ps NEQ switching protocol, which does not require sampling at the QML level of theory. This method is expected to deliver accurate results, if the warning signs stated above are monitored closely.

# Bibliography


[1] Behler, J., and Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Physical Review Letters 98*, 14 (2007), 1–4.

[2] Bennett, C. H. Efficient Estimation of Free Energy Differences from Monte Carlo Data. *Journal of Computational Physics 22*, 2 (1976), 245–268.

[3] Bento, A. P., Gaulton, A., Hersey, A., Bellis, L. J., Chambers, J., Davies, M., Krüger, F. A., Light, Y., Mak, L., McGlinchey, S., Nowotka, M., Papadatos, G., Santos, R., and Overington, J. P. The ChEMBL bioactivity database: An update. *Nucleic Acids Research 42*, D1 (2014), 1083–1090.

[4] Boresch, S., and Woodcock, H. L. Convergence of single-step free energy perturbation. *Molecular Physics 115*, 9-12 (2017), 1200–1213.

[5] Brauer, B., Kesharwani, M. K., Kozuch, S., and Martin, J. M. The S66x8 benchmark for noncovalent interactions revisited: Explicitly correlated ab initio methods and density functional theory. *Physical Chemistry Chemical Physics 18*, 31 (2016), 20905–20925.

[6] Choi, H., Kang, H., and Park, H. New solvation free energy function comprising intermolecular solvation and intramolecular self-solvation terms. *Journal of Cheminformatics 5*, 2 (2013), 1–13.

[7] Cohen, A. J., Mori-Sánchez, P., and Yang, W. Insights into Current Limitations of Density Functional Theory. *Science 321*, 5890 (2008), 792–794.

[8] Cournia, Z., Allen, B., and Sherman, W. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *Journal of Chemical Information and Modeling 57*, 12 (2017), 2911–2937.

[9] Crooks, G. E. Entropy production fluctuation theorem and the nonequilibrium work relation. *Physical Review E 60*, 3 (1999), 2721–2726.

[10] Dellago, C., and Hummer, G. Computing Equilibrium Free Energies Using Non-Equilibrium Molecular Dynamics. *Entropy 16*, 1 (2014), 41–61.

[11] Devereux, C., Smith, J. S., Davis, K. K., Barros, K., Zubatyuk, R., Isayev, O., and Roitberg, A. E. Extending the Applicability of the ANI Deep Learning Molecular Potential to Sulfur and Halogens. *Journal of Chemical Theory and Computation 16* (2020), 4192–4202.

[12] Eastman, P., Swails, J., Chodera, J. D., McGibbon, R. T., Zhao, Y., Beauchamp, K. A., Wang, L. P., Simmonett, A. C., Harrigan, M. P., Stern, C. D., Wiewiora, R. P., Brooks, B. R., and Pande, V. S. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Computational Biology 13*, 7 (2017), 1–17.

[13] Fink, T., Bruggesser, H., and Reymond, J. L. Virtual Exploration of the Small-Molecule Chemical Universe below 160 Daltons. *Angewandte Chemie - International Edition 44*, 10 (2005), 1504–1508.

[14] Fink, T., and Raymond, J. L. Virtual Exploration of the Chemical Universe up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discovery. *Journal of Chemical Information and Modeling 47*, 2 (2007), 342–353.

[15] Galvelis, R., Varela-Rial, A., Doerr, S., Fino, R., Eastman, P., Markland, T. E., Chodera, J. D., and De Fabritiis, G. NNP/MM: Fast molecular dynamics simulations with machine learning potentials and molecular mechanics. *arXiv* (2022). doi: 10.48550/ARXIV.2201.08110.

[16] Genheden, S. Solvation free energies and partition coefficients with the coarse-grained and hybrid all-atom/coarse-grained MARTINI models. *Journal of Computer-Aided Molecular Design 31*, 10 (2017), 867–876.

[17] Giese, T. J., and York, D. M. Development of a Robust Indirect Approach for MM $\rightarrow$ QM Free Energy Calculations That Combines Force-Matched Reference Potential and Bennett's Acceptance Ratio Methods. *Journal of Chemical Theory and Computation 15*, 10 (2019), 5543–5562.

[18] Goette, M., and Grubmüller, H. Accuracy and Convergence of Free Energy Differences Calculated from Nonequilibrium Switching Processes. *Journal of Computational Chemistry 30*, 3 (2009), 447–456.

[19] Hudson, P. S., Woodcock, H. L., and Boresch, S. Use of Nonequilibrium Work Methods to Compute Free Energy Differences Between Molecular Mechanical and Quantum Mechanical Representations of Molecular Systems. *Journal of Physical Chemistry Letters 6*, 23 (2015), 4850–4856.

[20] Hummer, G., and Szabo, A. Free energy reconstruction from nonequilibrium single-molecule pulling experiments. *Proceedings of the National Academy of Sciences of the United States of America 98*, 7 (2001), 3658–3661.

[21] Hummer, G., and Szabo, A. Free energy profiles from single-molecule pulling experiments. *Proceedings of the National Academy of Sciences of the United States of America 107*, 50 (2010), 21441–21446.

[22] Jarzynski, C. Nonequilibrium Equality for Free Energy Differences. *Physical Review Letters 78*, 14 (1997), 2690–2693.

[23] Jia, X., Ge, H., and Mei, Y. Free energy change estimation: The Divide and Conquer MBAR method. *Journal of Computational Chemistry 42*, 17 (2021), 1204–1211.

[24] Kearns, F. L., Warrensford, L., Boresch, S., and Woodcock, H. L. The good, the bad, and the ugly: "Hipen", a new dataset for validating (S)QM/MM free energy simulations. *Molecules 24* (2019), 1–28.

[25] Klimovich, P. V., Shirts, M. R., and Mobley, D. L. Guidelines for the analysis of free energy calculations. *Journal of Computer-Aided Molecular Design 29*, 5 (2015), 397–411.

[26] Lee, J., Miller, B. T., and Brooks, B. R. Computational scheme for pH-dependent binding free energy calculation with explicit solvent. *Protein Science 25*, 1 (2016), 231–243.

[27] Mey, A. S., Allen, B. K., Bruce Macdonald, H. E., Chodera, J. D., Hahn, D. F., Kuhn, M., Michel, J., Mobley, D. L., Naden, L. N., Prasad, S., Rizzi, A., Scheen, J., Shirts, M. R., Tresadern, G., and Xu, H. Best Practices for Alchemical Free Energy Calculations [Article v1.0]. *Living Journal of Computational Molecular Science 2*, 1 (2020), 1–48.

[28] Neal, R. M. Annealed Importance Sampling. *Statistics and Computing 11* (2001), 125–139.

[29] Pathak, Y., Mehta, S., and Priyakumar, U. D. Learning Atomic Interactions through Solvation Free Energy Prediction Using Graph Neural Networks. *Journal of Chemical Information and Modeling 61*, 2 (2021), 689–698.

[30] Pohorille, A., Jarzynski, C., and Chipot, C. Good Practices in Free-Energy Calculations. *The Journal of Physical Chemistry B 114*, 32 (2010), 10235–10253.

[31] Rufa, D. A., Bruce Macdonald, H. E., Fass, J., Wieder, M., Grinaway, P. B., Roitberg, A. E., Isayev, O., and Chodera, J. D. Towards chemical accuracy for alchemical free energy calculations with hybrid physics-based machine learning / molecular mechanics potentials. *bioRxiv* (2020). doi: 10.1101/2020.07.29.227959.

[32] Schöller, A., Kearns, F., Woodcock, H. L., and Boresch, S. Optimizing the Calculation of Free Energy Differences in Nonequilibrium Work SQM/MM Switching Simulations. *Journal of Physical Chemistry B 126*, 15 (2022), 2798–2811.

[33] Shirts, M. R., and Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *Journal of Chemical Physics 129*, 12 (2008).

[34] Shirts, M. R., and Mobley, D. L. An Introduction to Best Practices in Free Energy Calculations. In *Biomolecular Simulations: Methods and Protocols*, L. Monticelli and E. Salonen, Eds. Humana Press, Totowa, NJ, 2013, pp. 271–311.

[35] Smith, J. S., Isayev, O., and Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chemical Science 8*, 4 (2017), 3192–3203.

[36] Stern, C. D., Bayly, C. I., Smith, D. G. A., Fass, J., Wang, L.-P., Mobley, D. L., and Chodera, J. D. Capturing non-local through-bond effects in molecular mechanics force fields I: Fragmenting molecules for quantum chemical torsion scans [Article v1.1]. *bioRxiv* (2022). doi: 10.1101/2020.08.27.270934.

[37] Vanommeslaeghe, K., Hatcher, E., Acharya, C., Kundu, S., Zhong, S., Shim, J., Darian, E., Guvench, O., Lopes, P., Vorobyov, I., and Mackerell, A. D. CHARMM General Force Field: A Force Field for Drug-Like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields. *Journal of Computational Chemistry 31* (2010), 671–690.

[38] Wan, S., Potterton, A., Husseini, F. S., Wright, D. W., Heifetz, A., Malawski, M., Townsend-Nicholson, A., and Coveney, P. V. Hit-to-lead and lead optimization binding free energy calculations for G protein-coupled receptors: Free energy calculations for GPCRs. *Interface Focus 10*, 6 (2020), 1–11.

[39] Wang, L., Wu, Y., Deng, Y., Kim, B., Pierce, L., Krilov, G., Lupyan, D., Robinson, S., Dahlgren, M. K., Greenwood, J., Romero, D. L., Masse, C., Knight, J. L., Steinbrecher, T., Beuming, T., Damm, W., Harder, E., Sherman, W., Brewer, M., Wester, R., Murcko, M., Frye, L., Farid, R., Lin, T., Mobley, D. L., Jorgensen, W. L., Berne, B. J., Friesner, R. A., and Abel, R. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *Journal of the American Chemical Society 137*, 7 (2015), 2695–2703.

[40] Wieder, M., Fass, J., and Chodera, J. D. Fitting quantum machine learning potentials to experimental free energy data: predicting tautomer ratios in solution. *Chemical Science 12*, 34 (2021), 11364–11381.

[41] Wieder, M., Fass, J., and Chodera, J. D. Teaching free energy calculations to learn from experimental data. *bioRxiv* (2021). doi: 10.1101/2021.08.24.457513.

[42] Wieder, M., Fleck, M., Braunsfeld, B., and Boresch, S. Alchemical free energy simulations without speed limits. A generic framework to calculate free energy differences independent of the underlying molecular dynamics program. *Journal of Computational Chemistry 43*, 17 (2022), 1151–1160.

[43] Xiong, H., Crespo, A., Marti, M., Estrin, D., and Roitberg, A. E. Free energy calculations with non-equilibrium methods: Applications of the Jarzynski relationship. *Theoretical Chemistry Accounts 116*, 1-3 (2006), 338–346.

[44] Yang, L., Horton, J. T., Payne, M. C., Penfold, T. J., and Cole, D. J. Modeling Molecular Emitters in Organic Light-Emitting Diodes with the Quantum Mechanical Bespoke Force Field. *Journal of Chemical Theory and Computation 17*, 8 (2021), 5021–5033.

[45] Zuckerman, D. M., and Chong, L. T. Weighted Ensemble Simulation: Review of Methodology, Applications, and Software. *Annual Review of Biophysics 46*, 1 (2017), 43–57.

[46] Zwanzig, R. W. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *Journal of Chemical Physics 22*, 1954 (2004), 1420–1426.

# A. Abstract

The accurate, computational assessment of free energy differences is challenging because of three principal difficulties: The molecular system has to be set up correctly, the potential energy function used has to accurately describe the molecular systems' energetics, and all relevant regions of the conformational space have to be sampled within the limits of one's computational resources. At present, most calculations of free energy differences rely on molecular dynamic (MD) simulations with a molecular mechanical description of energies and forces. One obvious improvement would be the correct treatment of inter-atomic forces using quantum mechanics (QM) instead of classical force fields. Due to the high computational cost, however, QM calculations cannot be used to compute free energy differences. Quantum machine learning (QML) potentials — which approach quantum mechanics level of theory accuracy while retaining classical mechanics computational costs — are a viable alternative. Although it is still an open question how to perform *alchemical* free energy calculations with QML potentials, endstate corrections from a classical mechanics force field to a QML description are possible and have been done. In this work, we performed such endstate corrections on 18 molecules in the gas phase. We used and compared three different methods (alchemical multi-state free energy calculations, non-equilibrium (NEQ) switching, and free energy perturbation (FEP)), as well as two free energy estimators (Bennett Acceptance Ratio and exponential averaging). The alchemical multi-state free energy calculations were treated as a "gold standard" reference for the free energy estimates' accuracy. The bidirectional NEQ switching protocol was the most accurate and stable approach. The bidirectional FEP results were also accurate for most of the studied systems. However, unidirectional NEQ switching delivered more reliable free energy estimates at a comparable computational cost. Therefore, we consider it the better choice.

# B. Kurzfassung

Eine akkurate Berechnung der Freien Energie Unterschiede ist aufgrund dreier Hauptschwierigkeiten eine Herausforderung: Das molekulare System muss korrekt aufgesetzt werden, die verwendete potentielle Energiefunktion muss die Energetik des molekularen Systems richtig beschreiben, und alle relevanten Bereiche des Konformationsraums müssen im Rahmen der zur Verfügung stehenden Rechenressourcen erkundet werden. Zurzeit beruhen die meisten Berechnungen Freier Energie Differenzen auf molekular-dynamischen (MD) Simulationen und einer klassisch-mechanischen Beschreibung von Energien und Kräften. Eine offensichtliche Verbesserung könnte mit Hilfe der Quantenmechanik (QM) erlangt werden, da diese eine korrekte Beschreibung interatomarer Kräfte anstelle der klassischen Beschreibung durch Kraftfelder ermöglicht. Aufgrund des hohen Rechenaufwands können QM-Rechnungen jedoch nicht zur Evaluierung Freier Energie Differenzen verwendet werden. Quantum machine learning (QML) Potentiale, die sich dem Niveau der theoretischen Genauigkeit der Quantenmechanik annähern und gleichzeitig die Rechenkosten der klassischen Mechanik beibehalten, sind eine günstige Alternative. Obwohl noch nicht vollständig geklärt werden konnte, wie *alchemische* Freie Energie Rechnungen mit QML-Potentialen durchgeführt werden können, sind Endzustand-Korrekturen von einem klassisch-mechanischen Kraftfeld zu einer QML-Beschreibung möglich und wurden bereits durchgeführt. In dieser Arbeit haben wir derartige Endzustand-Korrekturen für 18 Moleküle in der Gasphase verwendet. Wir verglichen drei verschiedene Methoden (alchemische Freie Energie Berechnungen, Nichtgleichgewichts(NEQ)-switching und free energy perturbation (FEP)) sowie zwei statistische Methoden zur Freien Energie Abschätzung (Bennett Acceptance Ratio und exponential averaging). Die alchemischen Freien Energie Werte wurden als „Goldstandard“-Referenz für die Genauigkeit der abgeschätzten Freien Energie angesehen. Das bidirektionale NEQ-switching Protokoll erwies sich als der genaueste und stabilste Ansatz. Auch mit der bidirektionalen FEP-Methode konnten für die meisten untersuchten Systeme akkurate Ergebnisse erhalten werden. Das unidirektionale NEQ-switching lieferte jedoch zuverlässigere Freie Energie Schätzungen bei vergleichbarem Rechenaufwand und erwies sich deshalb als effizienteste Methode.

# C. Supporting Materials

Due to a large number of supplementary materials, additional tables and figures are provided under `https://doi.org/10.5281/zenodo.7265114`.