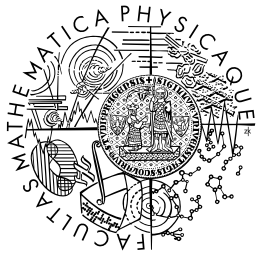

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MASTER THESIS

Martin Crhán

Full Centroid Molecular Dynamics through Machine Learning

Institute of Physics of Charles University

Supervisor of the master thesis: RNDr. Ondřej Maršálek, Ph.D.

Study programme: Theoretical Physics

Study branch: FTFP

Prague 2023

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In ..... date ....

Author's signature

I would like to express immense gratitude to my advisor Ondřej Maršálek, not only for all the help and consultation he has provided me over the course of writing this thesis but also for all of his invaluable professional and personal advice which has had a profound influence on putting me on the right track in life.

I also like to collectively thank all of my colleagues for always being open to discussion and providing a highly stimulating scientific environment.

On a personal note, I'd like to thank my parents, for always being there for me throughout all of this time. Further, I'd like to thank Anička for all the continuing love and support she has provided me, and all of the wonderful things she has done for me.

Title: Full Centroid Molecular Dynamics through Machine Learning

Author: Martin Crhán

Institute: Institute of Physics of Charles University

Supervisor: RNDr. Ondřej Maršálek, Ph.D., Institute of Physics of Charles University

Abstract: This thesis is concerned with developing new methodology for the more efficient execution of centroid molecular dynamics simulations — a method based on the theory of imaginary-time path integrals, commonly used in accurate computational prediction of vibrational spectra and other dynamical properties of condensed-phase molecular systems. This is done through the use of machine learning methods to explicitly construct the potential for the centroid. The results obtained through the new methodology are subsequently systematically compared with results obtained through the older, adiabatic, approach to centroid molecular dynamics. This is done for a range of low-dimensional model systems as well as realistic highly-dimensional molecular systems. The differences in these results and the advantages of the new approach are subsequently discussed. The properties of the potential which has been constructed have also been investigated.

Keywords: centroid, molecular dynamics, machine learning, path integrals, molecular systems, vibrational spectra

Contents

1 Introduction

Over the past decades, computational methods have taken a major role in theoretical studies in physics and other scientific fields. This rapid expansion has been facilitated by both advances in computer hardware and development in computational algorithms. One of the manifestations of these advances is the expansion of the possibilities of methods of molecular dynamics simulations, in which the statistical properties of many-body molecular systems are investigated by numerically solving their classical equations of motion (the methods can be used to investigate other systems as well, however, this work will not be concerned with these cases) [1]. An important direction in the development of these methods is the strive for a more accurate description of the simulated atoms and their interactions. In the past, the interatomic interactions were, for realistic systems, often available only in a highly approximate form as a sum of pairwise interactions. The recent developments in the field of *ab-initio* molecular dynamics [2] allow a highly accurate description of these interactions, obtained by solving the ground state electronic problem, even for moderately sized systems. Furthermore, the quantum mechanical effects on the equilibrium properties of the atomic nuclei can be included using the Feynman path integral formulation in imaginary time [3], [4]. This formulation allows one to map the problem of quantum statistics onto a problem of classical statistics where each particle is replaced by a ring polymer, which in turn can be simulated using molecular dynamics. These *path-integral molecular dynamics* simulations, however, require an accurate description of the interatomic interactions and are significantly more costly than their classical counterparts, and as such, they were until recently available only for rather simple systems or at a very high computational cost.

Despite this progress, the inclusion of quantum *dynamical* effects into molecular dynamics simulations still presents a major challenge. The main obstacle towards this goal is the numerical sign problem [5] — the direct calculation of the exact time evolution of a high-dimensional quantum mechanical system, be it through the usage of the Feynman path integral in real time, or through the explicit calculation of the time evolution operator, generally requires the calculation of a sum of a large number of mutually nearly canceling complex phase factors. Such a task is virtually impossible to do on a computer. This problem can be overcome by other means — most notably by diagonalizing the Hamiltonian, but these approaches tend to have highly unfavorable scaling with system size (the diagonalization of the Hamiltonian scales exponentially with system size). The size of the systems for which exact quantum dynamics can be computed is thus highly limited. In the study of molecular systems, this presents an issue for the theoretical prediction of the results of a number of experiments, such as infrared or Raman spectroscopy, which investigate the dynamical properties of the system.

1.1 Approximations to quantum dynamics

To solve this issue, a number of approximations have been developed which allow one to include quantum dynamical effects in molecular dynamics simulations of condensed systems without running into the numerical sign problem. These approximations generally take advantage of the fact that decoherence times tend to be small in condensed systems, and thus the effects of quantum coherence are significantly suppressed. The *semiclassical* class of methods introduces quantum corrections to classical dynamics by performing a systematic expansion in the powers of $\hbar$. Methods such as the Herman-Kluk propagator [6] or the linearized semiclassical initial value representation (LSC-IVR) [7] fall into this category. The general drawback of these methods is that they do not conserve the quantum Boltzmann distribution, and as such, the dynamical information obtained using them degrades with simulation time.

Another class of methods, which takes what could be called an "opposite" approach to semiclassical methods, are methods based on imaginary-time path-integral molecular dynamics. In imaginary-time path-integral molecular dynamics, each particle is replaced by a classical *ring polymer* of P replicas of the particle, connected by harmonic couplings, and the resulting system is simulated using molecular dynamics in order to determine its statistical properties. It can be shown that in the limit $P \to \infty$, the partition function of such a classical system *exactly* coincides with the quantum-mechanical partition function, and the equilibrium distribution of replicas exactly matches the quantum mechanical distribution of the particle positions. In numerical simulations, the number of replicas of course cannot be infinite, but it can be systematically increased to achieve convergence. Imaginary-time path-integral molecular dynamics can thus be used to calculate the equilibrium properties of quantum mechanical systems *exactly*. The dynamics are used in this scheme merely as a sampling tool to obtain these equilibrium averages. However, it turns out that a number of methods can be constructed, which use this classical dynamics to approximate the underlying quantum dynamics. The primary motivating feature behind using the auxiliary classical dynamics as a source of dynamical information is the fact that it exactly conserves the quantum Boltzmann distribution. The approximative methods in question prominently include ring polymer molecular dynamics (RPMD) [8], thermostatted ring polymer molecular dynamics (TRPMD) [9], and the subject of this work — centroid molecular dynamics (CMD) [10]. These methods can be shown to be exact in the classical, harmonic, and short-time limits. Furthermore, the series of approximations that takes one from exact quantum dynamics to RPMD and CMD can be established [11], with the help of *Matsubara dynamics* [12], as an intermediate step. Despite this, there is no good justification for taking these specific approximation steps and as such,

they remain rather ad-hoc in nature. Furthermore, both methods perform poorly for operators which are non-linear functions of the position and momentum operators [11]. The mentioned Matsubara dynamics presents a less drastic approximation to exact quantum dynamics, while still being classical, and quantum Boltzmann conserving. However, its practical usage is severely restricted by the fact that it still involves weighting contributions by a complex phase and as such still suffers from the numerical sign problem.

1.2 Centroid molecular dynamics

As a method, centroid molecular dynamics traces its roots to the seminal paper by Feynmann and Kleinert [13], in which it was shown that the imaginary-time path integral *centroid* can be used to obtain the equilibrium statistical properties of quantum mechanical systems. In numerical applications, the centroid corresponds to the center of mass of the ring polymer used in path-integral molecular dynamics and is a *classical* object. The potential "felt" by the centroid is the centroid potential of mean force — a free energy surface obtained by integrating out all non-centroid degrees of freedom of the ring polymer. In their original paper, Feynmann and Kleinert were able to find a scheme to yield accurate approximations to this potential of mean force. Furthermore, they were able to use the scheme to calculate approximations to the imaginary-time correlation functions in the systems studied by them and to analytically extend them to obtain real-time correlation functions. CMD takes this success as a motivating feature and approximates the quantum dynamics of a particle by the classical dynamics of the centroid, moving on the *exact* centroid potential of mean force. To determine the potential of mean force and its derivatives at each point visited in centroid molecular dynamics, a full converged path-integral molecular dynamics simulation with the particle centroids fixed would need to be run at every such point. Running centroid molecular dynamics according to such a scheme would be computationally demanding even for small systems, and especially so for complex condensed-phase molecular systems. To sidestep these computational demands, centroid molecular dynamics is usually run according to the adiabatic scheme [14]. In this scheme, imaginary-time path-integral molecular dynamics is run with all the non-centroid normal modes of the ring polymer shifted higher, such that they are adiabatically separated from the physical modes present in the system. Due to this separation, the centroid approximately moves on the potential of mean force, which is thus effectively generated on the fly. The higher the adiabatic separation is, the closer the potential effectively felt by the centroid is to the true potential of mean force, however, higher adiabatic separation also implies the usage of a shorter time step, which increases the computational effort needed. The effect of the adiabatic separation was investigated in the past and it was suggested that full adia-

batic separation is in certain cases not necessary to yield reasonable dynamics, yielding the computationally cheaper partially adiabatic centroid molecular dynamics (PACMD) method [15].

1.3 Machine learning

In recent years, methods of machine learning have gained significant popularity in many areas of computer science as well as other fields for solving a wide range of complex problems. In the computational study of complex molecular systems, these methods have shown themselves to be particularly useful for constructing the so-called *machine learning potentials* [16]. These are machine learning models that are able to correctly reproduce potential energy surfaces for atomic and molecular systems, calculated using a chosen method of interest. These are necessary for running molecular dynamics, and for ab initio molecular dynamics in particular, this presents a significant speedup, as this eliminates the necessity to run an electronic structure calculation at each step of the molecular dynamics simulation. Of course, to construct these models, sample data from electronic structure calculations (or another reference method of choice) is needed, which the model "learns". However, the number of these reference calculations needed tends to be much lower than what would be required by a typical molecular dynamics simulation.

1.4 Outline

This work will be mainly concerned with constructing an alternative approach to adiabatic centroid molecular dynamics, in which the potential of mean force is predicted at each simulation point by machine learning methods. It is proposed that the framework of machine learning potentials can not only be used to run CMD in a fashion that is closer to its original formulation even for condensed-phase molecular systems, but at the same time to achieve a significant improvement in performance over the previously used adiabatic CMD. To this end, focus will be given to detailing the implementation of *full* CMD, investigating its properties as well as those of the adiabatic CMD, and comparing the results to previous calculations and experimental results, where possible. These investigations will be performed on both realistic molecular systems of high dimensionality and on model one- or two-dimensional systems, where the main features of the methods can be demonstrated more clearly.

The discussion will begin by outlining the established theoretical background of the physical problems at hand, the approximations involved, and the computational methods used to solve these problems. This will be followed by detailing the implementation itself and the computational tools in-

volved. Lastly, the results obtained will be presented and discussed.

2 Theoretical background

This section will deal primarily with reviewing the established theory and methods used to describe molecular many-body systems. This theory has been sourced from the literature as well as other resources, and written down by the author in his own words. First, the basic formulation of the problem will be given, with attention paid to the approximations commonly involved. Next, an overview of both equilibrium and non-equilibrium statistical mechanics relevant to this work will be presented. The section will conclude with a general description of the computational methods used to solve the problems at hand. The practical details of the implementation of these methods will be left for the next chapter.

2.1 The molecular problem

The information presented in this section was mostly sourced from the book 2.

The most elementary non-relativistic description of a molecular system is given by the time-dependent Schrödinger equation:

$$i\frac{\partial}{\partial t}\psi(x,X;t) = \hat{H}\psi(x,X;t)\,. \tag{2.1}$$

Where x is used to denote the $3N$ coordinates of the electrons and X corresponds to the $3M$ coordinates of the nuclei. In writing this equation, the choice $\hbar = 1$ was made. Such a choice is a part of the atomic system of units:

$$\begin{aligned} \hbar &= 1 & 4\pi\epsilon_0 &= 1 \\ e &= 1 & m_e &= 1, \end{aligned}$$

which will be used throughout this section. In this system of units, the molecular Hamiltonian takes the compact form:

$$\begin{aligned} \hat{H} = &-\sum_{i=1}^{N}\frac{1}{2}\nabla^2_{\vec{x}_i} - \sum_{i=1}^{M}\frac{1}{2M_i}\nabla^2_{\vec{X}_i} + \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{|\vec{x}_i-\vec{x}_j|} \\ &+ \sum_{i=1}^{M}\sum_{j>i}^{M}\frac{Z_iZ_j}{|\vec{X}_i-\vec{X}_j|} - \sum_{i=1}^{N}\sum_{j=1}^{M}\frac{1}{|\vec{x}_i-\vec{X}_j|}, \end{aligned} \tag{2.2}$$

where $\vec{x}_i$ and $\vec{X}_i$ denote the three-dimensional coordinates of the individual electrons/nuclei. Z_i and M_i are the nuclear charges and masses, respectively. This equation can be rewritten and eventually approximated by first employing the following ansatz, introduced by Born [17]:

$$\psi(x,X;t) = \sum_{i=0}^{\infty}\chi_i(x;X)\phi_i(X;t). \tag{2.3}$$

The functions $\chi_i(x,X)$ are the solutions of the stationary *electronic* Schrödinger equation in which the nuclei are considered clamped:

$$\hat{H}_{\text{elec}}\chi_i(x;X) = E_i(X)\chi_i(x;X) \tag{2.4}$$

$$\hat{H}_{\text{elec}} = -\sum_{i=1}^{N}\frac{1}{2}\nabla^2_{\vec{x}_i} + \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{|\vec{x}_i-\vec{x}_j|} + \sum_{i=1}^{M}\sum_{j>i}^{M}\frac{Z_iZ_j}{|\vec{X}_i-\vec{X}_j|} - \sum_{i=1}^{N}\sum_{j=1}^{M}\frac{1}{|\vec{x}_i-\vec{X}_j|}. \tag{2.5}$$

They thus depend parametrically on the positions of the nuclei X and so do their corresponding eigenvalues $E_i(X)$. Assuming the spectrum of $\hat{H}_{\text{elec}}$ is non-degenerate, these functions are orthogonal and can be chosen to be normalized to unity. The functions $\phi_i(X;t)$ can be viewed as time-dependent coefficients. Using this ansatz in equation 2.1, multiplying it by $\chi_j^*(x;X)$, integrating over all coordinates x, and using the orthonormality of the functions χ_i yields

$$i\frac{\partial}{\partial t}\phi_j = \left(-\sum_{i=1}^{M}\frac{1}{2M_i}\nabla^2_{\vec{X}_i} + E_j(X)\right)\phi_j + \sum_{i=0}^{\infty}C_{ji}\phi_i, \tag{2.6}$$

where

$$C_{ji} = \sum_{k=1}^{M}\int \mathrm{d}x\chi_j^*\left(-\frac{1}{2M_k}\nabla^2_{\vec{X}_k}\right)\chi_i - \sum_{k=1}^{M}\frac{1}{M_k}\int \mathrm{d}x\chi_j^*\nabla_{\vec{X}_k}\chi_i\nabla_{\vec{X}_k}. \tag{2.7}$$

At this point, the *adiabatic approximation* will be made, according to which the non-diagonal elements of the matrix C_{ji} will be omitted. Under such an assumption, the electronic part of the system remains in a single state k during the time evolution. This also implies that the equations for the functions ϕ_i are now fully decoupled. If the wavefunctions χ_i are considered to be real, the diagonal elements C_{jj} themselves contain only a single term:

$$C_{jj} = \sum_{k=1}^{M}\int \mathrm{d}x\chi_j^*\left(-\frac{1}{2M_k}\nabla^2_{\vec{X}_k}\right)\chi_j, \tag{2.8}$$

due to:

$$\int \mathrm{d}x\chi_j\nabla_{\vec{X}_k}\chi_j = \frac{1}{2}\nabla_{\vec{X}_k}\left(\int \mathrm{d}x\chi_j\chi_j\right) = \nabla_{\vec{X}_k}1 = 0. \tag{2.9}$$

The wavefunctions can be safely considered real if the Hamiltonian is time reversible. Since this will be the case throughout this work the wavefunctions will be taken as real and the simplification presented by the preceding equation will be used. A final simplification will be made by omitting the diagonal elements C_{kk} as well. In the end, this finally yields the well-known *Born-Oppenheimer* approximation [18], which stipulates that the motion of the nuclei and the electrons is fully decoupled, i.e. the electrons adapt infinitely fast to the movement of the nuclei. Taken together, this results in the

following equations for the nuclear wavefunctions ϕ_i:

$$i\frac{\partial}{\partial t}\phi_i = \left(-\sum_{i=1}^{M}\frac{1}{2M_i}\nabla^2_{\vec{X}_i} + E_i(X)\right)\phi_i \equiv \hat{H}_{\text{nuc}}\phi_i. \tag{2.10}$$

Each ϕ_i is thus a solution of the Schrödinger equation, with the potential equal to the eigenvalue $E_i(X)$ of the electronic Schrödinger equation. The approximate solutions to equation (2.10) are the main concern of the present work. Attention will also be given exclusively to the electronic ground state ($i = 0$) case. As the typical first excitation energy of a molecular system is of the order of 1 eV, the temperature at which the $i \neq 0$ states would start to contribute significantly is $1eV/k_B \approx 10000$ K, which is well above the temperature ranges considered in this work. Aside from thermal excitation, the system could also be excited into the higher electronic states by interaction with electromagnetic radiation or other external perturbations. However, these cases will not be considered in this work.

A somewhat drastic approximation to the last equation is to invoke the correspondence principle (see appendix B for details) and to consider the nuclei to be classical particles, Hamilton's function being equal to:

$$H = \sum_{i=1}^{M}\frac{\vec{p}_i^2}{2M_i} + E_0(X). \tag{2.11}$$

Making this step presents an enormous simplification, as the partial differential equation 2.10 is effectively replaced by an ordinary differential equation. For highly-dimensional systems the original partial differential equation is generally impossible to solve (numerically or otherwise) and even though methods exist to circumvent this difficulty in certain cases, as will be shown in the next subsection, making the classical approximation is still the most common way to describe such systems. This level of approximation will also be used to motivate the molecular dynamics method in subsection 2.4, and as a useful reference, due to its common usage.

A necessary ingredient to perform any computation at the level of approximation outlined previously is the ground state electronic energy as a function of the coordinates of the nuclei $E_0(X)$. Determining this *potential energy surface* (PES) is the task of electronic structure methods. As investigation of these methods was not part of the goals of this work, and their inner workings thus of relatively low importance, the details of these methods will be omitted from this discussion. The basic ideas can be found summarized in appendix A. Instead, it will be assumed that the PES $E_0(X)$, as well as its derivatives with respect to the nuclear coordinates $\nabla_X E_0(X)$, are readily available as a function of X for all relevant positions of the nuclei, for instance through the use of *machine learning potentials* which will be further described in subsection 3.3. In any case, these quantities can be obtained through direct solution of the electronic problem at worst.

2.2 Path integral formulation of quantum mechanics

This section has been written down by the author based on the established knowledge, obtained at the lecture "Vybrané partie teorie kvantovaných polí I" taught by RNDr. Jiří Novotný, Csc., as well as the paper 13, and book 1.

Macroscopic observables can be obtained from the microscopic description of the system, introduced in the preceding chapter, using the methods of statistical mechanics. In particular, using the canonical model of thermodynamic equilibrium, the full statistical description of a system is contained in the canonical density operator:

$$\hat{\rho}(\beta) = \exp\left(-\beta \hat{H}\right), \tag{2.12}$$

where $\beta = 1/k_{\mathrm{B}}T$. The expectation value of an observable corresponding to an operator $\hat{A}$ is equal to

$$\langle \hat{A} \rangle = \frac{1}{Z} \mathrm{Tr}\left(\hat{A}\hat{\rho}\right), \tag{2.13}$$

the partition function Z is calculated as

$$Z(\beta, \lambda) = \mathrm{Tr}\left(\rho(\hat{\lambda})\right) = \mathrm{Tr}\left[\exp\left(-\beta \hat{H}(\lambda)\right)\right]. \tag{2.14}$$

In this expression, λ are some general parameters of the Hamiltonian. Important thermodynamic and statistical quantities (such as free energy or internal energy, among others) can be obtained from this function and its derivatives using the standard methods of statistical mechanics.

The matrix elements of the canonical density operator could in principle be directly calculated by diagonalizing the Hamiltonian. For example in the position basis:

$$\hat{\rho}(\beta)(x, x') = \sum_{i=0}^{\infty} \exp\left(-\beta E_i\right) \langle x|E_i\rangle \langle E_i|x'\rangle = \sum_{i=0}^{\infty} \exp\left(-\beta E_i\right) \psi_i^*(x')\psi_i(x), \tag{2.15}$$

however, this is seldom possible for a general many-particle molecular system. They can, nonetheless, be calculated in the position representation even for these systems using the imaginary-time path integral approach.

The path integral formulation is in principle a separate formulation of quantum mechanics developed in its modern form by Feynman [19] (it is thus also often explicitly referred to as the *Feynman* path integral formulation), however, note that the main ideas behind it can be traced back to Dirac [20]. This formulation can be viewed in two ways: as a consequence (and a useful reformulation) of the regular operator formulation of quantum mechanics, or as a separate formulation “guessed” from first principles, which can be a posteriori shown to be equivalent to the “regular” operator formulation of quantum mechanics [3]. The former approach will be adopted here.

The main objects of interest of the formulation are the matrix elements of the evolution operator in the position basis:

$$\langle x|\,\hat{U}(t)\,|x'\rangle = \left\langle x \middle| \exp\left(-\frac{i}{\hbar}\hat{H}t\right) \middle| x' \right\rangle. \tag{2.16}$$

(Time-homogeneity of the process has been assumed.) Note that the matrix elements of the canonical density operator can be obtained from the matrix elements of the evolution operator by performing an analytic extension $t \to -i\hbar\beta$, usually called the *Wick rotation*:

$$\hat{\rho}(\beta) = \exp\left(-\beta\hat{H}\right) = \exp\left(-\frac{i}{\hbar}\hat{H}(-i\hbar\beta)\right) = \hat{U}(-i\hbar\beta). \tag{2.17}$$

Attention will thus be given primarily to the evolution operator in the derivation, keeping in mind that the canonical density operator can be obtained in the end by performing such an analytic extension.

The main difficulty in evaluating the matrix elements of the evolution operator comes from the fact, that the expression contains an exponential of an operator composed of two non-commuting parts:

$$\begin{gathered} \hat{H} = \frac{\hat{p}^2}{2m} + \hat{U}(\hat{x}) \\ \left[\frac{\hat{p}^2}{2m}, \hat{U}\right] \neq 0. \end{gathered} \tag{2.18}$$

The operator exponential thus cannot be written simply as a product of exponentials of the individual parts. There are multiple ways to evaluate such an exponential, the one which the derivation relies upon being the *Lie–Trotter product formula* [21]

$$\exp(A+B) = \lim_{n\to\infty}\left[\exp\left(\frac{A}{n}\right)\exp\left(\frac{B}{n}\right)\right]^n, \tag{2.19}$$

where A and B are certain linear operators. For the details and limitations of the applicability of this formula see for instance [22]. The expression for the matrix elements thus becomes:

$$\left\langle x \middle| \exp\left(-\frac{i}{\hbar}\hat{H}t\right) \middle| x' \right\rangle = \lim_{P\to\infty} \left\langle x \middle| \left[\exp\left(\frac{-i\hat{p}^2}{2m\hbar P}\right)\exp\left(\frac{-i\hat{U}(\hat{x})}{\hbar P}\right)\right]^P \middle| x' \right\rangle. \tag{2.20}$$

In this expression, the operator exponentials only contain functions of either the $\hat{x}$ operator or the $\hat{p}$ operator. These exponentials can thus be turned into exponentials of ordinary numbers by inserting P completeness relations of the form:

$$\hat{\mathbb{1}} = \int \mathrm{d}^n p\,|p\rangle\,\langle p| \tag{2.21}$$

and $P-1$ completeness relations of the form:

$$\hat{\mathbb{1}} = \int \mathrm{d}^n x \, |x\rangle \langle x| , \tag{2.22}$$

where n is the total number of dimensions (for example, for a molecular system of M nuclei as considered in this work, this number would be $3M$). This yields:

$$\begin{aligned} \left\langle x \left| \exp\left(-\frac{i}{\hbar}\hat{H}t\right) \right| x' \right\rangle &= \lim_{P\to\infty} \int \mathrm{d}^n p_P \prod_{i=1}^{P-1} \mathrm{d}^n x_i \mathrm{d}^n p_i \\ &\times \prod_{j=1}^{P} \exp\left(\frac{-ip_j^2}{2m\hbar P}\right) \exp\left(\frac{-iU(x_{j-1})}{\hbar P}\right) \langle x_j \,|\, p_j \rangle \langle p_j \,|\, x_{j-1} \rangle , \end{aligned} \tag{2.23}$$

where it is understood that $x_P \equiv x$ and $x_0 \equiv x'$. Using

$$\langle x \,|\, p \rangle = \frac{1}{(2\pi\hbar)^{\frac{n}{2}}} \exp\left(\frac{i}{\hbar} p \cdot x\right) \tag{2.24}$$

one gets

$$\begin{aligned} &\left\langle x \left| \exp\left(-\frac{i}{\hbar}\hat{H}t\right) \right| x' \right\rangle \\ &= \lim_{P\to\infty} \int \mathrm{d}^n p_P \prod_{i=1}^{P-1} \frac{\mathrm{d}^n x_i \mathrm{d}^n p_i}{(2\pi\hbar)^{nP}} \prod_{j=1}^{P} \exp\left[\frac{i}{\hbar} p_j \cdot (x_j - x_{j-1}) - H(p_j, x_{j-1})\right]. \end{aligned} \tag{2.25}$$

This expression is known as the phase space path integral. For all forms of the potential $U(x)$ considered in this work, the function H in this expression is equal to the classical Hamiltonian. Note, however, that this is generally not the case — if a magnetic field is present, for example, this term may contain quantum corrections as a consequence of the *ordering ambiguity*. The final form of the path integral formulation is obtained by performing the P Gaussian integrations over the coordinates p_j:

$$\begin{aligned} &\left\langle x \left| \exp\left(-\frac{i}{\hbar}\hat{H}t\right) \right| x' \right\rangle \\ &= \lim_{P\to\infty} \int \prod_{i=1}^{P-1} \mathrm{d}^n x_i \left(\frac{mP}{2\pi i\hbar t}\right)^{\frac{nP}{2}} \exp\left\{\frac{i}{\hbar} \sum_{j=1}^{P} \left[\frac{mP}{2t}(x_j - x_{j-1})^2 - \frac{t}{P} U(x_{j-1})\right]\right\}. \end{aligned} \tag{2.26}$$

The expression in the exponential can be viewed as a discrete approximation to the classical action,

$$\lim_{P\to\infty} \sum_{j=1}^{P} \frac{t}{P} \left[\frac{m}{2}\left(\frac{x_j - x_{j-1}}{\frac{t}{P}}\right)^2 - U(x_{j-1})\right] = \int_0^t \mathrm{d}t' \left(\frac{1}{2} m\dot{x}^2 - V(x)\right) = S(t), \tag{2.27}$$

and the whole expression can be understood as an "integral over all possible *classical* paths starting from x at time 0 and ending at x' at time t" of $\exp\left(\frac{i}{\hbar}S(x(t))\right)$ (hence the name of the formulation) and is usually written compactly as

$$\int_{x(0)=x}^{x(t)=x'} \mathscr{D}x \exp\left(\frac{i}{\hbar}S(x(t))\right). \tag{2.28}$$

Note, however, that this is just a useful notation. After performing the Wick rotation the following is obtained:

$$\begin{aligned}&\left\langle x\middle|\exp\left(-\beta\hat{H}\right)\middle|x'\right\rangle \\ &= \lim_{P\to\infty}\int\prod_{i=1}^{P-1}\mathrm{d}^n x_i\left(\frac{mP}{2\pi\hbar^2\beta}\right)^{\frac{nP}{2}}\exp\left\{-\sum_{j=1}^{P}\left[\frac{mP}{2\hbar^2\beta}(x_j-x_{j-1})^2+\frac{\beta}{P}U(x_{j-1})\right]\right\},\end{aligned} \tag{2.29}$$

where $x_P = x$ and $x_0 = x'$. This can again be formally understood as

$$\int_{x(0)=x}^{x(\beta)=x'} \mathscr{D}x \exp\left(-S_\mathrm{E}(x(\beta))\right), \tag{2.30}$$

where β now plays the role of an imaginary time and S_E is the Euclidean action

$$S_\mathrm{E} = \int_0^\beta \mathrm{d}\tau L_\mathrm{E}, \tag{2.31}$$

$$L_\mathrm{E} = \frac{1}{2}m\dot{x}^2 + U(x) = H. \tag{2.32}$$

The equation (2.29) for the matrix elements of the canonical density operator is what is to be taken as the most relevant result of this derivation for this work.

Some important properties and consequences of equation (2.29) will now be shown. The partition function is calculated according to equation (2.14) as the trace of the density operator. Carrying out this trace in the position basis and using equation (2.29) one gets:

$$\begin{aligned}Z &= \int \mathrm{d}^n x\left\langle x\middle|\exp\left(-\beta\hat{H}\right)\middle|x\right\rangle \\ &= \lim_{P\to\infty}\int\prod_{i=1}^{P}\mathrm{d}^n x_i\left(\frac{mP}{2\pi\hbar^2\beta}\right)^{\frac{nP}{2}}\exp\left\{-\sum_{j=1}^{P}\left[\frac{mP}{2\hbar^2\beta}(x_j-x_{j-1})^2+\frac{\beta}{P}U(x_{j-1})\right]\right\},\end{aligned} \tag{2.33}$$

with a cyclic boundary condition $x_P = x_0$. Returning to the "integral over paths" interpretation of the current formulae, the partition function can thus

be understood as the integral of $\exp(-S_E)$ over all paths which return to their starting point after the imaginary time interval $\hbar\beta$, integrated over all possible starting points x. The mean of any operator $\hat{A}(\hat{x})$ can also be computed using this formula and equation 2.13 as

$$\begin{aligned}\langle\hat{A}\rangle &= \frac{1}{Z}\int \mathrm{d}^n x A(x)\left\langle x\middle|\exp\left(-\beta\hat{H}\right)\middle|x\right\rangle \\ &= \frac{1}{Z}\lim_{P\to\infty}\int\prod_{i=1}^{P}\mathrm{d}^n x_i A(x_P)\left(\frac{mP}{2\pi\hbar^2\beta}\right)^{\frac{nP}{2}} \\ &\times\exp\left\{-\sum_{j=1}^{P}\left[\frac{mP}{2\hbar^2\beta}(x_j-x_{j-1})^2+\frac{\beta}{P}U(x_{j-1})\right]\right\},\end{aligned} \tag{2.34}$$

again with the condition $x_P = x_0$.

Next, the *centroid* will be defined as

$$x_c = \frac{1}{\beta}\int_0^\beta d\tau\, x(\tau), \tag{2.35}$$

which can also equivalently be written as

$$x_c = \lim_{P\to\infty}\frac{1}{P}\sum_{i=1}^{P} x_i. \tag{2.36}$$

This quantity corresponds to the average position of the particle along its path through imaginary time. Using this quantity, the equation 2.33 can be formally rewritten as

$$Z = \int \mathrm{d}^n x\left(\frac{m}{2\pi\hbar^2\beta}\right)^{\frac{n}{2}}\exp\left(-\beta W(x)\right), \tag{2.37}$$

where

$$W(x) = -\frac{1}{\beta}\ln\left(\rho_c(x)\right), \tag{2.38}$$

$$\begin{aligned}\rho_c(x) &= \lim_{P\to\infty}\int\prod_{i=1}^{P}\mathrm{d}^n x_i P^{\frac{n}{2}}\left(\frac{mP}{2\pi\hbar^2\beta}\right)^{\frac{n(P-1)}{2}}\delta(x_c(x_i)-x) \\ &\times\exp\left\{-\sum_{j=1}^{P}\left[\frac{mP}{2\hbar^2\beta}(x_j-x_{j-1})^2+\frac{\beta}{P}U(x_{j-1})\right]\right\}.\end{aligned} \tag{2.39}$$

Here, $\rho_c(x)$ is the centroid probability distribution and $W(x)$ is its Boltzmann inverse — it is thus a free energy of the system with the centroid fixed at $x_c = x$. Note that this formal substitution turns the problem of quantum

statistics into a rather ordinary problem of classical statistics, of the same dimensionality. This is only a formal simplification, as practically obtaining the exact free energy surface $W(x)$ amounts to solving the original problem. This approach can, however, prove useful as the free energy surface $W(x)$ can either be approximated [13] or as is newly proposed in this work – numerically constructed. This will be explained in detail in later sections. Later in subsection 2.4, the centroid and the free energy surface $W(x)$ will also be shown to play a significant role in approximating quantum dynamics.

As a last note — the issue of exchange symmetry of the particles has not been discussed yet. As it turns out, though, for molecular systems at reasonable temperatures, the atomic nuclei can be safely approximated as distinguishable particles (as has been assumed so far). For more details regarding this fact, see appendix C.

2.3 Linear response theory

This subsection will review mostly established knowledge. For the classical formulation, this knowledge has been obtained by the author at the lecture "Vybrané kapitoly z nerovnovážné statistické fyziky I" taught by RNDr. Karel Netočný, Ph.D. For the quantum formulation the derivation, as presented, was put together by the author based on the books 1 and 23, as well as the article 24.

Having established an equilibrium statistical description of the systems at hand, attention will now be directed to the dynamical behavior of these systems. The primary focus will be given to the dynamics of macroscopic observables. In equilibrium, these observables can be calculated using the methods of the preceding chapter. The goal will thus be to establish how the dynamics of the macroscopic observables of the system can be calculated from the knowledge of the microscopic dynamics and the equilibrium statistical description. In general, this is the topic of *non-equilibrium* statistical mechanics.

A useful framework developed within this field to treat such problems is *linear response theory*. Within this framework, the system is assumed to be perturbed from equilibrium by a time-dependent force weak enough to keep only terms linear in this force in the total Hamiltonian:

$$H = H_0(\mathbb{X}) - B(\mathbb{X})F(t). \tag{2.40}$$

Here, $B(\mathbb{X})$ is a function on the phase space (the minus sign being a convention) and H_0 is a time-independent Hamiltonian (which determines the equilibrium). The symbol $\mathbb{X}$ denotes the phase space coordinates of the system. It is more common to work at the classical level in linear response theory, however, it can still be formulated at the quantum level — the functions H, H_0,

and B becoming operators. The following discussion will start at the classical level and show several relevant formulae such as the famous *fluctuation–dissipation theorem*. Afterward, the quantum formulation will be discussed.

2.3.1 Classical formulation

As was stated previously, within linear response theory, only terms linear in the perturbing force are kept in the total Hamiltonian (equation 2.40). Without loss of generality, the mean value of some macroscopic observable A at some time t can be written as

$$\langle \Delta A(t) \rangle = \int_{-\infty}^{\infty} \mathrm{d}s R_{AB}(t,s) F(s), \tag{2.41}$$

where R_{AB} is called the response function and ΔA is the deviation of the quantity A from its equilibrium value. Assuming time homogeneity, the following must hold:

$$R_{AB}(t,s) = R_{AB}(t-s,0) = R_{AB}(0,s-t) \equiv R_{AB}(t-s). \tag{2.42}$$

Furthermore, should R_{AB} respect causality, the following must hold also:

$$R_{AB}(t-s) = 0; \qquad \forall\, t-s < 0. \tag{2.43}$$

To show how the response function R_{AB} can be calculated from the knowledge of equilibrium properties of the system, consider the following special $F(t)$:

$$F(t) = F\theta(-t). \tag{2.44}$$

In such a case, $\langle \Delta A \rangle$ does not depend on time for $t < 0$ and can be expressed as

$$\langle \Delta A \rangle = \frac{1}{Z} \int \mathrm{d}\Omega(\mathbb{X}) \Delta A(\mathbb{X}) \exp\big[-\beta \left(H_0(\mathbb{X}) - B(\mathbb{X})\, F\right)\big], \tag{2.45}$$

$\mathrm{d}\Omega$ being the normalized phase-space volume element. Keeping only orders linear in F, this reduces to

$$\begin{aligned} \langle \Delta A \rangle &= \frac{1}{Z}\beta \int \mathrm{d}\Omega(\mathbb{X}) F B(\mathbb{X}) \Delta A(\mathbb{X}) \exp\big(-\beta H_0\big) = \beta F \langle \Delta A(\mathbb{X}) B(\mathbb{X}\rangle_0 \\ &= \beta F \langle \Delta A(\mathbb{X}) \Delta B(\mathbb{X}) \rangle_0. \end{aligned} \tag{2.46}$$

The brackets $\langle A \rangle_0$ denote an average with respect to the equilibrium ensemble. The last equality follows from $\langle \Delta A \rangle_0 = 0$:

$$\langle \Delta A(\mathbb{X}) B(\mathbb{X}) \rangle_0 = \langle \Delta A(\mathbb{X}) \Delta B(\mathbb{X}) \rangle_0 + B_0 \langle \Delta A(\mathbb{X}) \rangle_0 = \langle \Delta A(\mathbb{X}) \Delta B(\mathbb{X}) \rangle_0. \tag{2.47}$$

B_0 has been used to denote the equilibrium average of B. At the same time, it should hold for $t < 0$ that

$$\langle \Delta A \rangle = \int_0^{\infty} \mathrm{d}s R_{AB}(s) F; \qquad \forall\, t < 0, \tag{2.48}$$

implying that

$$\int_0^\infty \mathrm{d}s R_{AB}(s) = \frac{1}{\beta}\langle \Delta A(\mathbb{X})\Delta B(\mathbb{X})\rangle_0 \tag{2.49}$$

for $t < 0$.

In the case of $t \geq 0$, making use of the fact that the distribution in $t = 0$ is an equilibrium one, determined by $H(\mathbb{X})$:

$$\langle \Delta A(t)\rangle = \int \mathrm{d}\Omega(\mathbb{X})\Delta A(\mathbb{Y})P(\mathbb{X}(t) = \mathbb{Y}|\mathbb{X}(0) = \mathbb{X})\exp\big[-\beta(H_0 - B(\mathbb{X})F)\big]. \tag{2.50}$$

In this equation, $P(\mathbb{X}(t) = B|\mathbb{X}(0) = A)$ is used to denote the conditional probability that the system will be found at the phase-space point B at time t if it is known to have been at the point A at time 0. Keeping again only terms linear in F and using $\langle \Delta A(t)\rangle_0 = 0$ yields

$$\begin{aligned}\langle \Delta A(t)\rangle &= \beta \int \mathrm{d}\Omega(\mathbb{X})FB(\mathbb{X})\Delta A(\mathbb{Y})P(\mathbb{X}(t) = \mathbb{Y}|\mathbb{X}(0) = \mathbb{X})\exp\big(-\beta H_0\big) \\ &= \beta F\langle \Delta A(t)B(0)\rangle_0 = \beta F\langle \Delta A(t)\Delta B(0)\rangle_0.\end{aligned} \tag{2.51}$$

Comparing this again to the response function prescription

$$\langle \Delta A(t)\rangle = \int_0^\infty \mathrm{d}s R_{AB}(s)F(s) = F\int_t^\infty \mathrm{d}s R_{AB}(s) \tag{2.52}$$

yields

$$\int_t^\infty \mathrm{d}s R_{AB}(s) = \frac{1}{\beta}\langle \Delta A(t)\Delta B(0)\rangle_0. \tag{2.53}$$

Taken together with equation 2.49, the final result can be written as:

$$R_{AB}(t) = -\beta\theta(t)\frac{\mathrm{d}C_{AB}(t,0)}{\mathrm{d}t}, \tag{2.54}$$

where $\langle \Delta A(t)\Delta B(0)\rangle_0$ has been denoted as $C_{AB}(t,0)$ and will be from now on referred to as a *time correlation function*. Due to the linearity of the whole procedure in F, this relation holds not only for the special form of $F(t)$ considered above but for any $F(t)$. This result is known as the *fluctuation–dissipation theorem*. The theorem establishes a relation between the response of a system to a weak external perturbation (encoded in R_{AB}) and the equilibrium fluctuations of the system (encoded in C_{AB}). The time correlation functions have the following properties. Assuming time homogeneity, the time correlation functions are functions of one time variable only:

$$C_{AB}(t,s) = C_{AB}(t-s,0) \equiv C_{AB}(t-s). \tag{2.55}$$

This further implies that

$$C_{AB}(t) = \langle \Delta A(t)\Delta B(0)\rangle_0 = \langle \Delta A(0)\Delta B(-t)\rangle_0 = C_{BA}(-t). \tag{2.56}$$

Furthermore, it follows from the time-reversal symmetry of the microscopic dynamics that

$$C_{AB}(t) = (C_{AB}(t))^* = C_{B^*A^*}(t), \tag{2.57}$$

where $*$ denotes time inversion which in the classical case corresponds to the substitution $A(x,p)^* = A(x,-p)$. The last equality follows from the time inversion of equation 2.51. Assuming that the quantities A and B change under time reversal in the simple manner:

$$Q^* = \epsilon_Q Q; \qquad \epsilon_Q = \begin{cases} +1 \\ -1 \end{cases}, \tag{2.58}$$

ϵ being the *parity* of the observable, this equation takes the form

$$C_{AB}(t) = \epsilon_A \epsilon_B C_{BA}(t). \tag{2.59}$$

Together with equation (2.56), this yields

$$C_{AB}(t) = \epsilon_A \epsilon_B C_{AB}(-t). \tag{2.60}$$

As a consequence of this, time *autocorrelation* functions — time correlation functions with $A = B$ — of quantities with well-defined parity are even functions of time. A property expected from time correlation functions on physical grounds is

$$\langle A(t)B(0)\rangle_0 \overset{t\to\infty}{=} \langle A\rangle_0 \langle B\rangle_0, \tag{2.61}$$

meaning that the quantities A and B become decorrelated for large time separations. This then further implies:

$$\langle \Delta A(t) \Delta B(0)\rangle_0 \overset{t\to\infty}{=} 0. \tag{2.62}$$

In general, multiple perturbing forces may be present, or a perturbing force may be a vector. Even though the present discussion has worked with only a single quantity A and a single quantity B, the generalization of the framework to multiple variables A and multiple variables B (corresponding to multiple weakly perturbing forces or a multicomponent one) is straightforward:

$$\langle A_i(t)\rangle = \sum_{j=1}^{N} \int_0^\infty \mathrm{d}s R_{A_i B_j}(s) F_{B_j}(t-s). \tag{2.63}$$

The matrix elements $R_{A_i B_j}$ are in relation to the time correlation functions $C_{A_i B_j}$ according to the fluctuation–dissipation theorem. Another useful set of relations can be derived from the fluctuation–dissipation theorem. Working again with just a single A and B (keeping in mind the straightforward extension to multiple variables), which are assumed to have the same parity, one can define:

$$j_B = \frac{\mathrm{d}B}{\mathrm{d}t}, \tag{2.64}$$

and similarly j_A. It then holds that

$$\langle j_A \rangle_0 = \frac{\mathrm{d}\langle A \rangle_0}{\mathrm{d}t} = 0, \tag{2.65}$$

$$\langle j_A \rangle = \int_{-\infty}^{t} \mathrm{d}s R^j_{AB}(t-s)F(s), \tag{2.66}$$

$$R^j_{AB} = \frac{\mathrm{d}R_{AB}}{\mathrm{d}t}, \tag{2.67}$$

combining this with the fluctuation–dissipation theorem (equation (2.54)) one gets

$$\begin{aligned} R^j_{AB}(t-s) &= \frac{\mathrm{d}}{\mathrm{d}t}(-\beta\theta(t-s)\frac{\mathrm{d}}{\mathrm{d}t}C_{AB}(t-s)) = -\beta\delta(t-s)\frac{\mathrm{d}}{\mathrm{d}t}C_{AB}(t-s) \\ &+ \beta\frac{\mathrm{d}^2}{\mathrm{d}t\mathrm{d}s}C_{AB}(t-s) = \beta\frac{\mathrm{d}^2}{\mathrm{d}t\mathrm{d}s}\langle \Delta A(t)\Delta B\rangle_0 = \beta\langle j_A(t)\, j_B(s)\rangle_0 \\ &\equiv \beta C^j_{AB}(t-s). \end{aligned} \tag{2.68}$$

The term with $\delta(t-s)\frac{\mathrm{d}}{\mathrm{d}t}C_{AB}(t-s)$ is zero, because the function C_{AB} is even (due to the quantities A and B having the same parity), which implies its derivative is odd and thus is zero for $t-s=0$. This result is known as the *Green–Kubo relations* and will be important later.

An important application of the fluctuation–dissipation theorem is the study of the dynamics of a Brownian particle. For simplicity, work will be carried out in one dimension, the generalization to more dimensions being again straightforward. A Brownian particle is a particle immersed in a heat bath, which exerts a force on the particle. This force can be split into a systematic part, and a random part with a mean value equal to zero (corresponding to thermal noise):

$$f_B(X(t)) = \langle f_B \rangle (X(t)) + \Delta f_B, \tag{2.69}$$

$X(t)$ denoting the particle position. The systematic part can be calculated using the fluctuation–dissipation theorem. The system considered to be weakly perturbed from equilibrium will be the heat bath in direct contact with the particle, and the weak "force" will be the particle position:

$$H = H_0 + f_B(X(t))(X(s) - X(t)). \tag{2.70}$$

Using linear response theory and the fluctuation-dissipation theorem yields

$$\langle f_B \rangle = \langle f_B \rangle_0 + \int_0^\infty \mathrm{d}s \left(-\beta\theta(s)\frac{\mathrm{d}C_{ff}(s)}{\mathrm{d}s} \right)(X(t-s) - X(t)). \tag{2.71}$$

The equilibrium average $\langle f_B \rangle_0$ is zero for a homogeneous bath, the second term can be integrated by parts:

$$\langle f_B \rangle = -\beta \int_0^\infty \mathrm{d}s \theta(s) C_{ff}(s) \dot{X}(t-s). \tag{2.72}$$

The boundary terms do not contribute as for $s = 0$ $X(t) - X(t) = 0$ and for $s = \infty$ $C_{ff} = 0$. Having obtained the expression for the systematic part of the force exerted on the Brownian particle by the bath, the equations of motion can be put together:

$$\begin{aligned} \dot{X} &= \frac{P}{m}, \\ \dot{P} &= -\frac{\partial U}{\partial X} - \int_0^\infty \mathrm{d}s\Gamma(s)\dot{X}(t-s) + \xi(t), \end{aligned} \tag{2.73}$$

where $\Gamma(s) = \beta\theta(s)C_{ff}(s)$ is the *memory kernel* and $\xi(t)$ is a random noise term with properties consistent with Δf_B:

$$\begin{aligned} \langle\xi(t)\rangle &= 0, \\ \langle\xi(t)\xi(s)\rangle &= C_{ff}(t-s). \end{aligned} \tag{2.74}$$

The second equation of equations 2.73 is called the *generalized Langevin equation* (GLE). If the velocity of the Brownian particle changes little over the characteristic relaxation time of C_{ff} (a timescale on which C_{ff} decays to zero), then the second term in this equation may be further approximated as

$$\int_0^\infty \mathrm{d}s\Gamma(s)\dot{X}(t-s) \approx \dot{X}(t)\int_0^\infty \mathrm{d}s\Gamma(s) \equiv \gamma\dot{X}(t), \tag{2.75}$$

which is the same as writing $\Gamma(s) = 2\gamma\delta(s) \implies C_{ff}(s) = \frac{2\gamma}{\beta}\delta(s)$ (the factor 2 is present due to $\int_0^\infty \mathrm{d}s\Gamma(s) = \frac{1}{2}\int_{-\infty}^\infty \mathrm{d}s\Gamma(s) = \gamma$). The stochastic term then has the following statistical properties:

$$\begin{aligned} \langle\xi(t)\rangle &= 0 \\ \langle\xi(t)\xi(s)\rangle &= C_{ff}(t-s) = 2\frac{\gamma}{\beta}\delta(t-s), \end{aligned} \tag{2.76}$$

implying it is a *white noise* term. This approximation is called the *memoryless* or *Markovian* approximation, and it results in the *Langevin equation* (LE) [25]:

$$\dot{P} = -\frac{\partial U}{\partial X} - \gamma\dot{X} + \xi(t). \tag{2.77}$$

This equation has the formal solution:

$$\exp\Big(-\frac{\gamma}{m}s\Big)\frac{\mathrm{d}}{\mathrm{d}s}\Big[P(s)\exp\Big(\frac{\gamma}{m}s\Big)\Big] = \xi(s) - \frac{\partial U(X(s))}{\partial X}, \tag{2.78}$$

or equivalently

$$P(t) = P(0)\exp\Big(-\frac{\gamma}{m}t\Big) + \int_0^t \mathrm{d}s\exp\Big[-\frac{\gamma}{m}(t-s)\Big]\Big(\xi(s) - \frac{\partial U(X(s))}{\partial X}\Big). \tag{2.79}$$

which has implications concerning the statistics of $P(t)$. Assuming the particle is free ($\frac{\partial U(X(t))}{\partial X} = 0$) yields

$$\langle P(t)\rangle = P(0)\exp\left(-\frac{\gamma}{m}t\right) \tag{2.80}$$

alongside

$$\langle P^2(t)\rangle = P^2(0)\exp\left(-2\frac{\gamma}{m}t\right) + \frac{m}{\beta}\left[1-\exp\left(-2\frac{\gamma}{m}t\right)\right], \tag{2.81}$$

which for $t >> \frac{m}{\gamma}$ gives

$$\langle P^2(t)\rangle = \frac{m}{\beta} \implies \left\langle \frac{P^2}{2m}\right\rangle = \frac{1}{2}\frac{1}{\beta}. \tag{2.82}$$

This implies that the distribution of the momenta approaches the Maxwell-Boltzmann distribution. Another consequence of equations 2.80 and 2.81 is that the quantity $P(t)$ is a normal distributed random number with a mean value $P(0)\exp\left(-\frac{\gamma}{m}t\right)$ and variance equal to $\frac{m}{\beta}\left(1-\exp\left(-2\frac{\gamma}{m}t\right)\right)$. The formal solution to the Langevin equation for a free particle can thus be written as:

$$P(t) = P(0)\exp\left(-\frac{\gamma}{m}t\right) + \sqrt{\left\{\frac{m}{\beta}\left[1-\exp\left(-2\frac{\gamma}{m}t\right)\right]\right\}}\zeta(t), \tag{2.83}$$

$\zeta(t)$ being a normal distributed random number with unit variance and zero mean. These points will become important later in subsection 2.4.

Another important special case of a time-dependent external perturbation within linear response theory is a harmonic potential:

$$F(t) = F\cos(\omega t) = F\,\mathrm{Re}(\exp(i\omega t)). \tag{2.84}$$

Using the formulae of linear response theory the following is obtained:

$$\begin{aligned}\langle \Delta\hat{A}\rangle &= F\,\mathrm{Re}\left[\int_{-\infty}^{\infty} \mathrm{d}s R_{AB}(s)\exp(i\omega(t-s))\right] \\ &= F\,\mathrm{Re}(R_{AB}(\omega))\cos(\omega t) - F\,\mathrm{Im}(R_{AB}(\omega))\sin(\omega(t)),\end{aligned} \tag{2.85}$$

where $R_{AB}(\omega)$ is the Fourier transform of $R_{AB}(t)$. From linear response theory, it is also known that

$$R_{AB}(t) = -\beta\theta(t)\frac{\mathrm{d}C_{AB}(t)}{\mathrm{d}t} \equiv \theta(t)\tilde{R}_{AB}(t), \tag{2.86}$$

where $\tilde{R}_{AB}$ is a real function. The presence of the $\theta(t)$ follows from causality, and the presence of this function implies that the following relations hold for the real and imaginary parts of $R_{AB}(\omega)$:

$$\begin{aligned}\mathrm{Re}(R_{AB}(\omega)) &= \frac{1}{\pi}\mathscr{P}\int_{-\infty}^{\infty}\mathrm{d}\omega'\frac{\mathrm{Im}(R_{AB})(\omega')}{\omega-\omega'} \\ \mathrm{Im}(R_{AB}(\omega)) &= -\frac{1}{\pi}\mathscr{P}\int_{-\infty}^{\infty}\mathrm{d}\omega'\frac{\mathrm{Re}(R_{AB})(\omega')}{\omega-\omega'},\end{aligned} \tag{2.87}$$

which are known as the *Kramers–Kronig relations*. The symbol $\mathscr{P}$ is used to denote a principal value integral. The function $\mathrm{Im}(R_{BB}(\omega))$ (B being the quantity "coupled" to the external force) has a thermodynamic interpretation — it is proportional to the mean heat dissipated to the system by a harmonic external force of frequency ω averaged over a single period:

$$\left\langle \frac{\mathrm{d}'Q}{\mathrm{d}t} \right\rangle = \frac{1}{2} F^2 \omega \,\mathrm{Im}(R_{BB}(\omega)). \tag{2.88}$$

The derivative on the left hand side is primed as a reminder of the non-potential nature of heat. Using integration by parts and equation (2.54), this relation can also be brought into the form:

$$\left\langle \frac{\mathrm{d}'Q}{\mathrm{d}t} \right\rangle = \frac{1}{4} \beta F^2 \omega^2 C_{BB}(\omega), \tag{2.89}$$

or equivalently:

$$\left\langle \frac{\mathrm{d}'Q}{\mathrm{d}t} \right\rangle = \frac{1}{4} \beta F^2 C_{j_B j_B}(\omega). \tag{2.90}$$

2.3.2 Quantum formulation

In the quantum case, the approach is similar, with only terms linear in the perturbing force kept in the Hamiltonian:

$$\hat{H} = \hat{H}_0 - \hat{B} F(t), \tag{2.91}$$

with the linear response prescription (equation (2.41)) still holding:

$$\langle \Delta \hat{A} \rangle(t) = \int_0^\infty \mathrm{d}s R_{AB}(s) F(t-s), \tag{2.92}$$

the only difference is that the left-hand side is an expectation value of an operator (over a statistical ensemble of states). It can be expected from the classical case, that the response function R_{AB} can be calculated from a quantum analogy of the time correlation function C_{AB}. To show this, alongside what this analogy should be, the special $F(t)$ used in the classical case — $F(t) = \theta(-t)F$ will be used. For $t < 0$ the average does not depend on time again and is equal to:

$$\langle \Delta \hat{A} \rangle = \frac{1}{Z} \mathrm{Tr}\{\Delta \hat{A} \exp[-\beta(\hat{H}_0 - \hat{B}F)]\} = \frac{\mathrm{Tr}\{\Delta \hat{A} \exp[-\beta(\hat{H}_0 - \hat{B}F)]\}}{\mathrm{Tr}\{\exp[-\beta(\hat{H}_0 - \hat{B}F)]\}}. \tag{2.93}$$

As the operators $\hat{H}_0$ and $\hat{B}$ do not generally commute, the exponential cannot be simply expanded as in the classical case. Instead, the following formula [23] will be used:

$$\exp[-\beta(\hat{H}_0 - \hat{B}F)] = \exp(-\beta \hat{H}_0) \left\{ 1 + F \int_0^\beta \mathrm{d}\lambda \exp(\lambda \hat{H}_0) \hat{B} \exp[-\lambda(\hat{H}_0 - \hat{B}F)] \right\}, \tag{2.94}$$

which reduces, keeping only terms linear in F, to

$$\exp\left[-\beta(\hat{H}_0-\hat{B}F)\right]=\exp\left(-\beta\hat{H}_0\right)\left[1+F\int_0^\beta \mathrm{d}\lambda\exp\left(\lambda\hat{H}_0\right)\hat{B}\exp\left(-\lambda\hat{H}_0\right)\right]. \tag{2.95}$$

The expression for the expectation value thus becomes

$$\langle\Delta\hat{A}\rangle=\frac{\mathrm{Tr}\left\{\Delta\hat{A}\exp\left(-\beta\hat{H}_0\right)\left[1+F\int_0^\beta \mathrm{d}\lambda\exp\left(\lambda\hat{H}_0\right)\hat{B}\exp\left(-\lambda\hat{H}_0\right)\right]\right\}}{\mathrm{Tr}\left\{\exp\left(-\beta\hat{H}_0\right)\left[1+F\int_0^\beta \mathrm{d}\lambda\exp\left(\lambda\hat{H}_0\right)\hat{B}\exp\left(-\lambda\hat{H}_0\right)\right]\right\}}. \tag{2.96}$$

Keeping again only terms linear in F and using that by definition $\mathrm{Tr}\left[\Delta\hat{A}\exp\left(-\beta\hat{H}_0\right)\right]=0$, this equation becomes

$$\begin{aligned}\langle\Delta\hat{A}\rangle&=\frac{F}{Z_0}\mathrm{Tr}\left[\Delta\hat{A}\exp\left(-\beta\hat{H}_0\right)\int_0^\beta \mathrm{d}\lambda\exp\left(\lambda\hat{H}_0\right)\Delta\hat{B}\exp\left(-\lambda\hat{H}_0\right)\right]\\&=\frac{F}{Z_0}\mathrm{Tr}\left\{\left[\int_0^\beta \mathrm{d}\lambda\exp\left(\lambda\hat{H}_0\right)\Delta\hat{B}\exp\left(-\lambda\hat{H}_0\right)\right]\Delta\hat{A}\exp\left(-\beta\hat{H}_0\right)\right\}.\end{aligned} \tag{2.97}$$

For $t\geq 0$, using the fact that at $t=0$ $\hat{\rho}=\frac{1}{Z}\exp\left(-\beta(\hat{H}_0)-\hat{B}F\right)$ and that given a time-independent Hamiltonian, the density operator evolves according to:

$$\hat{\rho}(t)=\exp\left(-\frac{i}{\hbar}\hat{H}t\right)\hat{\rho}(0)\exp\left(\frac{i}{\hbar}\hat{H}t\right), \tag{2.98}$$

the following is obtained:

$$\begin{aligned}\langle\Delta\hat{A}\rangle(t)&=\frac{1}{Z}\mathrm{Tr}\left\{\Delta\hat{A}\exp\left(-\frac{i}{\hbar}\hat{H}_0t\right)\exp\left[-\beta(\hat{H}_0-\hat{B}F)\right]\exp\left(\frac{i}{\hbar}\hat{H}_0t\right)\right\}\\&=\frac{1}{Z}\mathrm{Tr}\left\{\exp\left(\frac{i}{\hbar}\hat{H}_0t\right)\Delta\hat{A}\exp\left(-\frac{i}{\hbar}\hat{H}_0t\right)\exp\left[-\beta(\hat{H}_0-\hat{B}F)\right]\right\}\\&=\frac{1}{Z}\mathrm{Tr}\{\Delta\hat{A}(t)\exp\left[-\beta(\hat{H}_0-\hat{B}F)\right]\},\end{aligned} \tag{2.99}$$

where $\Delta\hat{A}(t)$ now denotes a Heisenberg operator evolved to time t. Keeping only terms linear in F and using the same formulae as in the $t<0$ case yields

$$\langle\Delta\hat{A}\rangle(t)=\frac{F}{Z_0}\mathrm{Tr}\left\{\left[\int_0^\beta \mathrm{d}\lambda\exp\left(\lambda\hat{H}_0\right)\Delta\hat{B}\exp\left(-\lambda\hat{H}_0\right)\right]\Delta\hat{A}(t)\exp\left(-\beta\hat{H}_0\right)\right\}. \tag{2.100}$$

A new function $K_{BA}(t)$ can thus be defined as

$$K_{BA}(t)=\frac{1}{\beta Z_0}\mathrm{Tr}\left\{\left[\int_0^\beta \mathrm{d}\lambda\exp\left(\lambda\hat{H}_0\right)\Delta\hat{B}\exp\left(-\lambda\hat{H}_0\right)\right]\Delta\hat{A}(t)\exp\left(-\beta\hat{H}_0\right)\right\}, \tag{2.101}$$

which will be referred to as the *Kubo-transformed* time correlation function. Comparing the obtained results with the linear response prescription 2.92, one obtains similar expressions to the classical case:

$$\begin{aligned}\langle \Delta \hat{A} \rangle(t) &= \int_0^\infty \mathrm{d}s R_{AB}(s) F(t-s) = F \int_t^\infty \mathrm{d}s R_{AB}(s) \\ &= F\beta K_{BA}(t),\end{aligned} \tag{2.102}$$

which implies

$$R_{AB} = -\beta\theta(t)\frac{\mathrm{d}K_{BA}(t)}{\mathrm{d}t}. \tag{2.103}$$

The only difference is that the classical time correlation function is replaced by the quantum Kubo-transformed correlation function. Note that the Kubo-transformed time correlation function shares many properties of the classical time correlation function [8]. In particular, a property that was already assumed in the definition 2.101 is stationarity (if time homogeneity is also assumed):

$$K_{BA}(t,0) = K_{BA}(t+s,s) = K_{BA}(t). \tag{2.104}$$

Furthermore, it holds that

$$K_{BA}(t) = K_{AB}(-t), \tag{2.105}$$

and if $\hat{A}$ and $\hat{B}$ are Hermitian operators

$$K_{AB}(-t) = K_{BA}(t)^*. \tag{2.106}$$

Both equations can be verified by direct computation. Together, these two equations imply that $K_{BA}(t)$ is real. Equation 2.105 also implies that a Kubo-transformed autocorrelation function is even, as is the classical autocorrelation function. For a time-inversion symmetric Hamiltonian, it is also true that

$$K_{BA}(t) = \epsilon_A \epsilon_B K^*_{BA}(-t), \tag{2.107}$$

where ϵ_Q is again the operator parity, now defined as:

$$\hat{\Theta}\hat{Q}\hat{\Theta}^{-1} = \epsilon_Q \hat{Q} \qquad \epsilon_Q = \begin{cases} +1 \\ -1 \end{cases}, \tag{2.108}$$

Θ being the time inversion operator. Together with the previous two equations, this implies that the Kubo-transformed time correlation function is generally even for $\hat{A}$ and $\hat{B}$ of the same parity. Although the Kubo-transformed time correlation function is the most useful quantum equivalent of the classical correlation function in the context of the present work, multiple non-equivalent candidate analogues of the classical time correlation function exist in the quantum case. This is due to the (general) non-commutativity of $\hat{A}$

and $\hat{B}$. In particular using the formula [24]

$$\left[\exp\left(-\beta\hat{H}_0\right),\hat{A}\right]=\exp\left(-\beta\hat{H}_0\right)\int_0^\beta \mathrm{d}\lambda\frac{\mathrm{d}\hat{A}(-i\hbar\lambda)}{\mathrm{d}t}, \tag{2.109}$$

and using

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}K_{BA}&=\frac{1}{\beta Z_0}\mathrm{Tr}\left\{\left[\int_0^\beta \mathrm{d}\lambda\exp\left(\lambda\hat{H}_0\right)\Delta\hat{B}\exp\left(-\lambda\hat{H}_0\right)\right]\frac{\mathrm{d}\Delta\hat{A}(t)}{\mathrm{d}t}\exp\left(-\beta\hat{H}_0\right)\right\}\\&=\frac{1}{\beta Z_0}\mathrm{Tr}\left[\int_0^\beta \mathrm{d}\lambda\Delta\hat{B}(-i\hbar\lambda)\frac{\mathrm{d}\Delta\hat{A}(t)}{\mathrm{d}t}\exp\left(-\beta\hat{H}_0\right)\right]\\&=-\frac{i}{\hbar\beta Z_0}\mathrm{Tr}\left[\int_0^\beta \mathrm{d}\lambda\frac{\mathrm{d}\Delta\hat{B}(-i\hbar\lambda)}{\mathrm{d}t}\Delta\hat{A}(t)\exp\left(-\beta\hat{H}_0\right)\right],\end{aligned} \tag{2.110}$$

one obtains

$$\begin{aligned}R_{AB}(t)&=\frac{i}{\hbar Z_0}\mathrm{Tr}\left(\left[\exp\left(-\beta\hat{H}_0\right),\Delta\hat{B}\right]\Delta\hat{A}(t)\right)\\&=\frac{i}{\hbar Z_0}\mathrm{Tr}\left[\left[\Delta\hat{B},\Delta\hat{A}(t)\right]\exp\left(-\beta\hat{H}_0\right)\right]=\frac{i}{\hbar}\left\langle\left[\Delta\hat{B},\Delta\hat{A}(t)\right]\right\rangle_0.\end{aligned} \tag{2.111}$$

This expression can be commonly found in literature and it presents another candidate equivalent of the classical time correlation function — the commutator of $\Delta\hat{B}$ and $\Delta\hat{A}(t)$. The usefulness of the Kubo-transformed time correlation function comes from the fact that it shares the same properties and symmetries as the classical time-correlation function. Within linear response theory, the transition from classical to quantum description is done simply by changing $C_{AB}\rightarrow K_{AB}$ and vice versa. For instance, using 2.102 and 2.54 the quantum equivalent of the Green-Kubo relation can be immediately written down as

$$R^j_{AB}(t)=\beta K^j_{AB}(t), \tag{2.112}$$

where K^j_{AB} is a Kubo-transformed time correlation function of $\hat{j}_A=\frac{\mathrm{d}\hat{A}}{\mathrm{d}t}$ and $\hat{j}_B=\frac{\mathrm{d}\hat{B}}{\mathrm{d}t}$. The operators $\hat{A}$ and $\hat{B}$ are again assumed to have the same parity.

An application of linear response theory that is of particular interest in the context of the present work, is the theoretical modeling of *vibrational spectroscopy*. Several different experimental methods fall in the category of vibrational spectroscopy. The theoretical description of these different techniques is similar and will be illustrated on the example of *infrared absorption spectroscopy*.

In infrared spectroscopy, a sample is irradiated by a weak (in comparison to the energy scales of intermolecular and intramolecular interactions) electromagnetic radiation of a long wavelength (again, compared to the characteristic length scales present in the system). The measured quantity in such

an experiment is the portion of the energy flux of the radiation absorbed by the sample as a function of frequency ω.

This experimental setup can be theoretically described using linear response theory, taking the weak perturbation to be the electric component of an electromagnetic wave in the long-wave limit:

$$\hat{H} = \hat{H}_0 - \mu \cdot E \cos(\omega t), \tag{2.113}$$

where μ is the total dipole moment of the system. The direction of the electric field can be taken without loss of generality to be in the z direction. The portion of the radiation absorbed can be calculated using equation (2.90):

$$\left\langle \frac{\mathrm{d}'Q}{\mathrm{d}t} \right\rangle = \frac{1}{4}\beta E^2 C_{\dot{\mu}_z \dot{\mu}_z}(\omega). \tag{2.114}$$

For an isotropic system, all spatial directions are equal and thus

$$C_{\dot{\mu}_z \dot{\mu}_z}(\omega) = \frac{1}{3} C_{\dot{\mu}\dot{\mu}}(\omega). \tag{2.115}$$

If it is the case that the constituent particles can be assigned a definite charge, the total dipole moment can be decomposed into contributions from the individual particles as

$$\dot{\mu} = \sum_{i=1}^{N} e_i v_i \tag{2.116}$$

and thus

$$\begin{aligned} C_{\dot{\mu}\dot{\mu}}(\omega) &= \int_{-\infty}^{\infty} \mathrm{d}t \exp(-i\omega t) \langle \dot{\mu}_z(t) \dot{\mu}_z(0) \rangle_0 \\ &= \sum_{i=1}^{N} \sum_{j=1}^{N} e_i e_j \int_{-\infty}^{\infty} \mathrm{d}t \exp(-i\omega t) \langle v_{zi} v_{zj} \rangle_0 = \sum_{i=1}^{N} \sum_{j=1}^{N} e_i e_j C_{v_i v_j}(\omega). \end{aligned} \tag{2.117}$$

Note, however, that this can be used only rarely and was not in fact used in this work, as the particles generally cannot be ascribed a well-defined charge. However, the behavior of the particle velocities still remains relevant, as the dipole can be, without loss of generality, decomposed using:

$$\dot{\mu} = \sum_{i=1}^{N} \frac{\partial \mu}{\partial \vec{x}_i}(t) \frac{\mathrm{d}\vec{x}_i}{\mathrm{d}t} = \sum_{i=1}^{N} \frac{\partial \mu}{\partial \vec{x}_i}(t) \vec{v}_i. \tag{2.118}$$

The explicit time dependence of the derivative of the total dipole with respect to the position of a specific particle $\frac{\partial \mu}{\partial \vec{x}_i}$ has been highlighted, as this implies that the infrared spectra will not be constructed from the time correlation functions $C_{v_i v_j}$ of the velocities themselves, but of the velocities multiplied

by these derivatives. If these derivatives were constant, the calculation would reduce to the previous case.

The diagonal elements $C_{\vec{v}_i\vec{v}_i}(\omega)$ of the Fourier transform of the velocity–velocity correlation functions are within the context of molecular systems referred to as the vibrational density of states (VDOS) of particle i. All the quantities $C_{\vec{v}_i\vec{v}_j}(\omega)$ taken together are referred to as generalized VDOS [26]. These functions are also connected to the Fourier transforms of the position–position correlation functions through the relation

$$C_{\vec{v}_i\vec{v}_j}(\omega) = \omega^2 C_{\vec{x}_i\vec{x}_j}(\omega), \tag{2.119}$$

which follows from the properties of the Fourier transform.

In this work, the VDOS of a chosen particle, or set of particles, was one of the objects of main interest. Its significance comes mainly from the fact that the main characteristics of the motion of the constituent particles, relevant in the calculation of vibrational spectra, are well captured in it. Vibrational spectra include the aforementioned infrared spectrum, but also the related Raman spectrum, among others. These differ in the derivatives multiplying the velocities, for instance for the Raman spectrum these would be $\frac{\partial\alpha}{\partial\vec{x}_i}$, where α is the total polarizability of the system. Since the VDOS is free from the multiplication by these derivatives, it encapsulates the features in these spectra arising from the dynamics of the particles themselves. A peak in the VDOS might have a different intensity or even position in a specific vibrational spectrum, or in fact not be present in some of them at all, but this peak can usually be mapped to a specific peak in the VDOS, and interpreted as arising from a specific motion of a specific particle.

2.4 Molecular dynamics

The established knowledge, which will be shown in this section has been sourced from the book 1 as well as from the lectures "Molecular dynamics I" and "Molecular dynamics II" taught by RNDr. Ondřej Maršálek, Ph.D.

In classical statistical mechanics of the canonical ensemble, an average of some quantity $A(x)$, which is only a function of particle coordinates x, can be calculated by performing the averaging

$$\langle A\rangle = \int_{-\infty}^{\infty} \mathrm{d}^n x A(x)\rho(x) \tag{2.120}$$

where $\rho(x)$ is a probability distribution function given by

$$\rho(x) = \frac{1}{Q}\exp\big(-\beta U(x)\big), \tag{2.121}$$

where $U(x)$ is the potential energy. Q is the configurational integral - the configurational contribution to the partition function Z which can be calculated

using

$$Q = \int_{-\infty}^{\infty} \mathrm{d}^n x \exp\bigl(-\beta U(x)\bigr). \tag{2.122}$$

Carrying out such an integral analytically is, however, possible for only a limited set of $U(x)$ and $A(x)$. In particular, for a molecular system treated classically, the total Hamiltonian is given by equation 2.11, and the potential $U(x)$ is equal to the electronic energy for the given configuration of the nuclei, which can be rather general. Numerical methods are thus needed to perform such integrals. The most often used methods of carrying out such integrals numerically for systems of many particles are Monte-Carlo methods, in particular, the Metropolis algorithm [27], and methods of *molecular dynamics*.

The main motivating feature behind the method of molecular dynamics is the fact that any phase space probability distribution $\rho(\mathbb{X})$ is preserved by classical Hamiltonian dynamics. This is the well-known Liouville's theorem and can be written in mathematical form as

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} + \{\rho, H\} = 0, \tag{2.123}$$

where $\{...,...\}$ denotes the Poisson bracket. One could thus obtain the average of a certain quantity $A(\mathbb{X}) = A(x)$ over the canonical distribution by first generating some, relatively low, number of samples from the distribution and then letting them evolve according to Hamiltonian dynamics. This process is usually called a "simulation" and it yields a sequence of phase space configurations $p(t), x(t)$ collectively known as a *trajectory* which are still distributed according to the original distribution. The average value of the quantity of interest $A(x)$ can then be calculated by simply averaging over these configurations. Other probability distributions could in principle be used, such as the microcanonical or grand canonical ensembles. Which one is used in practice depends on the apriori known thermodynamic properties of the studied system. In this work the canonical ensemble has been used exclusively, and as such *NVT molecular dynamics* will be used.

It is often the case that the system one wants to study using molecular dynamics is macroscopic in size. Such a system would be too large to simulate on any existing computer. Therefore, a different approach is used. A relatively small system, defined by its particle number and the potential energy $U(x)$, is usually considered. This system is to be constructed in such a way that all intensive quantities of interest are equal to their macroscopic limit.

Since generating the initial samples is generally not easy itself, *NVT* molecular dynamics is most often performed by numerically integrating Hamilton's equations from a single initial condition while using an additional numerical algorithm (referred to as a *thermostat*) to enforce the canonical ensemble. This is most often done by perturbing the particle momenta to keep them consistent with the Maxwell–Boltzmann distribution. A useful perspective

on this procedure is that the thermostat effectively generates the initial samples "on the fly". Molecular dynamics itself is thus effectively used to generate the initial samples. Generating the one initial condition itself might still be a problem, however as the canonical ensemble contains technically *all* phase space points — most just have exceedingly low probability — the initial condition may be in principle any configuration. Practically, choosing a highly improbable initial condition will bias the beginning of the simulation, until the thermostat has enough time to ensure representative sampling. This period is known as "equilibration" and the samples obtained from it are usually discarded. The length of this period can be decreased by choosing a phase–space configuration which is as close to a typical representative structure from the ensemble as possible. For momenta, this simply means randomly sampling them from the Maxwell–Boltzmann distribution. For positions, however, an a priori insight into the structure of the system is usually needed to do this.

The most commonly used algorithm for numerical integration of Hamilton's equations is the velocity Verlet [28] algorithm. This algorithm can be formulated in multiple different ways which all yield the same trajectories. The formulation which will be used here is due to Berne and Tuckerman [29]. To obtain this algorithm, one can start from the general relation

$$A(t) = \exp(iLt)A(0), \tag{2.124}$$

which holds within classical mechanics. In this relation, $A(t)$ is some function on phase space — the most important such function being the phase space coordinates themselves, and L is the Liouville operator defined as

$$iL = \{\ldots, H\} = \frac{\partial H}{\partial p}\frac{\partial}{\partial x} - \frac{\partial H}{\partial x}\frac{\partial}{\partial p} = iL_x + iL_p. \tag{2.125}$$

For a Hamiltonian in of the form $H = \frac{p^2}{2m} + U(x)$, the operators iL_q and iL_p are equal to:

$$iL_x = \frac{p}{m}\frac{\partial}{\partial x} = v\frac{\partial}{\partial x}, \tag{2.126}$$

$$iL_p = -\frac{\partial U}{\partial x}\frac{\partial}{\partial p} = F\frac{\partial}{\partial p}. \tag{2.127}$$

The exponential in equation 2.124 thus cannot be simply evaluated for such a Hamiltonian, as the operators iL_x and iL_p do not commute. A way around this is to use the Lie–Trotter product formula [21]

$$\exp(iLt) = \lim_{n\to\infty}\left\{\exp\left[iL_p\left(\frac{t}{2n}\right)\right]\exp\left[iL_x\left(\frac{t}{n}\right)\right]\exp\left[iL_p\left(\frac{t}{2n}\right)\right]\right\}^n \tag{2.128}$$

This formula holds exactly for $n \to \infty$, however, it can be truncated for numerical purposes at some high enough n. The leading correction term is in

such a case $O\left(\left(\frac{t}{n}\right)^2\right)$. This formula can be conveniently written using individual operators $\exp(iL\epsilon)$ which evolve the trajectory through a single time step $\epsilon = t/n$:

$$\begin{aligned} \exp(iLt) &= \exp(iL\epsilon)^n + O(\epsilon^2) \\ \exp(iL\epsilon) &= \exp\left(iL_p\frac{\epsilon}{2}\right)\exp(iL_x\epsilon)\exp\left(iL_p\frac{\epsilon}{2}\right) + O(\epsilon^3) \end{aligned} \tag{2.129}$$

The individual exponentials can now be applied directly to the function of interest using the property:

$$\exp\left(l\frac{\mathrm{d}}{\mathrm{d}x}\right)f(x) = f(x+l), \tag{2.130}$$

which implies that

$$\exp(iL_x\epsilon)f(x,p) = f(x+v\epsilon,p), \tag{2.131}$$

$$\exp\left(iL_p\epsilon\right)f(x,p) = f(x,p+F\epsilon). \tag{2.132}$$

The velocity Verlet algorithm is thus obtained, in which the following steps are repeated as long as needed:

$$\begin{aligned} p &\leftarrow p + F\frac{\epsilon}{2} \\ x &\leftarrow x + v\epsilon \\ F(x) &\leftarrow F(x+v\epsilon) \\ p &\leftarrow p + F\frac{\epsilon}{2}. \end{aligned} \tag{2.133}$$

The notation $x \leftarrow y$ denotes a numerical update of the quantity x and is to be understood as replacing the value of x by the value of y (the name of the variable stays the same). This notation is often encountered in literature on molecular dynamics (for example [30]) and will thus be also adopted here. To start off the loop, the forces need to be calculated one additional time.

This algorithm is time-reversal symmetric, and as such it cannot produce systematic drifts in the total energy (within the limits of finite precision arithmetic). Another advantage of the velocity Verlet algorithm is its symplectic nature — the transformation:

$$\begin{pmatrix} x(0) \\ p(0) \end{pmatrix} \rightarrow \begin{pmatrix} x(\epsilon) \\ p(\epsilon) \end{pmatrix} \tag{2.134}$$

constitutes a canonical transformation, preserving the form of Hamilton's equations. Stated in another way — propagating one step with the velocity

Verlet algorithm preserves the canonical 2-form $\mathrm{d}x \wedge \mathrm{d}p$. Stated in yet another way, this implies that Liouville's theorem holds for such an integrator and the phase-space volume is preserved by it.

It is worth noting that any consistent numerical algorithm for integrating ordinary differential equations shares these properties when applied to Hamilton's equations in the limit $\epsilon \to 0$. The important advantage of the velocity Verlet algorithm is that it possesses these properties for (in principle) *any* ϵ. Within molecular dynamics simulations, it is desirable to use as high as possible ϵ, as this gives the most simulation time (and thus the most information) given limited computational resources — hence the advantage of the velocity Verlet algorithm. In practice, this time step is limited by the shortest timescale present in the simulation. Another important point is that in the majority of cases, the most time-consuming step of the velocity Verlet algorithm is the calculation of the forces. For molecular systems, as outlined in the preceding section, these come from the solution of the electronic problem. But even if a pair potential approximation for the electronic energy is employed, the computational costs of evaluating the forces typically dominate over the cost of evolving the system in time. This highlights another advantage of the velocity Verlet algorithm — it requires only one evaluation of the forces per step. In principle, more accurate algorithms could be devised using the same derivation as was showcased here, but using a symmetric Lie–Trotter product formula of a higher order. These would also share all the previously mentioned advantages of the velocity Verlet algorithm. Such algorithms are however rarely used, as they would require more evaluations of the forces per step.

A thermostat will now be introduced by artificially adding another term to the Liouville operator:

$$L = L_x + L_p + L_T. \tag{2.135}$$

The Lie–Trotter formula can now be used to symmetrically split the exponential of the Liouville operator, while preserving the basic velocity Verlet structure, in two different possible ways. The first scheme is:

$$\begin{aligned}\exp(iL\epsilon) = &\exp\left(iL_p\frac{\epsilon}{2}\right)\exp\left(iL_x\frac{\epsilon}{2}\right)\exp(iL_T\epsilon)\exp\left(iL_x\frac{\epsilon}{2}\right)\exp\left(iL_p\frac{\epsilon}{2}\right)\\ &+ O(\epsilon^3),\end{aligned} \tag{2.136}$$

while the other is:

$$\begin{aligned}\exp(iL\epsilon) = &\exp\left(iL_T\frac{\epsilon}{2}\right)\exp\left(iL_p\frac{\epsilon}{2}\right)\exp(iL_x\epsilon)\exp\left(iL_p\frac{\epsilon}{2}\right)\exp\left(iL_T\frac{\epsilon}{2}\right)\\ &+ O(\epsilon^3).\end{aligned} \tag{2.137}$$

There are multiple possible choices for the operator $\exp(iL_T\epsilon)$ itself. A simple yet useful choice is a thermostat based on the Langevin equation 2.77 [31]. As

was shown beforehand in equation 2.82, the distribution of momenta of particles obeying the Langevin equation converges for to the Maxwell–Boltzmann distribution $t \to \infty$, and as such, this is a valid approach for constructing a thermostat. To this end, the equation 2.83 will be useful while making a reassignment $\frac{\gamma}{m} \to \gamma$ often employed in literature on Langevin thermostats.

According to this equation, the thermostat consists of two components — a friction part and a stochastic part, both being connected through the parameter γ. In a single thermostat step, the friction part is first applied by rescaling the momenta:

$$p \leftarrow \exp(-\gamma\epsilon)p \tag{2.138}$$

and afterward, the stochastic part is applied by adding a random vector to the momenta:

$$p \leftarrow p + \sqrt{\frac{m}{\beta}\left[1 - \exp(-2\gamma\epsilon)\right]}\zeta(t), \tag{2.139}$$

where ζ is a vector of normally distributed random numbers of zero mean, unit variance, and same size as the vector of momenta p. For a Langevin thermostat, the Liouville operator splitting in equation 2.136 was found to possess better numerical properties than the alternative in equation 2.137 [31]. The parameter γ controls the rate of convergence of the momenta to the thermal distribution and is usually referred to as the "strength" of the thermostat. On timescales $\tau >> \frac{1}{\gamma}$, the distribution of the momenta of the particles reproduces the Maxwell–Boltzmann distribution, while on timescales $\tau << \frac{1}{\gamma}$ the momenta behave more like they would in Hamiltonian dynamics. This brings forth the idea that a simulation with a Langevin thermostat could in principle be used to calculate time correlation functions for timescales $t << \frac{1}{\gamma}$. However, the thermostat as described here still perturbs the Hamiltonian dynamics too much for such correlation functions to be of use. This is due to the thermostat being a *local* one. A thermostat is called local if, in each application of the thermostat, every momentum degree of freedom is acted upon by the thermostat separately (as is the case here). In contrast to this, a *global* thermostat acts on a set of momentum degrees of freedom (usually all of them) as a whole. A global thermostat is typically needed to obtain reasonable Hamiltonian dynamics for timescales $t << \frac{1}{\gamma}$, as it "distributes" the effect of the thermostat over multiple degrees of freedom and thus perturbs each individual one less. A global version of the Langevin thermostat (for more details of this version see appendix D) can be constructed [32], often called the *Canonical sampling through velocity rescaling* (CSVR) thermostat [32]. The

algorithm for a single thermostat step is:

$$\begin{aligned} \alpha^2 &= c + \frac{(1-c)\left(\sum_{i=1}^{N}\zeta_i^2\right)}{2\beta K} + 2\zeta_1\sqrt{\frac{c(1-c)}{2\beta K}} \\ \mathrm{sign}(\alpha) &= \mathrm{sign}\left(\zeta_1 + \sqrt{\frac{2\beta K c}{(1-c)}}\right) \\ p &\leftarrow \alpha p, \end{aligned} \tag{2.140}$$

where N is the number of momentum degrees of freedom, $c = \exp\left(-2\gamma\epsilon\right)$, K is the kinetic energy of the considered particles, and ζ_i are again normally distributed random numbers with zero mean and unit variance. In the case $N = 1$ this thermostat reduces to the local case:

$$\begin{aligned} \alpha^2 &= c + \frac{m_1(1-c)\zeta_1^2}{\beta p_1^2} + 2\zeta_1\sqrt{\frac{m_1 c(1-c)}{\beta p_1^2}} \\ &= \left(c + \frac{1}{p_1}\sqrt{\frac{m_1}{\beta}\left[1-\exp\left(-2\gamma\epsilon\right)\right]}\zeta_1\right)^2 \\ p_1 &\leftarrow p_1 + \sqrt{\frac{m_1}{\beta}\left[1-\exp\left(-2\gamma\epsilon\right)\right]}\zeta_1 \end{aligned} \tag{2.141}$$

The use of *NVT* molecular dynamics is not limited just to the problems of classical statistics. Using the path integral formulation, the partition function of a system of distinguishable quantum particles is given by equation (2.33):

$$\begin{aligned} Z &= \int \mathrm{d}^n x \left\langle x \middle| \exp\left(-\beta\hat{H}\right) \middle| x \right\rangle \\ &= \lim_{P\to\infty} \int \prod_{i=1}^{P} \mathrm{d}^n x_i \left(\frac{mP}{2\pi\hbar^2\beta}\right)^{\frac{nP}{2}} \exp\left\{-\sum_{j=1}^{P}\left[\frac{mP}{2\hbar^2\beta}(x_j - x_{j-1})^2 + \frac{\beta}{P}U(x_{j-1})\right]\right\}. \end{aligned} \tag{2.142}$$

With the condition $x_0 = x_P$. For the sake of numerical approximation, this expression may be truncated at some large enough but finite P. The expression then looks just like a classical configuration integral for a system composed of $\frac{nP}{3}$ particles in three dimensions with the potential energy equal to:

$$U(x) = \sum_{j=1}^{P} \frac{1}{2} m\omega_P^2 (x_j - x_{j-1})^2 + U(x_{j-1}) \tag{2.143}$$

at a higher temperature corresponding to β/P. The definition $\omega_P \equiv P/(\hbar\beta)$ was introduced. The analogy with the classical case can be completed by per-

forming nP reverse Gaussian integrations, the expression becoming:

$$Z = \int \prod_{i=1}^{P} d^3 p_i \prod_{i=1}^{P} d^3 x_i \cdot \\ \times \exp\left\{ -\frac{\beta}{P} \sum_{j=1}^{P} \left[\frac{p_j^2}{2m'} + \frac{1}{2} m\omega_P^2 (x_j - x_{j-1})^2 + U(x_j) \right] \right\}, \tag{2.144}$$

where formally $m' = \frac{m}{(2\pi\hbar)^2}$, however as an overall constant prefactor of the partition function has no effect on the thermodynamic quantities obtained from it, this parameter may be chosen freely. The quantum partition function thus can be calculated as the classical partition function of an extended system of P times the original number of particles with the Hamiltonian:

$$H(x) = \sum_{j=1}^{P} \frac{p_j^2}{2m'} + \frac{1}{2} m' \frac{m}{m'} \omega_P^2 (x_j - x_{j-1})^2 + U(x_j) \tag{2.145}$$

at a P times the temperature. Such a Hamiltonian corresponds to a series of "ring polymers" (each replacing one quantum particle) with each of its P "beads" (or alternatively "replicas") harmonically coupled to its two neighbors. This mapping between a problem of quantum statistics of a single distinguishable particle and a problem of classical statistics of a ring polymer is sometimes referred to as "classical isomorphism" [33]. Since this is now a problem of classical statistics, the thermodynamic and statistical properties of a quantum mechanical system of distinguishable particles can thus be obtained by simulating such a system using the established methods of molecular dynamics.

A point worth noting is the fact that the second term in equation 2.145 has been rather suggestively rewritten to stress the important role which m' plays in the dynamics of the ring polymer. Not only does it play the role of the inertial mass of the constituent beads of the ring polymer, but it also determines the frequency of the harmonic coupling between the beads, the frequency being equal to

$$\omega = \sqrt{\frac{m}{m'}} \omega_P \tag{2.146}$$

As this parameter may be chosen freely, as was previously discussed, so can the frequency be chosen freely. Different beads can even in principle have different choices of $m' \to m'_j$ and thus different coupling frequencies (the only difference still being an overall numerical prefactor of the partition function). This point will be further touched upon later in subsection 3.1.

The expectation value of any operator $\hat{A} = A(\hat{x})$ may be calculated following equation 2.34 as (integrating out the momenta again, to make the equa-

tion simpler)

$$\begin{aligned}\langle \hat{A}(\hat{x})\rangle = \frac{1}{Z}\int \prod_{i=1}^{P} \mathrm{d}^n x_i \left(\frac{mP}{2\pi\hbar^2\beta}\right)^{\frac{nP}{2}} A(x_P) \\ \times \exp\left\{-\frac{\beta}{P}\sum_{j=1}^{P}\left[\frac{1}{2}m\omega_P^2(x_j - x_{j-1})^2 + U(x_{j-1})\right]\right\},\end{aligned} \tag{2.147}$$

where the condition $x_P = x_0$ still applies. Since all of the ring polymer beads are equivalent, this relation may be rewritten as

$$\begin{aligned}\langle \hat{A}(\hat{x})\rangle = \frac{1}{Z}\int \prod_{i=1}^{P} \mathrm{d}^n x_i \left(\frac{mP}{2\pi\hbar^2\beta}\right)^{\frac{nP}{2}} \left(\sum_{j=1}^{P}\frac{1}{P}A(x_j)\right) \\ \times \exp\left\{-\frac{\beta}{P}\sum_{j=1}^{P}\left[\frac{1}{2}\omega_P^2(x_j - x_{j-1})^2 + U(x_{j-1})\right]\right\}.\end{aligned} \tag{2.148}$$

Calculating the mean in this fashion is advantageous as it increases the numerical sampling, and thus accelerates the convergence of $\langle \hat{A}\rangle$. The quantity $\frac{1}{P}\sum_{j=1}^{P} A(x_j)$ is referred to within the context of path-integral molecular dynamics as the estimator of the operator $\hat{A}$. To compute the expectation value of an operator that is not diagonal in the position basis, the off-diagonal elements of the density matrix are required. The calculation of these would necessitate the condition $x_P = x_0$ to be lifted, and the system would thus resemble classical polymers with free ends. An exception to this is the operator of quantum kinetic energy $T(\hat{p}) = \frac{\hat{p}^2}{2m}$, which can be calculated from ring polymer simulations. This results from the fact that the mean total energy can be calculated from the partition function using the relation

$$\begin{aligned}\langle E\rangle &= -\frac{1}{\beta}\ln(Z) \\ &= \frac{P}{2\beta} - \frac{1}{Z}\int \prod_{i=1}^{P} \mathrm{d}^n x_i \left(\frac{mP}{2\pi\hbar^2\beta}\right)^{\frac{nP}{2}} \frac{1}{P}\left[\sum_{j=1}^{P}\frac{1}{2}\omega_P^2(x_j - x_{j-1})^2 + U(x_{j-1})\right] \\ &\times \exp\left\{-\frac{\beta}{P}\left[\sum_{j=1}^{P}\frac{1}{2}\omega_P^2(x_j - x_{j-1})^2 + U(x_{j-1})\right]\right\},\end{aligned} \tag{2.149}$$

which is an average of the quantity

$$\frac{P}{2\beta} - \frac{1}{P}\left[\sum_{j=1}^{P}\frac{1}{2}\omega_P^2(x_j - x_{j-1})^2 + U(x_{j-1})\right] \tag{2.150}$$

over the canonical ensemble of the ring polymer configurations. In the third term of this expression, one can recognize the estimator of the potential energy operator $\hat{U}(\hat{x})$. The rest thus corresponds to an estimator of the kinetic energy operator and is accordingly called the *primitive kinetic energy estimator*. From a numerical point of view, this operator suffers from relatively large fluctuations due to its linear dependence on P. A kinetic energy estimator with fluctuations of smaller magnitude can be constructed by employing the virial theorem [34], which gives:

$$\frac{P}{2\beta} - \left\langle \frac{1}{P} \left[\sum_{j=1}^{P} \frac{1}{2} \omega_P^2 (x_j - x_{j-1})^2 \right] \right\rangle = \left\langle \frac{1}{P} \sum_{j=1}^{P} \frac{1}{2} x_j \frac{\partial U}{\partial x_j} \right\rangle. \tag{2.151}$$

The expression $\frac{1}{P} \sum_{j=1}^{P} \frac{1}{2} x_j \frac{\partial U}{\partial x_j}$ is thus referred to as the *virial kinetic energy estimator*.

The final discussion of this section will concern the usage of molecular dynamics for the calculation of time correlation functions. The technical details of these calculations themselves will however be left for subsection 3.5. As the current framework of molecular dynamics can generate both numerical approximations to classical trajectories, as well as samples from the canonical ensemble, it can be used to calculate classical time correlation functions $C_{AB}(t)$. The situation is quite different in the quantum case. Any quantum time-correlation function involves operators evolved according to the Heisenberg picture:

$$\hat{B}(t) = \exp\left(\frac{i}{\hbar} \hat{H} t\right) \hat{B}(0) \exp\left(-\frac{i}{\hbar} \hat{H} t\right). \tag{2.152}$$

The calculation of quantum time correlation functions thus requires integration over the complex exponential $\exp\left(\frac{i}{\hbar} \hat{H} t\right)$. The established methods of molecular dynamics (as well as Monte Carlo methods) cannot be used to calculate such integrals using the path integral approach, as the exponential is not positive definite, and as such all sampling schemes fail. The integrals also cannot be carried out efficiently using other numerical techniques (be it by using the path integral formulation or not) in high-dimensional systems due to convergence issues connected with the numerical sign problem. Loosely speaking, the numerical sign problem stems from the fact that such integrals contain a large number of mutually nearly-canceling terms, all of which need to be accounted for to yield a reasonable result. Such integrals can thus be carried out only for low-dimensional systems, as the number of terms that need to be considered scales with the total volume of the system, which scales exponentially with the dimension. The problem can be circumvented by diagonalizing the Hamiltonian, but this operation itself scales exponentially with the number of particles, and as such cannot be practically used for systems of more than a few particles.

No general solution or approximation which would be generally applicable, yet at the same time systematic and well-founded, exists. Within the con-

text of molecular systems in the condensed phase, several approximations have been developed. These mostly fall into two categories — semiclassical methods and methods based on imaginary-time path-integral molecular dynamics. Both approaches exploit the relatively short decoherence times which are to be expected in the condensed phase. This work will focus on a method falling within the latter of these categories — *centroid molecular dynamics* originally put forward by Cao and Voth [35, 10]. This method approximates the quantum Kubo-transformed time correlation function with the classical time correlation function (obtained from the same system treated at a classical level), but with the additional change $U(x) \rightarrow W(x)$, $W(x)$ being the centroid free energy introduced in equation 2.38. Although a formal chain of approximations going from the exact quantum dynamics to CMD can be established [11], using the *Matsubara dynamics* [12] introduced by Hele et al. as an intermediate, there is no formal justification for taking some of these approximation steps. As such, the method is rather *ad hoc* in nature, but several motivations can be put forward. The main motivating feature behind CMD (and all other methods based on imaginary-time path-integral molecular dynamics) is the fact that the resulting dynamics is at all times consistent with the quantum canonical distribution, and as such satisfies the quantum condition of detailed balance:

$$\frac{P(A \rightarrow B)_t}{P(B \rightarrow A)_t} = \frac{P(B)}{P(A)}, \tag{2.153}$$

where the probabilities $P(B)$ and $P(A)$ are given by the quantum canonical distribution, and A and B are some generic values of an observable. This also implies that quantum mechanical effects present at the level of static distributions — such as zero point energy and quantum tunneling — are reflected in the resulting dynamics. The fact that the dynamics is consistent with the condition of detailed balance also implies that it is microscopically reversible. As such, the classical time correlation functions calculated using these methods are even functions of time (a symmetry shared by the Kubo-transformed time correlation function). Furthermore, it can be shown that the approximation is exact in the classical and harmonic limits. Away from these limits, the position autocorrelation functions calculated using the method can be shown to be accurate up to $O(\hbar^3)$ [36] for short timescales up to $O(t^6)$ [37]. Despite these encouraging features, it is worth noting that CMD breaks down for operators which are non-linear functions of the position and momentum operators [11], which inherently limits the usage of CMD as a method. As this method is the main topic of this work it will be discussed in greater detail in the following sections.

3 Computational implementation

This section will go over the specific implementation choices made within the present work when applying the theory discussed in the preceding section. This will also include the methods used to process the data to yield the results, which will be presented in the next section. First, an efficient implementation of path-integral molecular dynamics will be discussed. This will be followed by a discussion on the possibilities of implementing centroid molecular dynamics. Here, the machine learning approach proposed by this work will be detailed. The final topic will be the calculation of time correlation functions and frequency spectra.

The information presented in this section is a combination of established knowledge and methods, which have been mostly obtained by the author from the cited articles, as well as from discussions with the thesis advisor – RNDr. Ondřej Maršálek, Ph.D., and newly proposed methodology. The newly proposed methodology is mostly contained in subsection 3.4, and its appearance will be emphasized in the upcoming text.

3.1 Efficient path-integral molecular dynamics

Even though path-integral molecular dynamics can in principle be carried out in the same fashion as ordinary molecular dynamics, this approach is most often not suitable. To see the reason for this, notice that the ring polymer Hamiltonian given in equation 2.145 can be written as:

$$H = \sum_{j=0}^{P-1} \frac{p_j^2}{2m'} + \sum_{0=1}^{P-1} \frac{1}{2} m\omega_P^2 x_j^T M_{ji} x_i + U(x_j), \tag{3.1}$$

where M is the real symmetric matrix

$$M = \begin{pmatrix} 2 & -1 & 0 & 0 & \cdots & 0 & -1 \\ -1 & 2 & -1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 2 & -1 & & \vdots & \vdots \\ & & \ddots & \ddots & \ddots & & \\ \vdots & \vdots & & -1 & 2 & -1 & 0 \\ 0 & 0 & \cdots & 0 & -1 & 2 & -1 \\ -1 & 0 & \cdots & 0 & 0 & -1 & 2 \end{pmatrix}. \tag{3.2}$$

Note that an advantageous shift of the summation indices has been performed. Due to its translational symmetry, this matrix can be diagonalized by vectors of the form

$$n_k = \exp\left(i\frac{2\pi kj}{P}\right) \qquad k \in \{0, 1, \ldots, (P-1)\} \tag{3.3}$$

with the eigenvalues

$$\lambda_k = 4\sin^2\left(\frac{\pi k}{P}\right). \tag{3.4}$$

A coordinate transformation matrix can be constructed out of the eigenvectors in equation 3.3, which form a basis of the eigenspace of the matrix M. It is convenient to take this matrix real (which can be done, since M is symmetric) and orthogonal. If U is the unitary matrix, where each column is one eigenvector from equation (3.3), then this matrix can be found as

$$C = \frac{1}{2\sqrt{P}}(U + U^\dagger), \tag{3.5}$$

and its matrix elements are

$$C_{ij} = \begin{cases} \sqrt{\frac{1}{P}}, & i = 0 \\ \sqrt{\frac{2}{P}}\cos\left(\frac{2\pi ij}{P}\right), & 1 \le i \le \frac{P}{2} - 1 \\ \sqrt{\frac{1}{P}}(-1)^j, & i = \frac{P}{2} \\ \sqrt{\frac{2}{P}}\sin\left(\frac{2\pi ij}{P}\right), & \frac{P}{2} + 1 \le i \le P - 1. \end{cases} . \tag{3.6}$$

Using this matrix, the *normal mode* coordinates and momenta can be introduced:

$$\tilde{x}_k = \sum_{j=0}^{P-1} C_{kj} x_j, \qquad \tilde{p}_k = \sum_{j=0}^{P-1} C_{kj} p_j, \tag{3.7}$$

in which the ring polymer Hamiltonian obtains the following form:

$$H = \sum_{k=0}^{P-1}\left(\frac{\tilde{p}_k^2}{2m'} + \frac{1}{2}m'\omega_k^2\tilde{x}_k^2 + \tilde{U}(\tilde{x})\right), \tag{3.8}$$

where the normal mode frequencies

$$\omega_k = 2\sqrt{\frac{m}{m'}}\omega_P \sin\left(\frac{k\pi}{n}\right) \tag{3.9}$$

were introduced. Note that the coordinate transformation leaves the kinetic term the same, due to the orthogonality of C, but the potential $\tilde{U}$ generally takes a different form than in Cartesian coordinates. It is worth noting that the $P = 0$ normal mode multiplied by $\sqrt{\frac{1}{P}}$ corresponds to a P-replica approximation to the centroid introduced in equation 2.35. It can thus be seen that the system contains P "stiff" harmonic modes spread over a broad frequency range. This results in poor energy exchange between the normal modes, as the only way for them to exchange energy is through anharmonicities in $\tilde{U}(\tilde{x})$ [38]. A strong enough thermostat thus needs to be attached to each normal mode [39]. The broad frequency range spanned by the normal

modes makes the system rather complex. The highest of these frequencies are also often much higher than any of the physical frequencies present in the system and thus severely limit the simulation time step. This limitation can be overcome by performing the molecular dynamics simulation entirely in normal modes, switching to Cartesian coordinates only to calculate forces and write output. The switch to Cartesian coordinates is done, because the potential $\tilde{U}$ is generally not known. In practice, only the forces are required to run molecular dynamics and these can be transformed to normal modes using the same matrix C from their Cartesian versions, and as such the $\tilde{U}$ is in fact not necessary. In this approach, the equations of motion of the quadratic part of the Hamiltonian alone are integrated exactly. These are then numerically corrected for the presence of $\tilde{U}(\tilde{x})$ [30].

This results in a time evolution scheme with a thermostat:

$$\begin{aligned}\exp(iL\epsilon) = &\exp\left(iL_\gamma\frac{\epsilon}{2}\right)\exp\left(iL_V\frac{\epsilon}{2}\right)\exp(iL_0\epsilon)\exp\left(iL_V\frac{\epsilon}{2}\right)\exp\left(iL_\gamma\frac{\epsilon}{2}\right)\\ &+O(\epsilon^3),\end{aligned} \tag{3.10}$$

where $iL_0\epsilon$ corresponds to the time evolution of the free ring polymer:

$$\begin{pmatrix}\tilde{p}_k\\ \tilde{x}_k\end{pmatrix} \leftarrow \begin{pmatrix}\cos(\omega_k\epsilon) & -m'_k\omega_k\sin(\omega_k\epsilon)\\ \frac{1}{m'_k\omega_k}\sin(\omega_k\epsilon) & \cos(\omega_k\epsilon)\end{pmatrix}\begin{pmatrix}\tilde{p}_k\\ \tilde{x}_k\end{pmatrix}, \tag{3.11}$$

$iL_p\frac{\epsilon}{2}$ corresponds to the evolution of the momenta under the influence of $\tilde{U}(\tilde{x})$:

$$\begin{aligned}\tilde{F}_k &\leftarrow \sum_{j=0}^{P-1} C_{kj}F_j\\ \tilde{p}_k &\leftarrow \tilde{p}_k + \tilde{F}_k\frac{\epsilon}{2}\end{aligned} \tag{3.12}$$

and $iL_\gamma\frac{\epsilon}{2}$ corresponds to applying a Langevin thermostat to each normal mode:

$$\tilde{p}_k \leftarrow \exp\left(-\gamma_k\frac{\epsilon}{2}\right)\tilde{p}_k + \sqrt{\frac{m'_k}{\beta}\left[1-\exp\left(-2\gamma_k\frac{\epsilon}{2}\right)\right]}\zeta_k(t). \tag{3.13}$$

As indicated, the parameter γ should in general be chosen differently for each normal mode. An optimal choice of these parameters is [30]:

$$\gamma_i = \begin{cases}\frac{1}{\tau_0}, & k=0\\ 2\omega_k, & k>0,\end{cases} \tag{3.14}$$

τ_0 being the desired timescale of thermalization of the centroid momenta. This algorithm is called PILE-L (PILE standing for path integral Langevin equation, and L for local). A global CSVR thermostat can alternatively be attached

to the particle centroids. This method is then called PILE-G. After the transformation to normal modes, the parameter m' retains its role as the inertial mass (now, of the normal modes), while determining their frequency. This parameter can also still be chosen freely, as was previously discussed in subsection 2.4. This can also be done separately for each normal mode, as has been already suggested in equations 3.11 and 3.13. A common use of this freedom is a shift of all the normal mode frequencies to a common lower frequency, which further loosens the requirements on the simulation time step and increases the accuracy of the numerical integration of the equations of motion. Note that if the forces are simple enough to calculate and a high number of particles and replicas are used, it may be the case that the transformations to and from normal modes may be the rate-limiting step of the simulation. Most often, the transformation of the coordinates is thus not done using matrix multiplication as described before, but by using a fast Fourier transform, which has a more advantageous scaling — $O(P\log(P))$, compared to the $O(P^2)$ scaling of matrix multiplication.

To perform path-integral molecular dynamics simulations for molecular systems, the i-Pi program [40] was used. This program (written in the Python programming language) performs the evaluation of the scheme described above, while taking the required forces from an external program through a socket interface. The CP2K [41] program was used to evaluate forces based on an existing machine learning model constructed using data from electronic structure calculations [42].

Path-integral molecular dynamics simulations were also performed for small model systems. These were done using a custom program written in the Fortran programming language for this work. Within this program, the Fastest Fourier Transform in the West library [43] was employed to perform the coordinate transformations to and from normal modes.

3.2 Centroid molecular dynamics in practice

As was outlined before, the CMD method is defined as classical molecular dynamics with the potential given by the path integral centroid free energy $W(x)$. This choice of potential yields the following equations of motion:

$$\frac{\partial p}{\partial t} = -\frac{\partial W(x)}{\partial x} \tag{3.15}$$

$$\frac{\partial x}{\partial t} = \frac{p}{m}. \tag{3.16}$$

The derivative $\frac{\partial W(x)}{\partial x}$ can be, in the discrete approximation, expressed using equations (2.38) and (2.39):

$$\begin{aligned}\frac{\partial W(x)}{\partial x} = \frac{1}{\rho_c(x)} \int \prod_{i=1}^{P} \mathrm{d}^n x_i \left(\frac{mP}{2\pi\hbar^2\beta}\right)^{\frac{nP}{2}} \delta(x_c(x_i) - x) \\ \times \left(\frac{1}{P}\sum_{i=1}^{P} \frac{\partial U}{\partial x_i}\right) \exp\left\{-\sum_{j=1}^{P}\left[\frac{mP}{2\hbar^2\beta}(x_j - x_{j-1})^2 + \frac{\beta}{P}U(x_{j-1})\right]\right\}.\end{aligned} \tag{3.17}$$

The forces acting on the particles thus correspond to the mean force acting on the centroid fixed at x. CMD should thus be performed by integrating Hamilton's equations (with or without a thermostat) while calculating forces at each step by performing an auxiliary path-integral molecular dynamics simulation with the centroid fixed at the current position of the particles to determine the mean force. The computational costs associated with such a scheme would, however, be extreme, and as such it is essentially never used. In practice, a different approach was adopted — a path-integral molecular dynamics simulation is performed with all the non-centroid normal modes shifted (by adjusting the parameters m'_k) to a common high frequency [14, 44]. There are two ways of looking at the role of the high frequency. The first is that it ensures that the non-centroid degrees of freedom are adiabatically separated from the physical frequencies present in the system. This suppresses the energy exchange between the non-centroid degrees of freedom and the centroid. The second perspective is that the high frequency makes it so that the force on the centroid, averaged over short enough timescales such that the centroid can be considered still in the vicinity of x, is close to its converged value for the given position of the centroid x. Both of these however describe essentially the same phenomenon. This method has been used in the past to yield results consistent with both other approximations to quantum dynamics for molecular systems in the condensed phase [45] and with experiment [46]. These successes notwithstanding, it still suffers from a series of drawbacks. The most readily apparent drawback is the fact that the high frequency becomes a new convergence parameter. Raising the frequency severely limits the time step and it is thus desirable to keep this frequency as low as possible while achieving adiabatic separation. Having found a suitable frequency, the time step should then be further optimized to accommodate this frequency, while being as long as possible. An additional issue is that the non-centroid normal modes in reality do not occupy only a single frequency, but a rather broad range of frequencies. This effect is caused by the thermostats attached to these modes. To alleviate this problem, the γ parameter of the thermostats may be lowered to eliminate any spurious effects caused by the presence of the thermostats and to accelerate the convergence with respect to the target frequency [45]. In summary, setting up this procedure requires a generally non-trivial optimization procedure in the three-dimensional space of the

time step, γ parameter, and the target frequency. The optimal time step necessary for running adiabatic CMD however still usually ends up being more than ten times shorter than the optimal time step for an equivalent classical simulation. This presents a substantial computational drawback. To circumvent these difficulties, it is proposed in this work that instead of performing adiabatic CMD, the centroid potential $W(x)$ (or respectively its derivatives) can be pre-calculated and the CMD simulations ran as purely classical molecular dynamics simulations. The construction of the potential $W(x)$ will be detailed in subsection 3.4.

3.3 Machine learning potentials

The recent years and decades saw a rapid expansion of the possibilities of methods of machine learning in a wide array of fields. In the context of molecular dynamics, perhaps the main impact of these methods was the introduction of *machine learning potentials*. As was previously mentioned, the calculation of the forces — which are derivatives of the underlying Born–Oppenheimer potential energy surface is the most costly part of molecular dynamics. Especially in the case when these are calculated directly by solving the electronic structure problem, the cost can become extreme. Machine learning potentials seek to solve this issue by constructing a machine learning model which can predict the forces and potential energy for a given structure without the necessity of solving the electronic problem every time.

These methods have been used successfully in the past to construct accurate models of potential energy surfaces while using only a small set of training structures (compared to the number of structures visited in a typical trajectory) and still remain the subject of active research.

As the theory of machine learning is rich and expansive, and its study was not the main goal of this work, the discussion of this theory will be kept rather brief, and only the most important (for the present work) properties of these methods will be presented. Furthermore, although other approaches to machine learning exist, the discussion will be limited only to feedforward artificial neural networks, on which the machine learning methods used in this work were based. An overview of these neural networks and specifically their use in the prediction of potential energy surfaces of atomic and molecular systems can be found for instance in the review by Behler [16]. These networks present highly flexible (meaning they contain a large number of free parameters, subject to optimization) non-linear models, which can be used to approximate any function of the input data (under certain assumptions) [47]. The parameters of these networks are determined in a process usually referred to as “training”, where a (case dependent) loss function of the neural network output is constructed using a known *training set* of pairs of reference input and output values. This loss function is then minimized with

respect to the model coefficients. An advantageous feature of these networks is that the resulting optimized model is available as an analytic function, and as such its derivatives with respect to the input parameters are also available.

The machine learning potentials used in this work were the so-called *high-dimensional neural network potentials* (HDNNPs). The defining feature of HDDNPs is the fact that the total energy or force is calculated as a sum of atomic contributions. This can be well illustrated on the historically first instance of HDNNPs — the Behler–Parrinello [48] potentials. In the Behler–Parrinello approach, each chemical element is assigned an artificial neural network, which takes the *atom centered symmetry functions* [49], calculated from the Cartesian coordinates of atoms within a cutoff distance of any given atom of this element as an input, and produces a potential energy value as an output. A slightly disadvantageous aspect of the Behler–Parrinello potentials is the fact that in the most common implementation of these potentials, the training of these potentials requires both the energies and the forces for the training data. The total potential energy is then calculated as a sum of atomic contributions from all atoms (for atoms corresponding to the same element this energy is calculated by evaluating the same neural network model) and the force as the derivative of this total energy with respect to the atomic positions. This is the approach on which the previously mentioned model [42] to predict the Born-Oppenheimer potential energy surface was based, albeit with the additional usage of *committees*. A point worth mentioning is, that the symmetry functions are constructed from only the local surroundings of each atom, long-range interactions are not explicitly present in these models. They are present in an implicit fashion since the local surroundings of each atom is influenced by these interactions, but even so, these models generally perform better for short-range interactions.

A second type of machine learning potentials has been used in this work — equivariant message-passing machine learning potentials [50]. In this slightly different approach, a single feed-forward neural network is used for all elements present in the system, however, it takes the element type as an additional input. Furthermore, as the name suggests these potentials employ *message passing*. This means that, roughly speaking, several times (in the so-called interaction layers) information about the state of the given atomic neural network (the single neural network is evaluated for each atom separately) is passed to all the other atomic neural networks within a cutoff distance, which they use as further input information. Lastly, the used approach was an *equivariant* one. Roughly speaking again, the neural network uses additional tensor information as an input and preserves its tensor nature throughout the model. This ensures that the model output (which can be in principle a tensor of any order) transforms properly under coordinate transformations (belonging to the Euclidean group). In contrast, the previous “non-equivariant” models were traditionally invariant with respect to these transformations,

and thus proper behavior was ensured by them only for rank zero tensors. Note, however, that for the prediction of potential energies and forces, an equivariant network is not necessary as the potential energy is a rank zero tensor. The force — minus its derivative — thus also transforms automatically as a rank one tensor even if an equivariant network is not used. An advantageous feature of this approach, in contrast to the Behler–Parrinello approach is the fact that in most of its implementations, the training can be done using forces only. These models likewise perform better for short-range interactions. The NequIP program [50] was used to construct these models and its plugin ("pair style") for the LAMMPS [51] package was used to evaluate them.

3.4 Construction of the centroid potential

In this work, it is proposed that the methodology of CMD can be improved by performing CMD as purely classical molecular dynamics on the potential $W(x)$. In this approach, the derivatives of the potential $\frac{\partial W(x)}{\partial x}$ are calculated beforehand as a function of the particle coordinates x for all of the coordinates which are expected to be visited in a typical simulation. The reason for pre-calculating just the forces and not the potential $W(x)$ itself is the fact that the potential itself is not needed (molecular dynamics requires just the forces, not the potential itself), and in fact hard to obtain. In contrast to this, calculating the force $-\frac{\partial W(x_i)}{\partial x}$ for a given x_i is relatively simple to obtain using equation 3.17 — an imaginary-time path-integral molecular dynamics simulation is run with the centroid fixed at x_i, and the average force on the centroid is subsequently read off from this simulation, which is the force $-\frac{\partial W(x_i)}{\partial x}$ needed. The most important question, however, is how to construct the forces for all x (within some relevant volume of the space) from forces calculated for some reasonable amount of configurations.

For systems of low dimensionality, this can be done rather easily. In this work for low-dimensional systems, the forces were constructed by calculating the forces explicitly, according to the procedure mentioned in the previous paragraph, for positions lying on a grid. The resulting forces were subsequently interpolated using cubic splines [52], using the SciPy library [53].

For realistic molecular systems, this approach could not be used due to their high dimensionality. Instead, methods of machine learning potentials were used. These can successfully reproduce highly-dimensional Born–Oppenheimer potential energy surfaces, and since the potential $W(x)$ is itself a type of "potential energy surface" (from the point of view of the classical dynamics which results), these methods were seen as ideal candidates. Since only the forces were available, not the absolute values of the free energies, the second of the methods mentioned in subsection 3.3 — equivariant message-passing neural network potentials — were used, since the available implementation allows

training on forces only. The training set for constructing machine learning potentials needs to be chosen with care, as the quality of the resulting potential keenly depends on it. The training structures used in this work were the same as those already used with success to construct generation 1 of the already mentioned committee neural network potential [42]. Furthermore, the model was trained to reproduce only the difference between the full centroid force $-\frac{\partial W(x_i)}{\partial x}$ and the underlying Born–Oppenheimer force $-\frac{\partial U(x)}{\partial x}$, as the correction $W(x) - U(x)$ was assumed to be more short-range than either potential alone. Of course, when the CMD simulation is run, the Born–Oppenheimer force needs to be added again to this difference.

3.5 Numerical calculation of time correlation functions

As follows from the definition of a classical time correlation function (given in equation 2.51), numerical calculation of such a time correlation function would proceed by first performing an equilibrium simulation with a thermostat, yielding representative structures from the canonical ensemble. Second, these structures would each be used as an initial condition for a separate simulation *without* a thermostat (the number of these being a convergence parameter). From these simulations, products of the type $A(0)B(t)$ are calculated, and averaged over these separate simulations:

$$C_{AB}(t) = \frac{1}{N}\sum_{k=1}^{N} A(x_k(0))B(x_k(t)). \tag{3.18}$$

In these simulations only the $x_k(0)$ configuration belongs to the canonical ensemble, and thus each trajectory contributes only once to the statistical average of $C_{AB}(t)$ for each t. However, if a weak global thermostat were attached to each simulation instead, then the simulation would still reproduce the canonical ensemble over long timescales, while still yielding reasonable Hamiltonian dynamics at short timescales. The subsequent configurations $x(t)$ thus could be used as separate initial conditions for the purposes of calculating a time correlation function:

$$C_{AB}(t) = \frac{1}{NM}\sum_{k=1}^{N}\sum_{j=1}^{M} A(x_k(j\Delta t))B(x_k(j\Delta t + t)). \tag{3.19}$$

To avoid oversampling, the spacing Δt should be chosen such that the contributions are statistically decorrelated. If $A = B$, this is ensured by choosing $t > \tau_A$, τ_A being the correlation time of A. The most often used method of calculating time correlation functions is using this approach while setting $N = 1$ (or possibly, to another low number), sidestepping the need for an initial simulation to generate the canonical ensemble altogether [1]. In this approach, the most commonly chosen value of Δt is ϵ — the simulation time

step itself. The expression for the time correlation function thus becomes

$$C_{AB}(n\epsilon) = \frac{1}{K-n}\sum_{m=1}^{K-n} A(x(n\epsilon))B(x((n+m)\epsilon)), \tag{3.20}$$

$K\epsilon$ being the total simulated time. The time correlation function in this approach thus almost corresponds to a convolution — the difference being the normalization factor $1/(K-n)$. As such, the function can be efficiently calculated using a fast Fourier transform. It is clear, that if $K\epsilon$ is the length of the simulation, then the time argument of the correlation function takes on the values $\epsilon, 2\epsilon, \ldots, K\epsilon$. The value of the time correlation function at each of these points is a statistical average of $K, K-1, \ldots, 1$ contributions respectively. The statistics thus becomes progressively worse the higher the argument t is. These contributions may simply be discarded, but a better approach might be chosen, anticipating that the Fourier transform of the time correlation function will be calculated (such as the VDOS, or the infrared spectrum, see equation 2.117). As the number of frequencies present in the Fourier spectrum is equal to the number of data points in the original time correlation function, simply leaving out the high t data would result in a spectrum with a low resolution (the maximum frequency is given by ϵ via the Nyquist–Shannon sampling theorem [54]). The high t data might simply be replaced with zeros (zeros might even be added beyond the original simulation length), but this would introduce a sharp jump into the correlation function, which would result in artifacts in the spectrum. A better choice is to apply a smooth *windowing function* to the time correlation function:

$$C_{AB}(t) \leftarrow C_{AB}(t)w(t). \tag{3.21}$$

The windowing function $w(t)$ should have the following properties:

$$\begin{aligned} w(t) &= 1 \quad \text{for} \quad t \lesssim t_{\max} \\ w(t) &= 0 \quad \text{for} \quad t \gtrsim t_{\max}, \end{aligned} \tag{3.22}$$

where $t_{\max}$ is the largest timescale of interest. The function should change slowly enough, such that artifacts are not introduced into the resulting spectrum. This procedure is referred to as *apodization* within signal processing.

If one has access to the eigenvalues and eigenstates of the Hamiltonian $\hat{H}$ of the studied system, it is also possible to calculate the Kubo-transformed time correlation from its definition in equation 2.101. Note that this is almost never the case for condensed-phase molecular systems, however, this approach can still be used as a reference when studying model systems of low dimensionality. Using the eigenstates as an orthonormal basis, equa-

tion 2.101 becomes (dropping the zero denoting the equilibrium ensemble):

$$\begin{aligned} K_{XY}(t) = \frac{1}{\beta Z}\sum_{AB}\int_0^\beta \mathrm{d}\lambda \exp[\lambda(E_B - E_A)]\,\langle E_B|\,\hat{X}\,|E_A\rangle \\ \times \exp\left[\frac{i}{\hbar}(E_A - E_B)\,t\right]\langle E_A|\,\hat{Y}\,|E_B\rangle \exp\big(-\beta E_B\big), \end{aligned} \tag{3.23}$$

the partition function Z being simply:

$$Z = \sum_A \exp\big(-\beta E_A\big). \tag{3.24}$$

The summations extend over the, in principle, infinite number of energy eigenstates, however, as the contributions of high-energy states are exponentially suppressed, this sum may be safely truncated at a finite number of states and carried out numerically. The integral over λ in equation 3.23 can be simply carried out:

$$\begin{aligned} K_{XY}(t) = \frac{1}{\beta Z}\sum_{AB}\frac{1}{E_B - E_A}\left[\exp\big(-\beta E_B\big) - \exp\big(-\beta E_A\big)\right] \\ \times \langle E_B|\,\hat{X}\,|E_A\rangle \cdot \exp\left[\frac{i}{\hbar}(E_A - E_B)\,t\right]\langle E_A|\,\hat{Y}\,|E_B\rangle\,. \end{aligned} \tag{3.25}$$

The last thing which needs to be addressed before implementing this calculation numerically is the evaluation of the $E_A = E_B$ terms in the sum. These involve a zero both in the denominator and the numerator, and as such cannot be evaluated on a computer. The relevant limit may however be carried out:

$$\begin{aligned} &\lim_{E_B\to E_A}\frac{1}{E_B - E_A}\left[\exp\big(-\beta E_B\big) - \exp\big(-\beta E_A\big)\right] \\ &= \lim_{E_B\to E_A}\frac{\exp\big(-\beta E_A\big)}{E_B - E_A}\left[\exp\big(-\beta(E_B - E_A)\big) - 1\right] \\ &= \lim_{E_B\to E_A}\frac{\exp\big(-\beta E_A\big)}{E_B - E_A}\left[-\beta(E_B - E_A) + O((E_B - E_A)^2)\right] \\ &= -\beta\exp\big(\beta E_A\big) \end{aligned} \tag{3.26}$$

where E_A was chosen to denote the common energy.

A useful tool in gaining insight into the properties and origins of features in a frequency spectrum is the *2D spectrum*, defined as:

$$\Gamma_A(\omega,\omega') = \frac{1}{t}\int_0^t \mathrm{d}t'\,A(\omega,t')\,A(\omega',t'), \tag{3.27}$$

where $A(\omega, t)$ is the *time-dependent* spectrum of the quantity A (a similar concept is also often used in sound processing, where it is known as a *spectrogram*). This spectrum is defined as:

$$A(\omega, t) = \int_{-\infty}^{\infty} \mathrm{d}t' C_{AA}^{[t]}(t') w(t' - t) \exp\bigl(-i\omega t'\bigr), \tag{3.28}$$

where $C_{AA}^{[t]}$ is the autocorrelation function of the quantity A, with the time t taken as the reference (zero) time. The windowing function is for this purpose taken to be rather narrow (compared to its use in apodization). The purpose of this windowing function is to restrict the spectrum, such that it captures the properties of the system in its state around the time t. As an example, the frequency spectrum of a molecule is dependent on its surrounding environment, which changes over time. In the calculation of the time-independent frequency spectrum, this dependence gets averaged out, which results in the broadening of the features present in the spectrum — an effect known as *inhomogeneous broadening*. In the time-dependent frequency spectrum, the effect of the surrounding environment is kept, being reflected by the dependence of the spectrum on the time t. The 2D spectrum then gives the degree of correlation between various features in the time-dependent spectrum. In a numerical application, the 2D spectrum can be calculated by discretizing the integral:

$$\Gamma_A(\omega, \omega') = \frac{1}{t} \sum_{n=1}^{t/\epsilon} A(\omega, n\epsilon) A(\omega', n\epsilon), \tag{3.29}$$

and the time-dependent spectrum can be calculated using a fast Fourier transform of the discretization of $C_{AA}^{[t]}(t') w(t' - t)$. Lastly, if two different time-dependent spectra are available — these may correspond, for example, to different atomic species each, a 2D *cross*-spectrum may be calculated analogously using

$$\Gamma_{AA'}(\omega, \omega') = \frac{1}{t} \sum_{n=1}^{t/\epsilon} A(\omega, n\epsilon) A'(\omega', n\epsilon), \tag{3.30}$$

where A and A' denote the two time-dependent spectra. The 2D cross-spectrum can analogously be interpreted as giving the degree of correlation between features in the two time-dependent spectra A and A'.

All methods within this subsection were implemented in the Python programming language, in particular using the NumPy library for numerical calculations [55]. To calculate the VDOS and 2D spectra of molecular systems an existing in-house Python library was used. The Fourier DVR Python code [56] was used to compute the energy eigenvalues and eigenfunctions of model one-dimensional and was extended in this work to obtain these quantities for two-dimensional systems as well. These were needed to compute the Kubo-transformed time correlation function according to equation (3.25). This method of calculating the function was also implemented in this work.

4 Results and discussion

4.1 Model systems

This subsection will work in reduced units, determined by the system parameters. They will thus not be explicitly written.

Four different simple model potentials were considered:

1. A one-dimensional harmonic oscillator:

$$U(x) = \frac{1}{2}x^2 \tag{4.1}$$

2. A one-dimensional quartic oscillator:

$$U(x) = \frac{1}{4}x^4 \tag{4.2}$$

3. A one-dimensional double-well potential:

$$U(x) = \frac{1}{4}(x^2 - 4)^2 \tag{4.3}$$

4. A two-dimensional Morse potential:

$$U(x, y) = D_0 \left\{ 1 - \exp\left[-\alpha \left(\sqrt{x^2 + y^2} - r_{\text{eq}} \right) \right] \right\}^2 \tag{4.4}$$

For all systems, work has been carried out in atomic units. For the two-dimensional potential, the three parameters were based on publication 57 and were $D_0 = 0.18748, \alpha = 1.1605, r_{\text{eq}} = 1.8324$. Additionally, the mass of the particle, in this case, was also chosen according to ref. 57 as $m = 1741.05198$, while for the one-dimensional systems, the mass was chosen as $m = 1$.

These potentials represent cases where CMD is known to be exact (the harmonic potential), where it is not exact but still reasonably sufficient (the quartic potential), and where CMD is known to break down or miss important physical processes (the double well and Morse potentials). The Morse potential is of particular interest as it can be thought of as a simplified model of an OH bond [57] — a common feature present in many molecular systems of interest. The performance and behavior of both implementations of CMD will be explored in these cases.

In the investigation of model systems, attention will be directed mainly to the position autocorrelation functions. For such small systems, these are simple enough, such that they can be easily read, and they contain valuable information on the underlying dynamics. The centroid correction to the force

and to the potential, defined as the difference between the centroid potential/force and the actual "physical" potential/force, will also be studied.

For all one-dimensional potentials, centroid molecular dynamics was performed by first calculating the average forces on the centroid fixed at points on a regular grid from $x = -10$ to $x = 10$ with a step of 0.1. The average force on the centroid was determined by averaging over 375 independent path integral simulations (with the centroid fixed) of length $t = 200$ with the first 10 % of the simulations left for equilibration. The simulations used 32 replicas and had the non-centroid normal modes set to a common frequency of 2, such that the time step could be kept at a relatively low value of 0.01. The forces were sampled every 10 simulation steps. The average forces were then interpolated with cubic splines using the SciPy library [53]. The interpolated force was subsequently used to drive molecular dynamics simulations necessary to calculate position autocorrelation functions.

For the two-dimensional potential, the procedure was the same but due to computational constraints, the grid was restricted to the area $x, y = -5$ to $x, y = 5$ with a step of 0.25 in both directions. The number of independent simulations was also reduced from 375 to 125. Additionally, as this system was in a quite different region of parameters from the one-dimensional systems the simulation parameters had to be adjusted. The simulations had a length of $t = 4000$ and a time step of 0.1, with the first $t_{eq} = 2100$ discarded for equilibration. The equilibration was observed to be more difficult, and as such longer equilibration period was used for safety. The target frequency used was 0.2. The number of replicas and sampling frequency of the force were kept at the values of 32 and 10 respectively.

The autocorrelation functions of simple systems were calculated according to the "strict" scheme, as outlined at the beginning of subsection 3.5. There are two main reasons behind this choice. First, for a system of non-interacting particles, a global thermostat does not make sense, and performing the "one-simulation" scheme would be rather problematic with a local thermostat. And second, the simulations of low-dimensional systems are simple enough, such that performing large amounts of them is not a drawback, but in fact, a benefit, as it allows one to parallelize them trivially.

To be more specific, the scheme to calculate a single time-correlation function was thus:

1. For the case of one-dimensional systems 375 source simulations were performed. These had a time length of $t = 100$ and a time step of 0.01. For the case of the two-dimensional system, 125 source simulations of length $t = 200000$ and a time step of 10 were used. From these, the first half has been discarded. All simulations were thermostatted with a Langevin thermostat with a time constant of $\gamma = 1$.

2. For the one-dimensional systems, every 500 steps, starting from step

2000, an initial condition was sampled. For the two-dimensional system this was done every 5000, starting from step number 20000.

3. From every initial condition, an unthermostatted simulation was started (the total number of these was thus 6000 per one-dimensional system and 2000 for the two-dimensional system). For the one-dimensional systems these had the same length and time step as the source simulation. For the two-dimensional system these had a length of $t = 10000$ and a time step of 1.

4. From these simulations, structures were sampled at each step. The value of the time correlation function at each of these time points was calculated as an average over all contributions according to equation 3.18.

In addition to these simulations, reference simulations performed according to the adiabatic CMD scheme were also performed. These followed the same overall scheme as was outlined in the previous four points. The difference was that all of the simulations are path-integral molecular dynamics simulations and thus needed to be adjusted accordingly. For the one-dimensional systems, the source simulations used 32 replicas, target frequency for the non-centroid modes equal to 2, length of $t = 200$, and a time step of 0.01. The first $t_{\mathrm{eq}} = 100$ of the simulation were discarded, and from this point on the initial structures were sampled in the same fashion as outlined before. The "production simulations" used the same number of replicas, a target frequency of 1000, a time step of 0.0001, and a length of $t = 100$. The structures were sampled with an interval of 100 simulation steps. For the two-dimensional system, the source simulation also used 32 replicas alongside a target frequency of 0.000002, the rest of the parameters and the sampling were the same, as in the interpolated case. The "production simulations" used the same number of replicas, a target frequency of 10, a time step of 0.01, and a simulation length of $t = 10\ 000$. Structures were sampled from these every 100 simulation steps.

Reference results obtained by diagonalizing the Hamiltonian and using matrix multiplication to calculate the Kubo transformed function according to equation 3.25 were also performed.

The harmonic oscillator

The quantum dynamics of the harmonic oscillator can be exactly described by CMD [10]. The system thus presents an opportunity to study the inherent differences between the two implementations of CMD, and their deficiencies, free from the inherent shortcomings of CMD or anharmonicities through which the non-centroid degrees of freedom could exchange energy with the centroid in the adiabatic implementation. As can be seen from figures 1 and 2, both implementations of CMD yield the expected result and

are able to reproduce the exact Kubo-transformed autocorrelation function (which in this case is just a harmonic function). The only slight deficiency which can be seen is the slight dampening of the adiabatic autocorrelation function. Due to the elimination of other sources of influence, this deviation has to come inherently from the "averaging on the fly" procedure which underlies the adiabatic implementation. Note however that this dampening is so small that it is for all practical purposes irrelevant. Overall, these results are encouraging, and show the implementation proposed in this work to be in this case consistent with both exact results and the previously used adiabatic implementation, while doing away with the difficulties and additional problems associated with the adiabatic implementation.

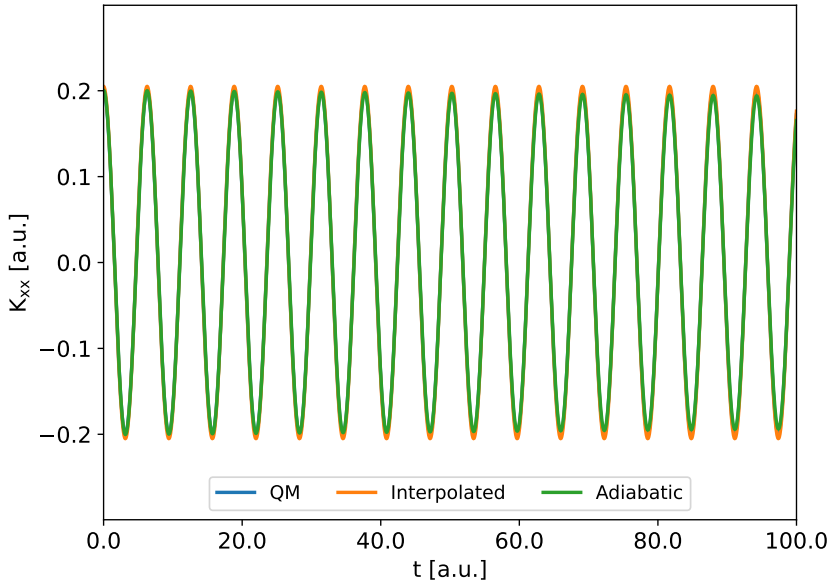

Figure 1: Position autocorrelation functions of the harmonic oscillator at a temperature $T = 0.2$ calculated using equation 3.25, as well as the two different implementations of CMD.

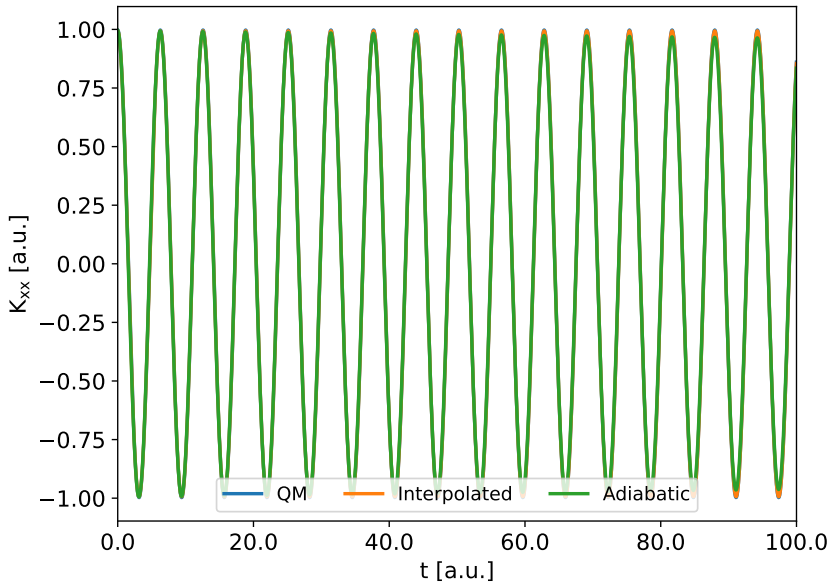

Figure 2: Position autocorrelation functions of the harmonic oscillator at a temperature $T = 1.0$ calculated using equation 3.25, as well as the two different implementations of CMD.

The quartic oscillator

The quartic oscillator is a much more difficult case to describe correctly using CMD, due to its strongly anharmonic nature. Despite this, CMD can be seen to perform at least partially well. At every considered temperature, the position autocorrelation functions calculated using both implementations of CMD follow the numerically exact result for at least one oscillation. For the subsequent oscillations the agreement breaks down, the CMD time correlation functions becoming damped. The agreement with the numerically exact time correlation function is better at lower temperatures, where the CMD time correlation functions, despite being damped, still agree well regarding the "main" oscillation frequency. Note, however, that certain subtle features are not reproduced by any implementation of CMD at all. This fact can be most easily seen from the VDOS (figure 3), where peaks present in the VDOS of the numerically exact solution are not present in the approximate CMD VDOSes at all. Both implementations of CMD also yield autocorrelation functions consistent with each other, the agreement breaking down slightly with increasing temperature. The disagreement was ascribed to the residual non-adiabaticity of the otherwise almost adiabatic CMD simulations, caused by the smearing of the frequency range of the non-centroid normal modes by the thermostats. As the temperature increases, this smearing gets broader and thus results in a progressively worse agreement. Note that the anhar-

monicity of the potential is key for this source of disagreement to come into play, as the non-centroid normal modes can only exchange energy with the centroid through the anharmonicities in the potential. As such, this effect was not present in the previous case of the harmonic oscillator.

The fact that CMD is more accurate at lower temperatures for the case of the quartic oscillator is a known, but rather unexpected one. Acknowledging the fact that CMD is exact in the classical limit and that CMD tends to yield poor results at low temperatures (this fact will be discussed alongside the following two simple models, which exhibit examples of this behavior), one would expect the agreement to increase with temperature. This disagreement with expectation was discussed previously by Pérez, Tuckerman, and Müser in ref. 58. There, they suggested that the unexpectedly good results are caused by the fact that at low temperatures only the first two energy eigenstates are thermally available for the dynamics and the problem is thus "effectively harmonic" — a problem that CMD can treat well.

From the previous discussion, one could thus expect that for at least simple potentials, CMD will exhibit reasonable agreement with exact results and between both of its implementations at low temperatures, where the problem is "effectively harmonic", while breaking down for higher temperatures where the anharmonicity of the system starts to come into play. Note, however, that for temperatures high enough that the system can be treated classically, the agreement between CMD and the numerically exact quantum solution has to be restored, as both should yield classical dynamics in this limit.

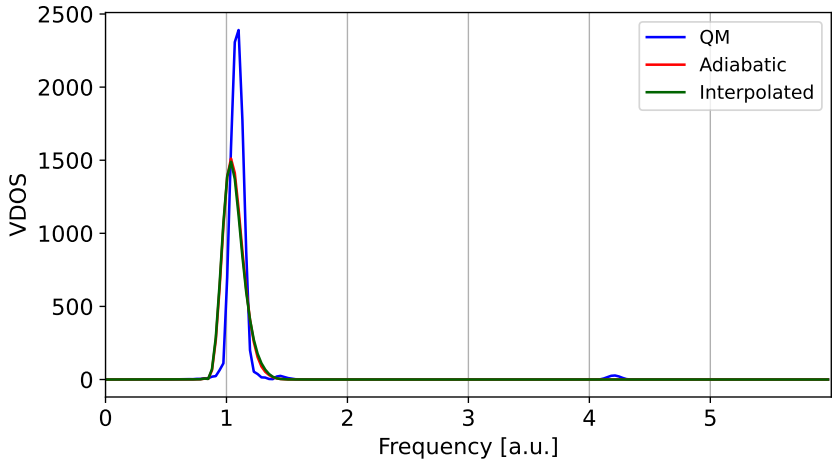

Figure 3: The VDOS of the quartic oscillator at a temperature $T = 0.2$ calculated from the velocity autocorrelation function calculated using equation 3.25, as well as the two different implementations of CMD.

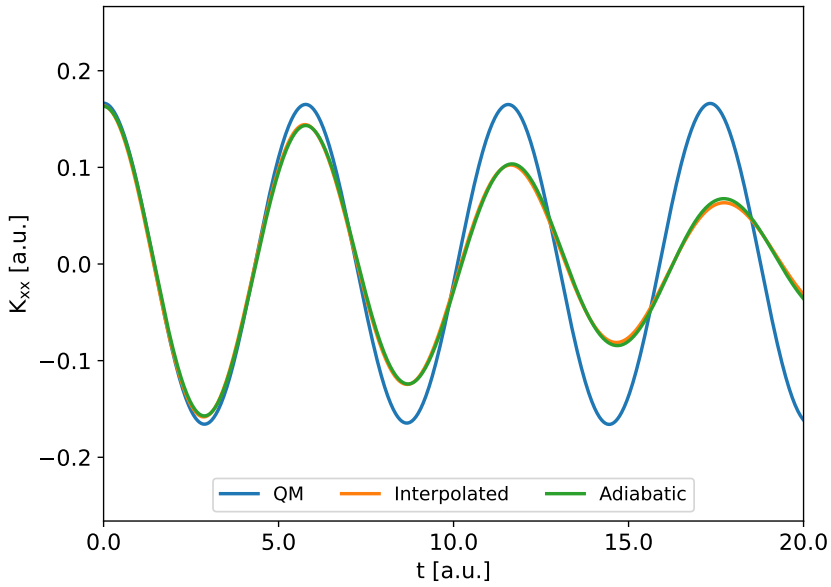

Figure 4: Position autocorrelation functions of the quartic oscillator at a temperature $T = 0.2$ calculated using equation 3.25, as well as the two different implementations of CMD.

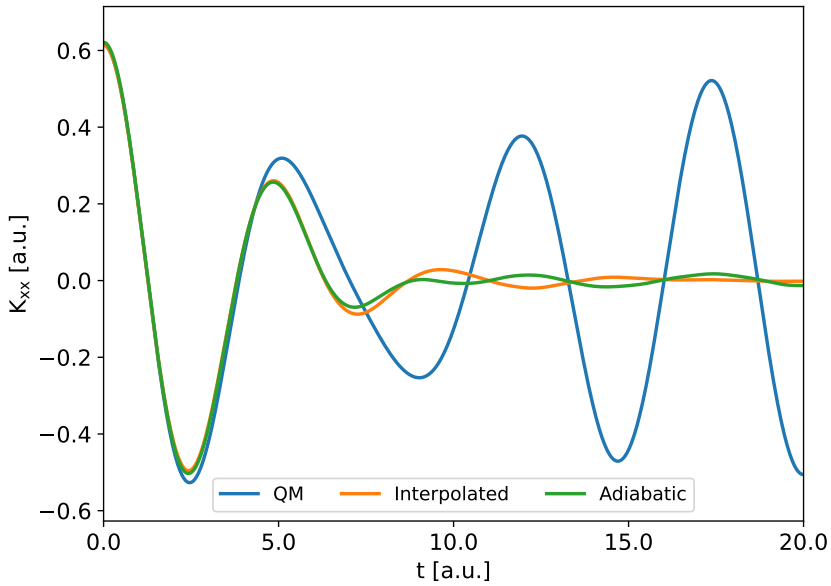

Figure 5: Position autocorrelation functions of the quartic oscillator at a temperature $T = 1.0$ calculated using equation 3.25, as well as the two different implementations of CMD.

For this potential, the position autocorrelation functions were published

in the past [58]. The presently proposed approach to performing CMD can thus be checked against these results. As can be seen from figure 6, results obtained both using the interpolated and adiabatic CMD within this work are consistent with the previously published result.

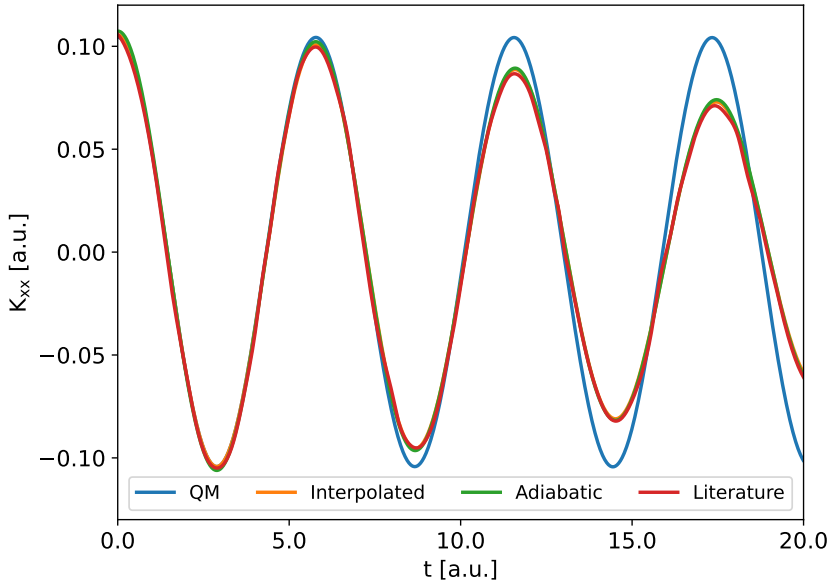

Figure 6: A comparison of the position autocorrelation function of the quartic oscillator at a temperature $T = 0.125$. The four curves correspond to the result obtained using equation 3.25, the two different implementations of CMD, and a result published in [58].

An advantage of the proposed "interpolated" approach to CMD is the fact that both the forces and the potential are readily available and thus can be easily analyzed. As a demonstration of this, the centroid correction to the quartic potential, as a function of temperature and position, can be found in figure 7, and the centroid correction to the corresponding force, as a function of the same variables, can be found in figure 8. The correction, for the case of the quartic oscillator, can be seen to make the potential more confining. The effect is qualitatively the same across all temperatures, but its magnitude decreases with increasing temperature (at least for positions close to $x = 0$).

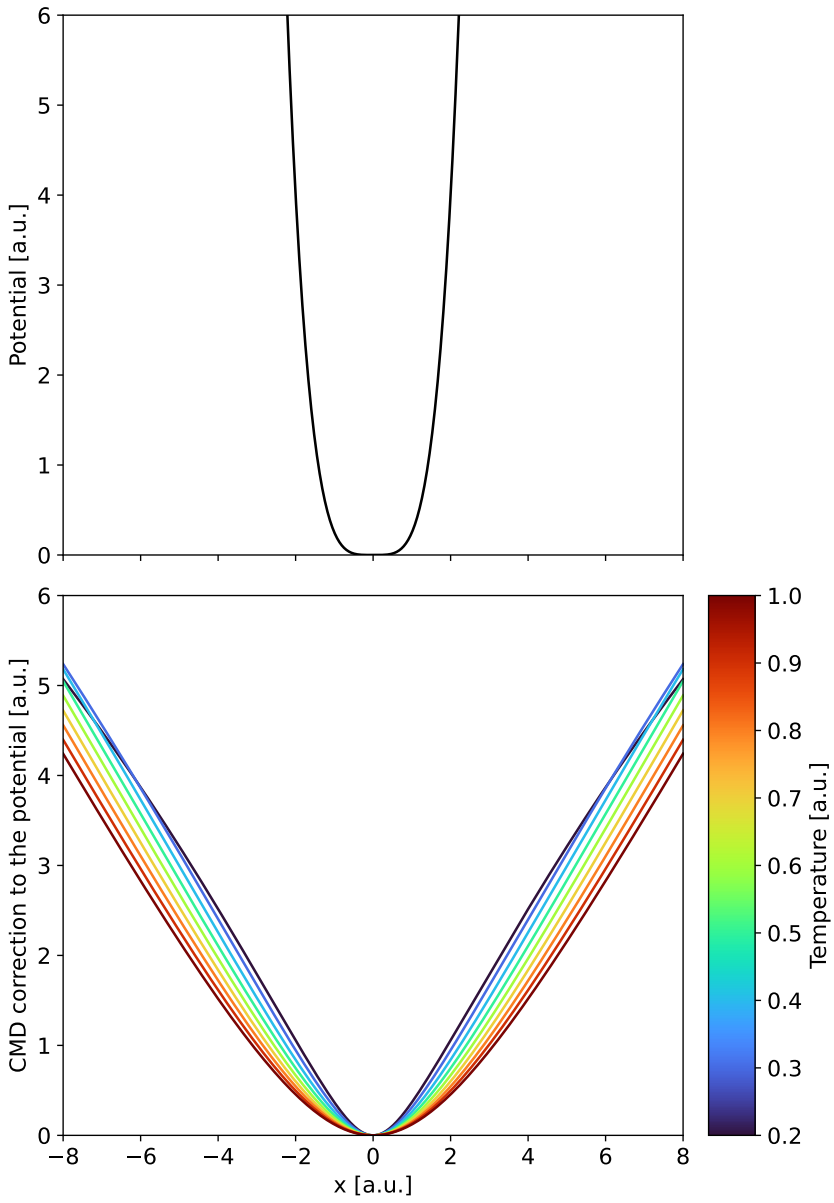

Figure 7: The dependence of the centroid correction to the quartic potential on the position and temperature. The top panel gives the quartic potential itself as a reference. All potentials have been aligned such that they are equal to 0 at $x = 0$.

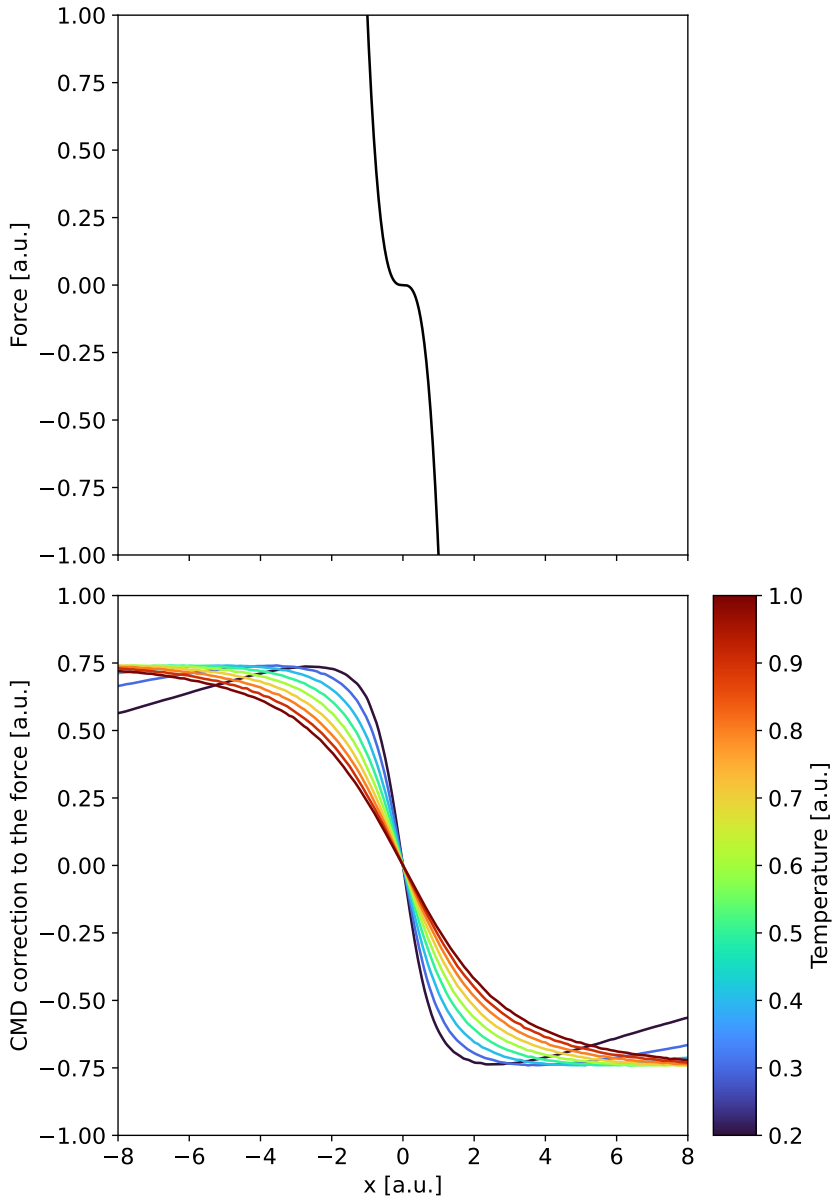

Figure 8: The dependence of the centroid correction to the force of the quartic potential on the position and temperature. The top panel gives the force of the quartic potential itself as a reference.

At this point, a slight detour which will demonstrate the importance of the adiabatic separation will be made. This will be showcased on the two previously shown one-dimensional models. The previously shown results for the adiabatic implementation were obtained using a generous target frequency of 1000. If this target frequency is systematically lowered (keeping all other parameters the same), and the adiabaticity of the simulation thus decreased,

the following results are obtained. For the harmonic oscillator, decreasing the target frequency has, at least for the considered frequencies no effect on the autocorrelation functions, regardless of the temperature (figures 9 and 10). For the quartic oscillator at low temperatures, the autocorrelation functions can be seen to become progressively more damped, systematically deviating from the numerically exact result and the result obtained from the interpolated implementation (figure 11). At higher temperatures, all results can be seen to be "equally bad" when it comes to reproducing the numerically exact result (figure 12).

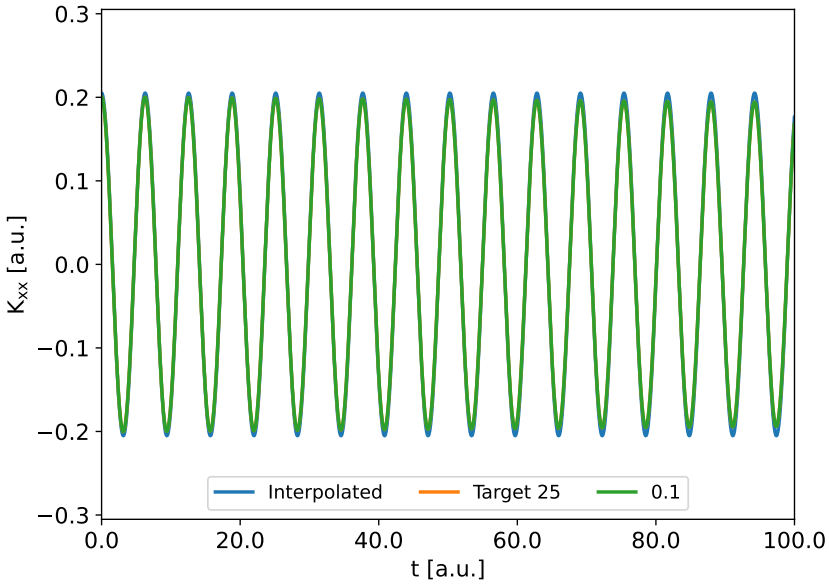

Figure 9: Position autocorrelation functions of the harmonic oscillator at a temperature $T = 0.2$ obtained using the "interpolated" implementation of CMD and the adiabatic CMD with different settings of the target frequency.

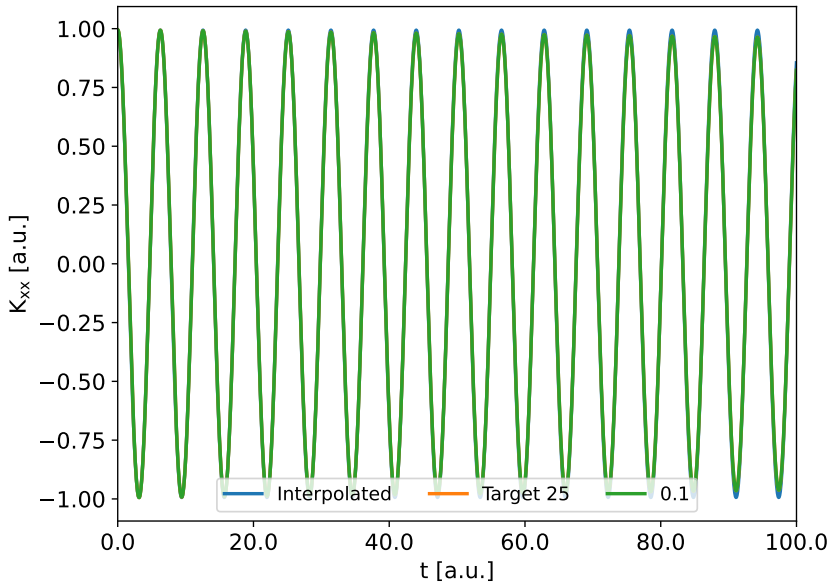

Figure 10: Position autocorrelation functions of the harmonic oscillator at a temperature $T = 1.0$ obtained using the "interpolated" implementation of CMD and the adiabatic CMD with different settings of the target frequency.

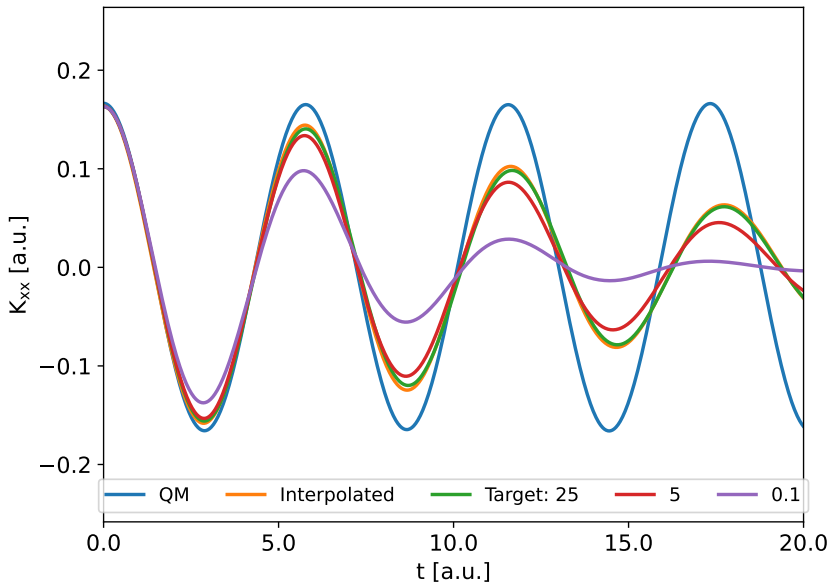

Figure 11: Position autocorrelation functions of the quartic oscillator at a temperature $T = 0.2$ obtained using the "interpolated" implementation of CMD and the adiabatic CMD with different settings of the target frequency.

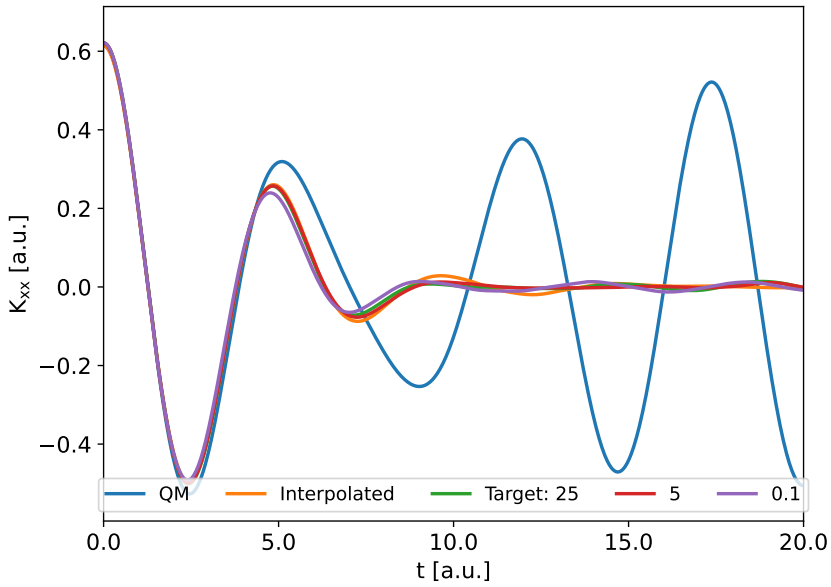

Figure 12: Position autocorrelation functions of the quartic oscillator at a temperature $T = 1.0$ obtained using the "interpolated" implementation of CMD and the adiabatic CMD with different settings of the target frequency.

The influence of the non-centroid modes on the centroid, in the case of the quartic oscillator at the lower temperature, can also be very well illustrated using the 2D cross-spectra. Roughly speaking, these give the degree of correlation between the motion of one degree of freedom at a frequency ω and the motion of another degree of freedom at a frequency ω' (for more details see subsection 3.5). Stated in another way, this spectrum gives the degree of correlation between the intensities (at the two given frequencies) in the respective VDOSes of the two degrees of freedom. The two degrees of freedom considered here were the centroid and one of the non-centroid normal modes (all of which are equivalent, as they share a common frequency). As this spectrum is used only for demonstration purposes, and the computational expenses associated with calculating such spectra are high, the resulting 2D spectrum has been averaged over only 375 different trajectories. The trajectories themselves are a subset of those which were used to calculate the autocorrelation functions in figure 11. This procedure has been carried out for the three different target frequencies considered, and the results can be seen in figure 13. In this figure, the physical degrees of freedom are located near the origin (recall that according to figure 3 the peak in the VDOS of the centroid is located at roughly 1 a.u.), while the non-centroid normal mode frequency is located further to the higher frequencies, getting gradually closer to the physical ones with decreasing target frequency. As can be seen, at decent adiabatic separation the centroid is weakly coupled to the non-centroid

normal mode, mainly through motion at frequencies of about 3 a.u. As the adiabatic separation gets lower, so does this correlation get higher, finally arriving at a situation where the target frequency and physical frequencies overlap.

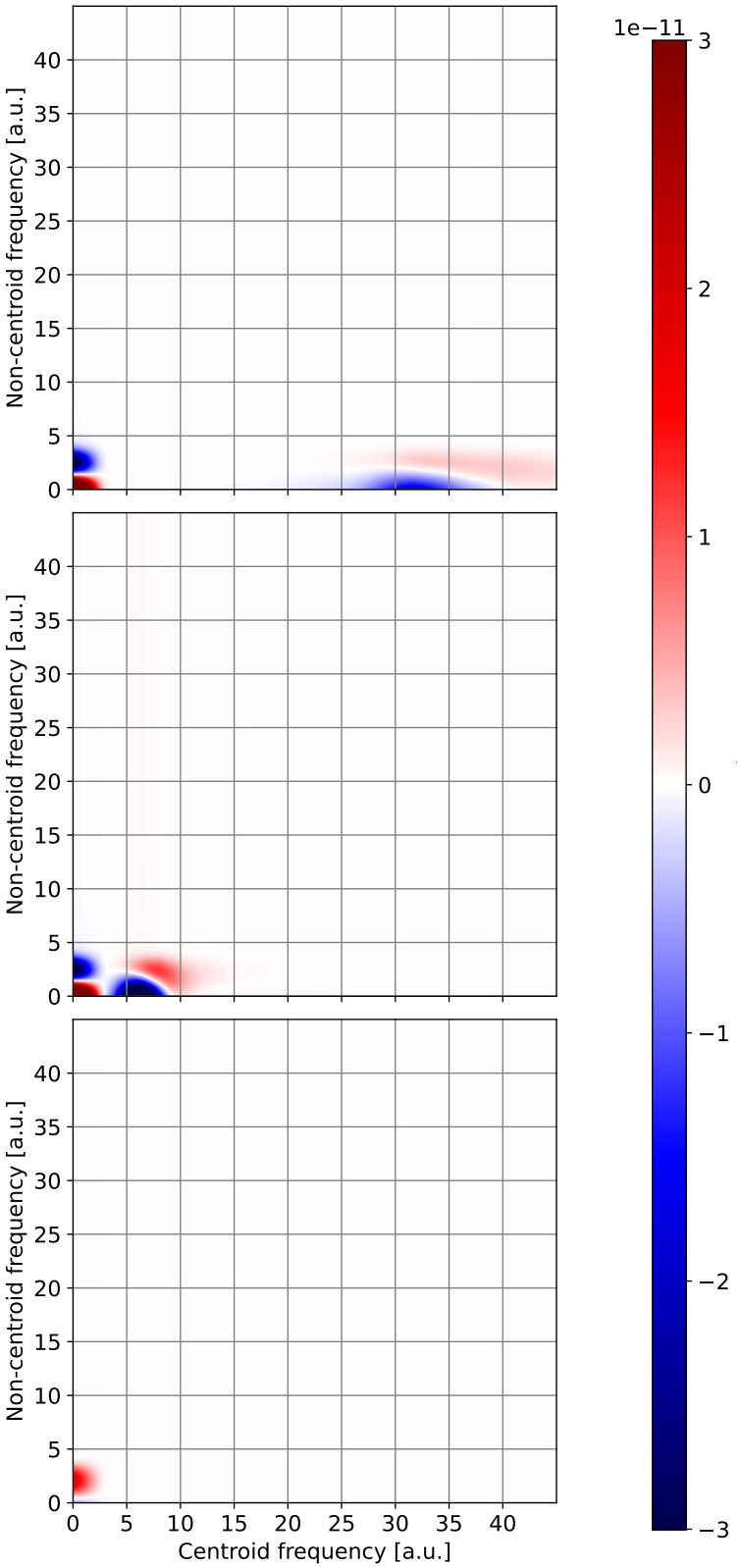

Figure 13: 2D cross spectra between the centroid and one of the equivalent non-centroid normal modes, calculated for the quartic oscillator. The three panels correspond to normal mode target frequency, going from top to bottom, of 25, 5, and 0.1 [a.u.].

These results underline the important role the anharmonicity plays in degrading the adiabaticity of the simulations. The results also showcase the fact that the adiabatic separation necessary to obtain a converged result is dependent on the temperature. In particular, the required separation decreases with temperature. This is to be expected, as at very high temperatures the adiabatic implementation should yield the classical result regardless of the adiabatic separation, as should all other methods.

The double well potential

For the double-well potential, a deficiency at low temperatures of not only CMD, but the whole class of methods based on the imaginary-time path-integral can be observed. The double well potential has a well-known spectrum in which, for low quantum numbers, states appear in nearly degenerate pairs (the two states in the pair being of even and odd parity, respectively). The small split between these two states is responsible for slow oscillations in which the particle tunnels from one well into the other. This split (and the corresponding motion) is notably difficult for all methods, which start with the classical limit as a reference, to treat correctly (in fact, the split itself can be shown to be non-perturbative in $\hbar$). Such is the case with CMD as well — see figure 14. CMD is able to approximate only the short time scale oscillations while failing to capture the slow oscillations completely. Note that these slow oscillations happen at longer timescales (see figure 15) and as such do not appear in full in figure 14. For the short time scale oscillations, a similar discussion as for the previous potential could be made. However, note that the agreement here is in fact better, as the potential is locally more harmonic. From the practical point of view, this deficiency can be expected to degrade the performance of CMD in predicting quantities such as reaction rate constants, in which similar effects can play a role. The problem can, however, be solved while still using the general theory of imaginary-time path-integrals using the concept of *instantons* [59].

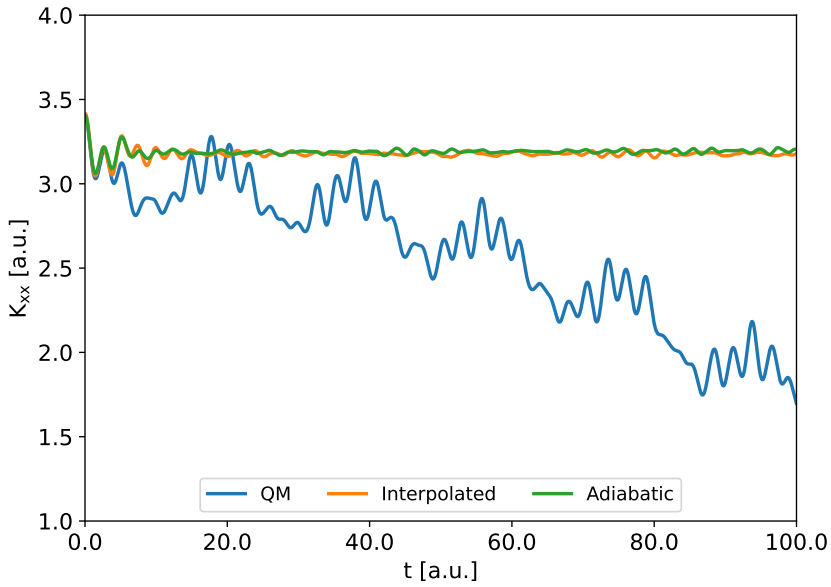

Figure 14: Position autocorrelation functions of the double well potential at a temperature $T = 1.0$ calculated using equation 3.25, as well as the two different implementations of CMD.

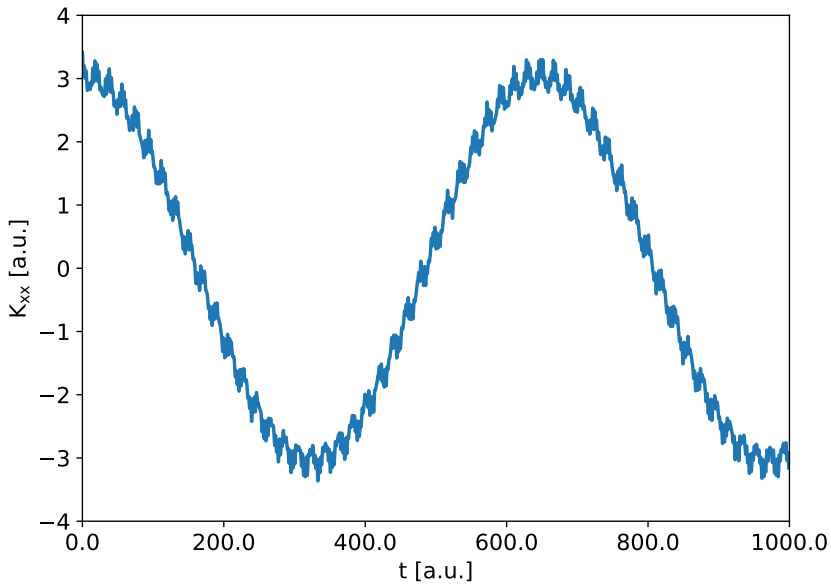

Figure 15: Position autocorrelation function of the double well potential at a temperature $T = 1.0$ calculated using equation 3.25.

The 2D Morse potential

For the 2D Morse potential, which can be thought of as a simple proxy for the OH molecular system, CMD exhibits a further deficiency known as the *curvature problem* [60]. The curvature problem is perhaps the most relevant drawback of CMD when it comes to its usage as a tool to calculate vibrational spectra of molecular systems, as it causes artificial redshifts of peaks corresponding to vibrations of OH bonds (which are very common) in the calculated infrared spectra. The curvature problem has its origin in the fact that at lower temperatures, the path integral ring polymers tend to be more spread out (the harmonic couplings become weaker at lower temperatures). For potentials that confine the ring polymer to move along a circle or a semi-circle, this results in the centroid (calculated as the average of the cartesian coordinates of the replicas) to appear in a classically highly improbable position. As an example, for the 2D Morse potential in the extreme case of the ring polymer spread out evenly over the whole circle, this in fact results in the centroid being located at the potential maximum.

For this potential, the position autocorrelation will be abandoned in favor of the VDOS. The reason for this is that the autocorrelation functions already for this potential start to be rather complicated and that the VDOS best captures the curvature problem present in CMD. The VDOSes for this potential were calculated from the position autocorrelation functions using equation 2.119, employing the Hann windowing function of width 16000 a.u. and an additional 100000 a.u. worth of padding with zeroes. Calculating the VDOS from the position autocorrelation function was found to possess better numerical properties for this case than calculating it from the momentum autocorrelation function. The VDOS of this model system, in the considered temperature range, contains two main features — a low-frequency rotation peak and a high-frequency vibration peak. As can be seen from figure 16, both implementations yield consistent results, with some differences. Both implementations suffer from the curvature problem, but the interpolated implementation can be seen to suffer more from this problem. For $T = 0.0015$ the vibration peak of the VDOS calculated using the interpolated implementation is blue-shifted with respect to both the adiabatic implementation and the “numerically exact” reference VDOS calculated using 3.25.

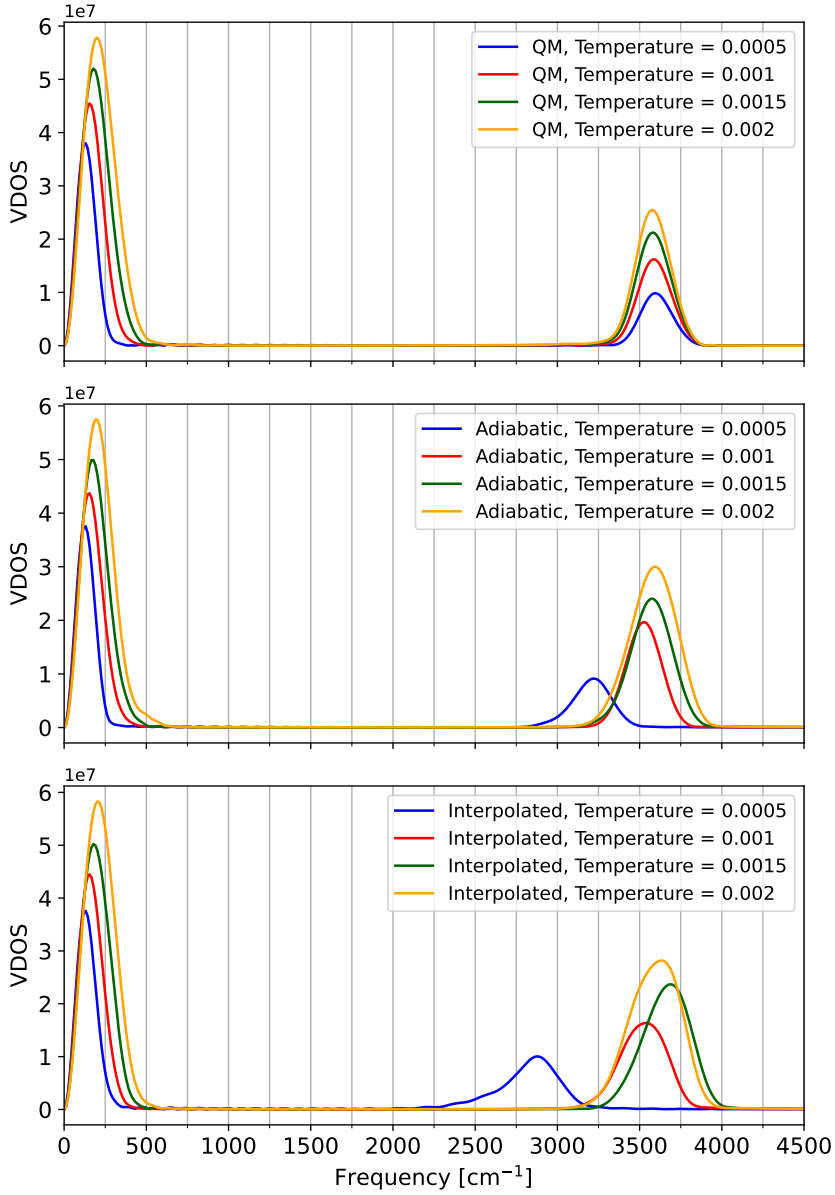

Figure 16: The VDOS of the 2D Morse potential at a sequence of temperatures. The three panels correspond to VDOSes calculated from the position autocorrelation function obtained using equation 3.25, as well as the two different implementations of CMD.

Several propositions have recently been put forward which seek to solve the curvature problem. These include the so-called quasi-centroid molecular dynamics (QCMD), based on path integrals in curvilinear coordinates [61], as well as the "T-PIGS" approach [62]. According to this method, a centroid potential calculated at a higher temperature should be used to run CMD at

a lower temperature. The basis for doing so is the fact that the vibrational modes present in the system are, for the temperature ranges of interest, effectively in the ground state. The shape of the potential, at least in the direction of the vibrational modes, thus should not be expected to change with decreasing temperature. This justifies one in taking the higher-temperature centroid potential, where the curvature problem is not yet present, but the vibrational modes are already in the ground state, and using it for lower temperatures. Note, however, that in the publication 62, the centroid potential is not really used, instead an effective potential obtained through methods of *coarse graining* for running CMD is used. In this work, it is assumed that this discussion can be applied to the centroid potential itself. This approach has been tested by taking the $T = 0.002$ centroid potential and performing CMD simulation for all four temperatures using this potential (the rest of the procedure was the same). The results can be found in figure 17. As can be seen, this approach indeed fixes the curvature problem for the vibrational peak, while leaving the temperature dependence the same. The fact that the proposed implementation of CMD can be easily combined with the "T-PIGS" method to remove the curvature problem is thus another possible advantage. Some care needs to be taken, however, when applying this method to molecular systems. This will be discussed in the next subsection.

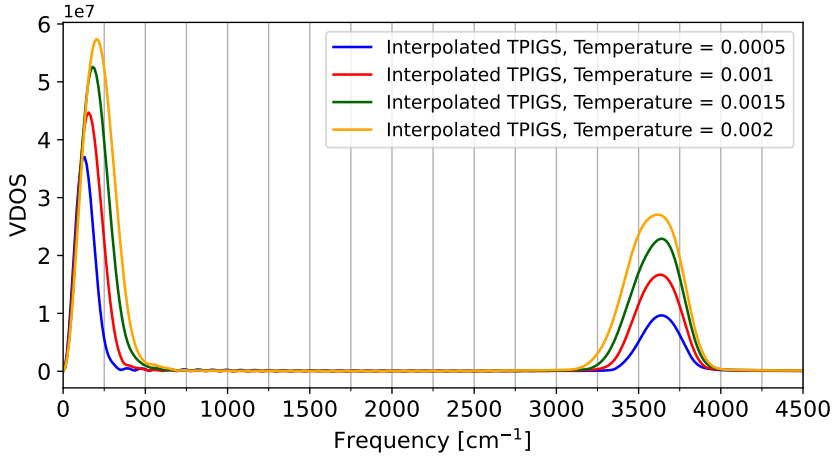

Figure 17: The VDOS of the 2D Morse potential at a sequence of temperatures calculated using the "T-PIGS" approach.

4.2 Molecular systems

For molecular systems, three different systems were considered — liquid water at 300 K, liquid heavy water at 300 K, and liquid water at 350 K. All systems were composed of 64 water molecules simulated in a cubic box of side length

12.42 Å. Periodic boundary conditions were employed, as is usually done to simulate a bulk liquid environment. Note, that this implies the omission of thermal volume expansion for the system at 350 K, however, this is justified as this effect is negligible over these temperature ranges. For all simulations, the previously mentioned machine learning model of the Born–Oppenheimer ground state potential energy surface was used [42].

The two liquid water systems were picked due to the ubiquitous, yet non-trivial nature of water. Two temperatures were chosen to show the temperature dependence of the results. The significance of the heavy water system comes from the fact that one of the most studied and readily apparent implications of the quantum nature of atomic nuclei is the *isotope effect*. This is a name for the change of behavior of substances under the exchange of isotopes of the constituent atoms. Within classical statistical mechanics, this effect should be in principle simple — equilibrium properties should be changed trivially, while non-equilibrium properties should be changed in a simple manner due to the fact that the inertial masses of the atoms are different. In quantum statistical mechanics, the effect is generally non-trivial and plays a role even at the level of static equilibrium properties (as an example, the freezing point of heavy water is approximately 4 K above the freezing point of water). For a general overview of the role of nuclear quantum effects in liquid water, including the isotope effect, see for example ref. 63.

The main object of interest in the study of these systems was the hydrogen atom VDOS, as it shares the main features of the experimentally measurable infrared spectrum and provides a great deal of information about the dynamics of the systems while remaining easily readable (the autocorrelation functions of realistic molecular systems tend to be too complex to be of use for this purpose). Attention was also given to the analysis of the centroid correction to the physical potential.

For each system, centroid molecular dynamics was performed according to the following protocol:

1. 111 path-integral molecular dynamics simulations with the centroids of all atoms fixed at different positions were performed. The positions were the same as used in the generation 1 training set for the machine learning model of the liquid water electronic ground state constructed in the original work [42]. Each of these simulations had a length of 10 ps, with the first 4 ps left for equilibration, and a time step of 0.2 fs. The simulations used 32 replicas and all non-centroid normal modes were shifted to a common frequency of 2000 cm^{-1} to allow a longer time step. Forces on the centroids were sampled every 8 fs.

2. An average over the sampled forces was performed, thus yielding 111 samples of converged centroid forces for all 64 water molecules of the

system. From these, the Born—Oppenheimer ground state force was subtracted, thus yielding only the centroid correction.

3. A NequIP machine learning model was trained on these 111 samples.

4. A CMD production simulation (which, at this point, is a purely classical simulation of 64 water molecules) was performed. The machine learning model obtained in the previous step was used to predict the centroid correction, which was added to the Born-Oppenheimer ground state force. For all considered systems, this simulation had a length of 80 ps and a timestep of 0.5 fs. A global CSVR thermostat was used, with a time constant of 1 ps. Velocities necessary for the computation of the VDOS were sampled every 2 fs.

The necessary simulation and equilibration lengths of the simulations in the first step of this procedure were conservatively chosen based on analysis of the root mean square deviation (the mean being performed over particles) of the running average of the force on the centroid from its final value, obtained by averaging over the entire trajectory (figure 18), for a sample configuration.

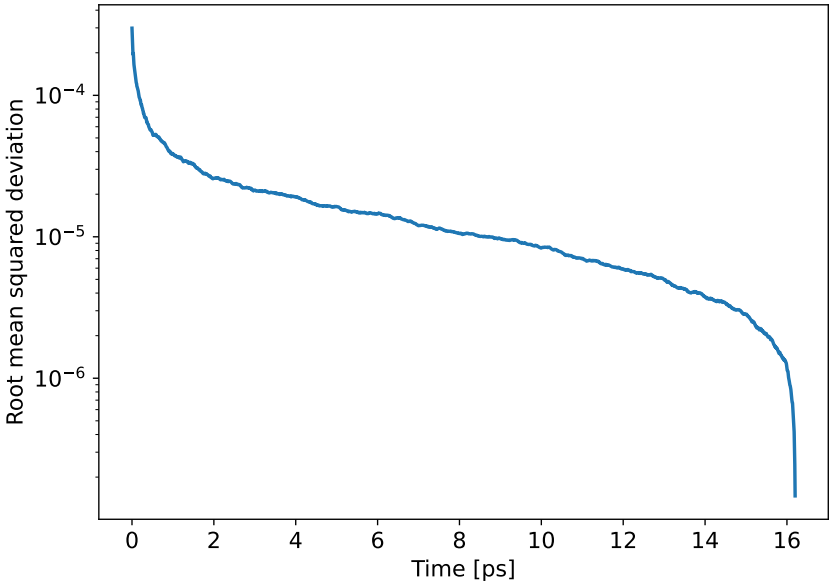

Figure 18: Root mean squared deviation of the running average of the centroid force from its final value.

To these CMD simulations, reference adiabatic CMD (henceforth referred to as “A-CMD”) and classical simulations were performed. The classical simulations were performed in exactly the same fashion as the production CMD

simulations, but the centroid correction was not used. The simulation settings of the adiabatic CMD simulations were based on the article 45. There, it was suggested that adiabatic separation can be achieved if the non-centroid degrees of freedom are shifted to a common target frequency as low as 8000 cm^{-1}. However, the time constants of the thermostats attached to these modes need to be simultaneously scaled down from the optimal values introduced in equation 3.14. From the supplementary information of the paper it follows that they should be scaled down by at least a factor of the order of magnitude 0.01. To verify the fact that the adiabatic separation is indeed fully reached, further benchmarks of the target frequency, thermostat scaling factor, and time step were performed (figure 21). The benchmark simulations had a length of 20 ps each and each differs in one parameter from the finally chosen A-CMD setup. An exception is the simulation with a higher target frequency, which had the time step lowered as well, which was necessary to integrate the high-frequency modes.

Based on these benchmarks, it was determined that an essentially full (or the largest practically possible) adiabatic separation is reached with the choice of parameters: target frequency 8000 cm^{-1}, thermostat scaling factor 0.001, and a time step of 0.05. These settings were used to perform reference adiabatic CMD simulations of the length 100 ps for the liquid water system at 300 K and 20 ps for the other two systems. In these simulations, a global CSVR thermostat with a time constant of 1 ps was also attached to the particle centroids. The velocities were again sampled every 2 fs.

The Hann windowing function [64] with the width parameter of 1000 fs was used to calculate the correlation functions according to subsection 3.5. The correlation functions were additionally padded with 2500 fs worth of zeroes to increase the numerical resolution of their Fourier transforms (which are the VDOSes).

The main results, compared with their classical counterparts, can be found in figure 19. All VDOSes displayed have a shape characteristic of liquid water infrared spectra — they consist of three dominant peaks, which correspond to (going from the lowest in frequency to the highest) the *libration* of the molecules, bending of the molecules, and vibration of the O-H bond. The term libration, commonly used in literature dealing with the structure of aqueous systems [63], refers to the hindered rotation of the molecules. The strongest source of this hindrance is caused by the hydrogen bonds formed by the molecule with the other molecules present in the system.

As can be seen from figure 19, the magnitude of the quantum correction rises with the frequency of the mode and the direction of the correction is in all cases towards lower frequencies. The magnitude of the correction is also smaller for the heavy water system and for the 350 K system. This is expected, as quantum behavior is expected to be suppressed if the temperature is higher, or the masses of the constituent particles are larger. Also, follow-

ing the discussion on nuclear quantum effects, notice on the first and third panel of figure 19 that the bulk of the difference between the liquid H_2O and D_2O VDOSes is already present at the classical level — the D_2O peaks being located at lower frequencies, due to the increased mass of the constituent atoms. The difference between quantum corrections of H_2O and D_2O contributes comparatively little and acts in the opposite direction. Overall, these results demonstrate that the proposed implementation is able to describe the quantum corrections to classical dynamics for at least a simple aqueous system while yielding behavior consistent with expectations and established knowledge.

The comparison of "interpolated" CMD and A-CMD can be found in figure 20. For all three systems, the two implementations of CMD yield almost identical results below 3000 cm^{-1}, however, for all three systems, the OH stretching peak (the highest frequency peak of the three most prominent peaks present) is slightly redshifted in the "interpolated" machine learning implementation of CMD with respect to A-CMD. Further differences between the two implementations can be seen in the logarithmic scale for frequencies above the OH stretching peak. Here the (weak intensity) *overtone* peaks can be seen to be smeared in the case of the adiabatic implementation, while for the machine learning implementation, these peaks are as sharp as in the classical case. These features are also shared across all three systems. The smearing is attributed to the imperfect adiabaticity of the adiabatic simulations, exacerbated by the smearing of the spectra of the non-centroid normal modes by the thermostats. In fact, a rather drastic example of this can be seen in the third panel of figure 20. The A-CMD VDOS can be seen to be essentially constant above 6000 cm^{-1}, in contrast to the VDOS obtained using machine-learning CMD.

Even though the overtone peaks are weak in intensity, they are not insignificant. Sensitive experimental methods exist which can study these features [65]. The fact that the proposed implementation of CMD is able to yield better-resolved peaks in this region, free from the influence of the non-centroid modes, is thus considered to be a major advantage.

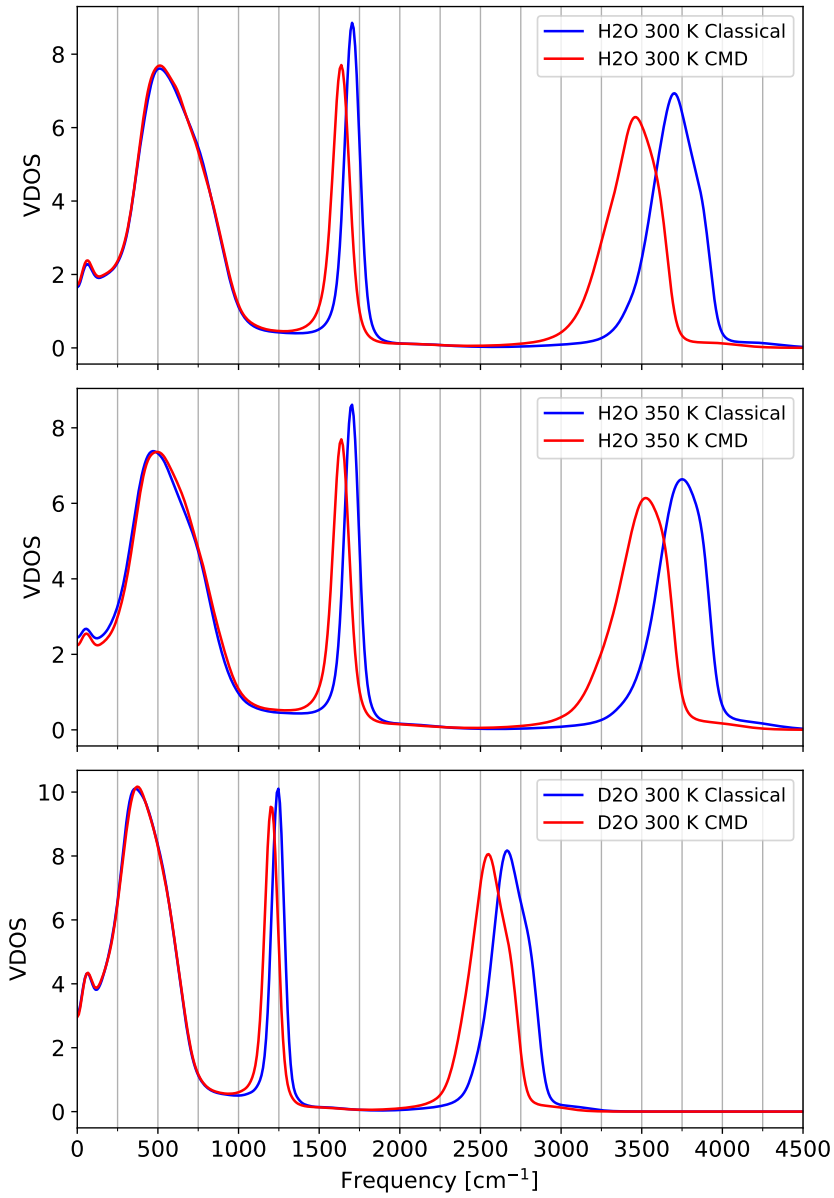

Figure 19: The H atom VDOS of liquid H_2O at 300 K, liquid H_2O at 350 K, and liquid D_2O respectively. The red and blue curves correspond to VDOSes calculated from simulations with or without the CMD correction, respectively.

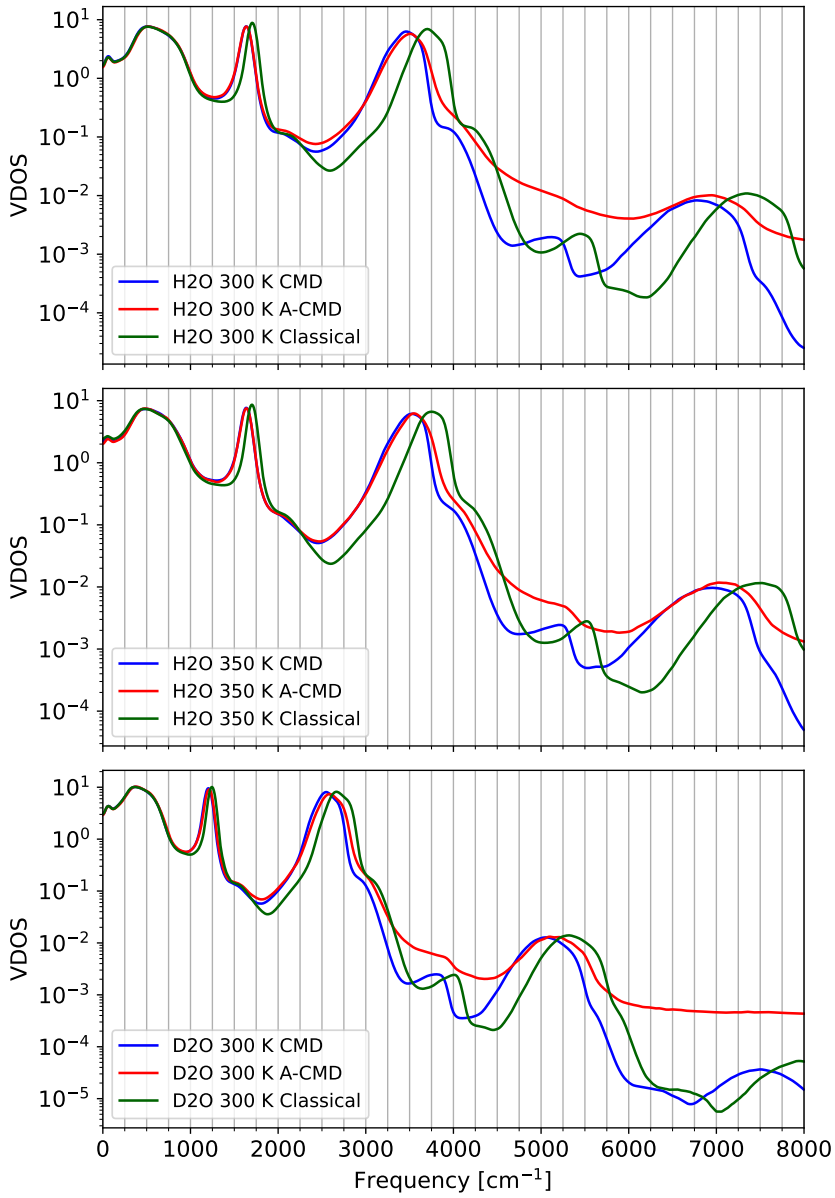

Figure 20: Comparison of the H atom VDOS for the H_2O 300 K system calculated from simulations performed with the CMD correction (blue), without the CMD correction (green), and using the A-CMD approach (red) for the three systems considered. The y-axis is in logarithmic scale.

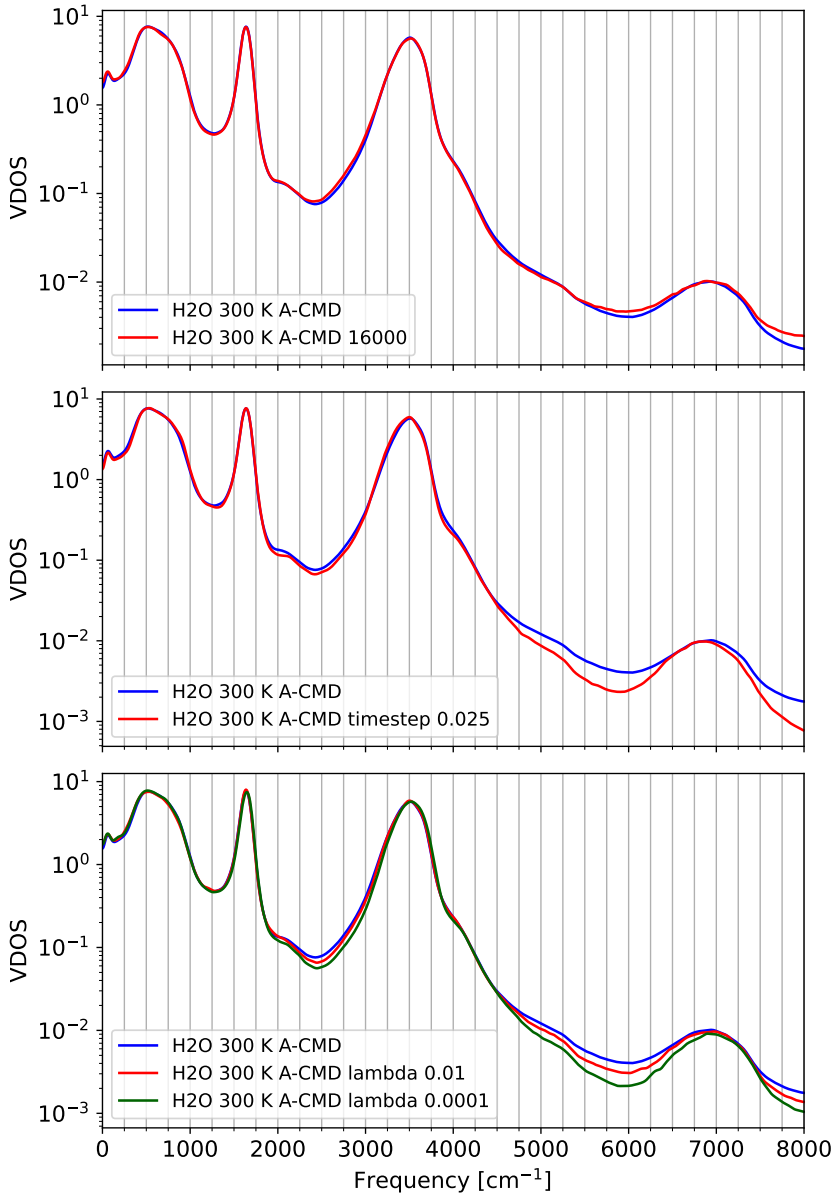

Figure 21: Benchmarks of the A-CMD setup on the case of H atom VDOS for the H_2O 300 K system. The blue curves correspond to the finally chosen setup. The first plot compares this setup to one with a higher target frequency (and shorter time step), the second to one with just a shorter time step, and the third to two different choices of the thermostat scaling factors (referred to as lambda). The y-axis is in logarithmic scale.

The proposed approach also allows further methods of investigation of the relative importance and behavior of the classical and quantum isotope effects. Since the potential for each CMD simulation is pre-calculated, a CMD simulation with one isotope of an element can in principle be run using a centroid potential calculated for a different one. In this work, this was done by running a CMD simulation of H_2O at 300 K using the centroid potential calculated for D_2O at 300 K (the rest of the parameters stayed the same). Running the simulation in such a fashion effectively isolates just the quantum contribution to the isotope effect present upon changing the hydrogen atoms in liquid water to deuterium atoms. The comparison of the VDOS obtained from such a CMD simulation to the VDOS obtained from the "ordinary" CMD simulation of liquid H_2O at 300 K can be seen in figure 22. As can be seen, the quantum isotope effect is quite small and causes the peaks in the VDOS to shift to higher frequencies. In contrast, as can be seen from figure 19, the classical isotope effect causes the peak to shift to lower frequencies.

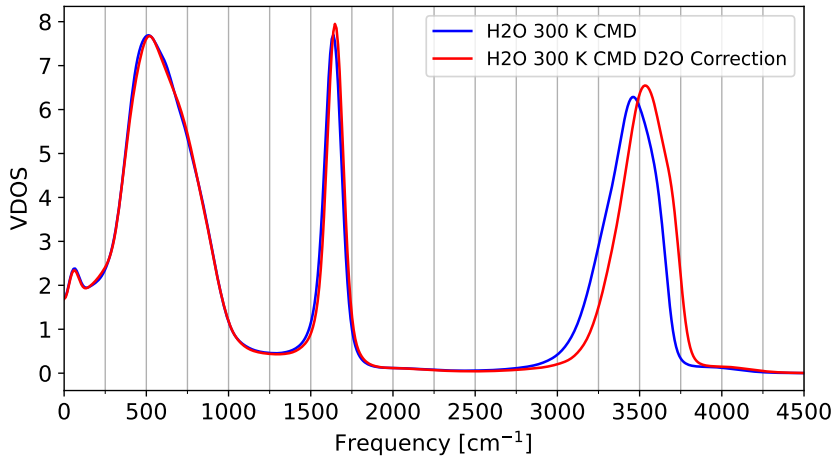

Figure 22: Comparison of the H atom VDOS for the H_2O 300 K system calculated from simulations with the H_2O 300 K centroid correction (blue) or the D_2O 300 K centroid correction (red).

The last point which will be discussed when it comes to the VDOSes is the recently proposed "T-PIGS" approach (already mentioned in the discussion of the 2D Morse potential), which was used as a simple model of the O-H bond. As mentioned previously, this approach has been suggested to alleviate the curvature problem for CMD simulations of liquid water, and it rests in using the CMD potential calculated at a higher temperature for a simulation at a lower temperature. For the 2D Morse potential, this approach has been shown to alleviate the spurious shift of the vibrational peaks to lower frequencies, while leaving the rotational peak largely untouched.

This approach has also been tested, in this work, for the liquid water system at 300 K — a CMD simulation has been performed for this system with a centroid potential calculated at an elevated temperature of 600 K. All other details of the procedure were exactly the same as in the previous cases. As can be seen from figure 23, this approach has indeed resulted in a VDOS in which the O-H stretching peak and the bending peak are shifted to higher frequencies compared to normal CMD. The librational peak remains largely unaffected, although some very subtle differences can be seen.

Combining this result with the previous result for the 2D Morse potential might seem like a great success for this approach. However, care needs to be taken when interpreting this result. Elevating the temperature, in contrast to the simple model potential, causes the effective potential felt by any given atom to change, as the entire static equilibrium distribution in the system changes. The shift of the two peaks thus contains artificial and uncontrolled contributions from the change of the environment on top of the "ground state correction" present for the 2D Morse potential. As a particular example of this, elevated temperature weakens hydrogen bonding in liquid water, which is known to cause the O-H stretching peak to shift to higher frequencies as well. This effect is present at both the classical and quantum levels of description.

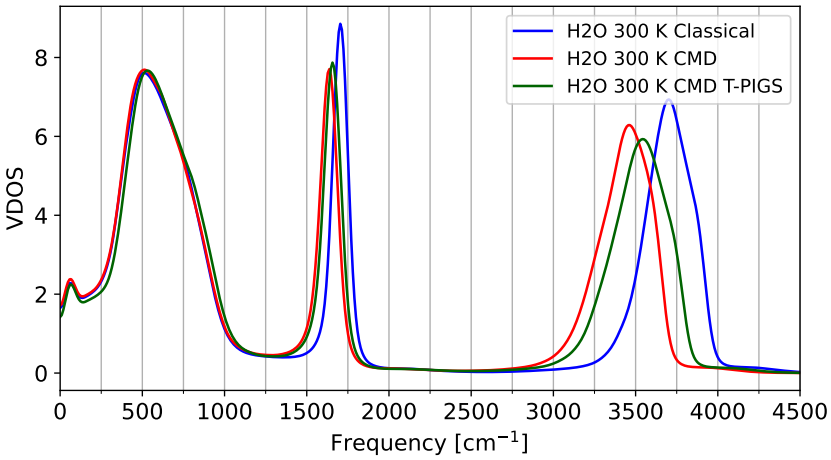

Figure 23: Comparison of the H atom VDOS for the H_2O 300 K system calculated from simulations with the 300 K centroid correction (blue), 600 K centroid correction (green), or no centroid correction (red).

To gain further insight into the structure of the CMD potential, the CMD correction to the force on the hydrogen atom has been calculated for 8000 representative thermal structures. These were obtained by sampling every 10 fs from the CMD simulations used to obtain the VDOSes. This force has then been split into a projection along the O-H bond and a component orthogonal

to this direction. The distributions of the projections and magnitudes of the orthogonal components alongside the same distributions for the raw Born–Oppenheimer (B–O) and total forces (calculated on the same structures) can be found in figures 24 and 25.

The dependence of these two quantities on the O-H bond length has also been studied, as both were expected to be sensitive to this parameter. The joint distributions of the O-H bond length and the two quantities of interest can be found in figures 26 and 27.

Several observations can be made. The distribution of the projection of the total force along the O-H bond peaks around 0. This is an expected result and it simply reflects the fact that the O-H covalent bonding largely defines the potential in this direction (this being the case at both classical and quantum levels). It is thus characterized by a well-defined minimum around which the particle is most likely to be found. This is also reflected in the fact that the projections of the B–O forces and the CMD corrections are roughly the same in magnitude and opposite in sign. Even though adding the CMD correction to a simulation has a relatively small impact on the calculated VDOS (compared to the frequencies at which the bending and vibrational peaks are located), this does not directly imply that the CMD correction is always smaller than the B–O force. In the orthogonal direction, this picture is somewhat restored — the distribution of the magnitude of the total force is not peaked around zero but around a finite value which in turn is mostly determined by the B–O force, while the CMD correction plays a smaller role.

As far as the dependence of the quantities on the O-H bond length is concerned — all magnitudes of the orthogonal components of the forces on the H atoms can be seen to be virtually independent of the bond length (figure 27). The projection of the total force forms a narrow distribution that follows a well-defined curve as a function of the bond length (figure 26). In contrast to this, the distributions of the CMD correction and the B–O force are again rather independent of the bond length. The fact that the distributions of the total forces are as narrow as they are is something that directly follows from the discussion in the previous paragraph. The particle in this direction is located in a sharp potential minimum around which it oscillates. As such, the force has to systematically rise as the bond length gets shorter and decrease as it gets longer according to the confining potential. The fact that the distribution is narrow also implies that the behavior of the O-H bond is relatively insensitive to the surrounding environment. The B–O force and the CMD correction then have to have distributions that cancel each other in a manner that yields this distribution. The same discussion applies to all three systems studied.

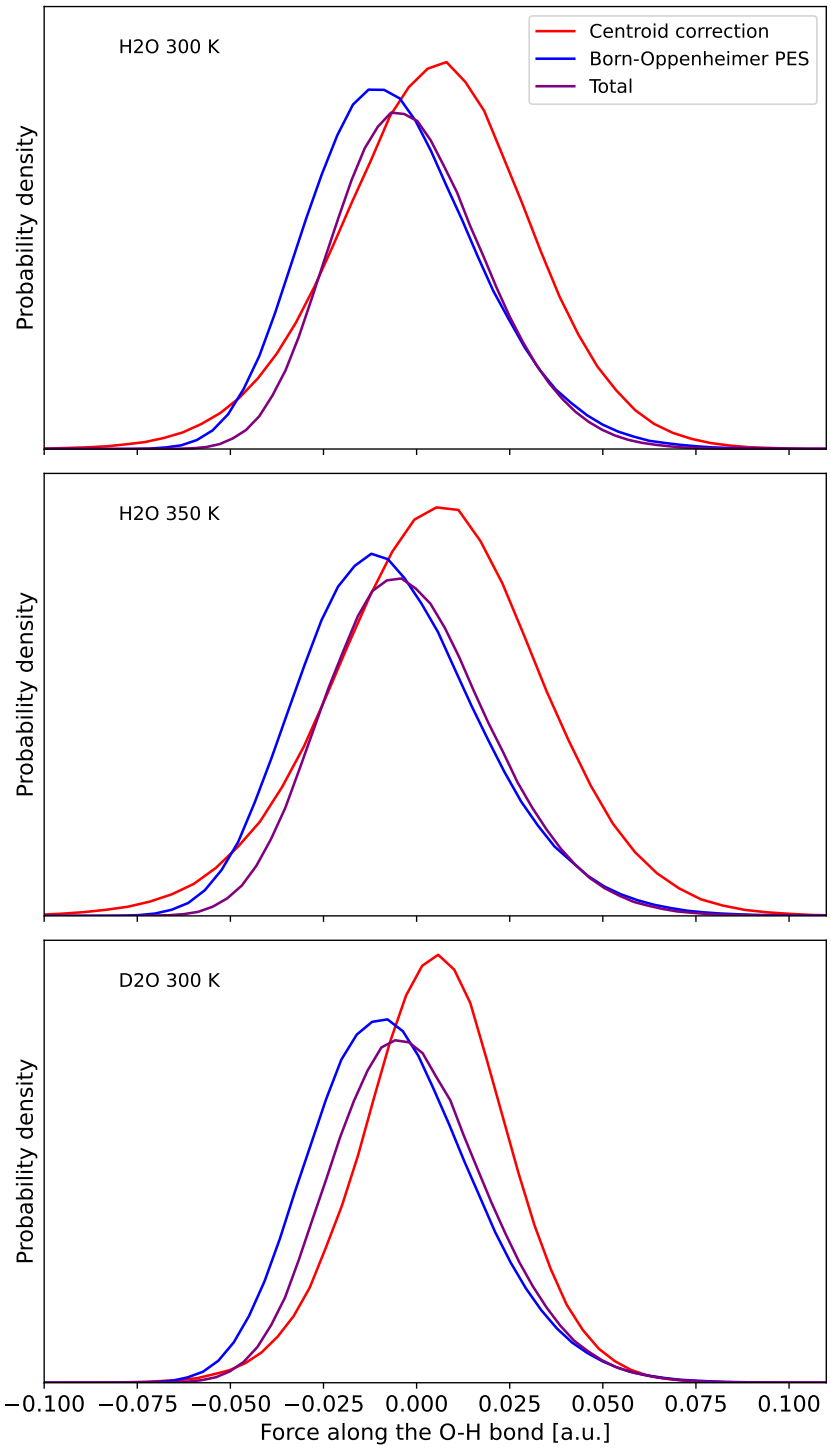

Figure 24: Distributions of the projections of the total, CMD correction and B–O forces on the H atom on the O-H bond for the three systems considered.

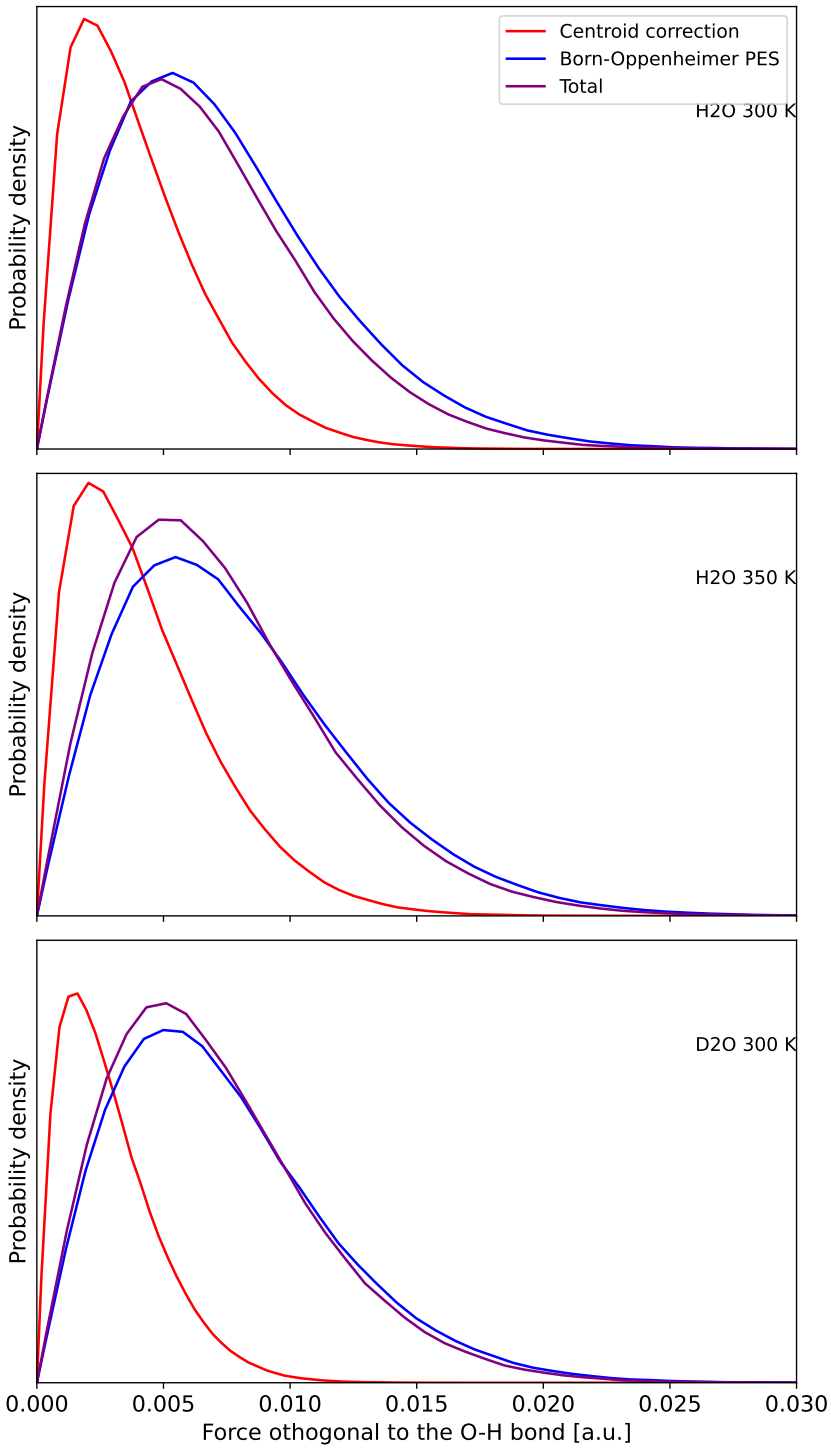

Figure 25: Distributions of the magnitudes of the components of the total, CMD correction, and B–O forces on the H atoms orthogonal to the O-H bond for the three systems considered.

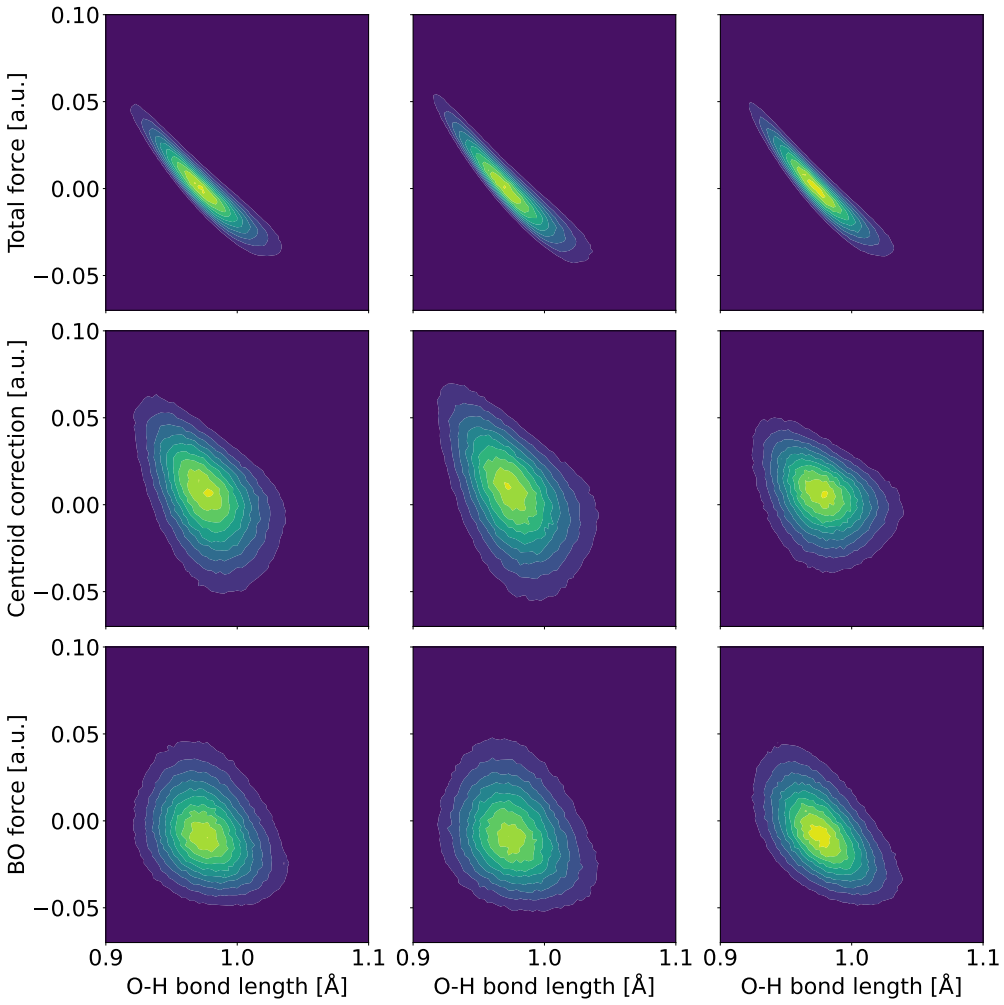

Figure 26: Joint distributions of the projection of the forces on the H atoms on the O-H bond, and the O-H bond length for the three different systems studied. The left-most column corresponds to the H_2O 300 K system, the middle one to the H_2O 350 k system, and the right-most one to the D_2O 300 K system.

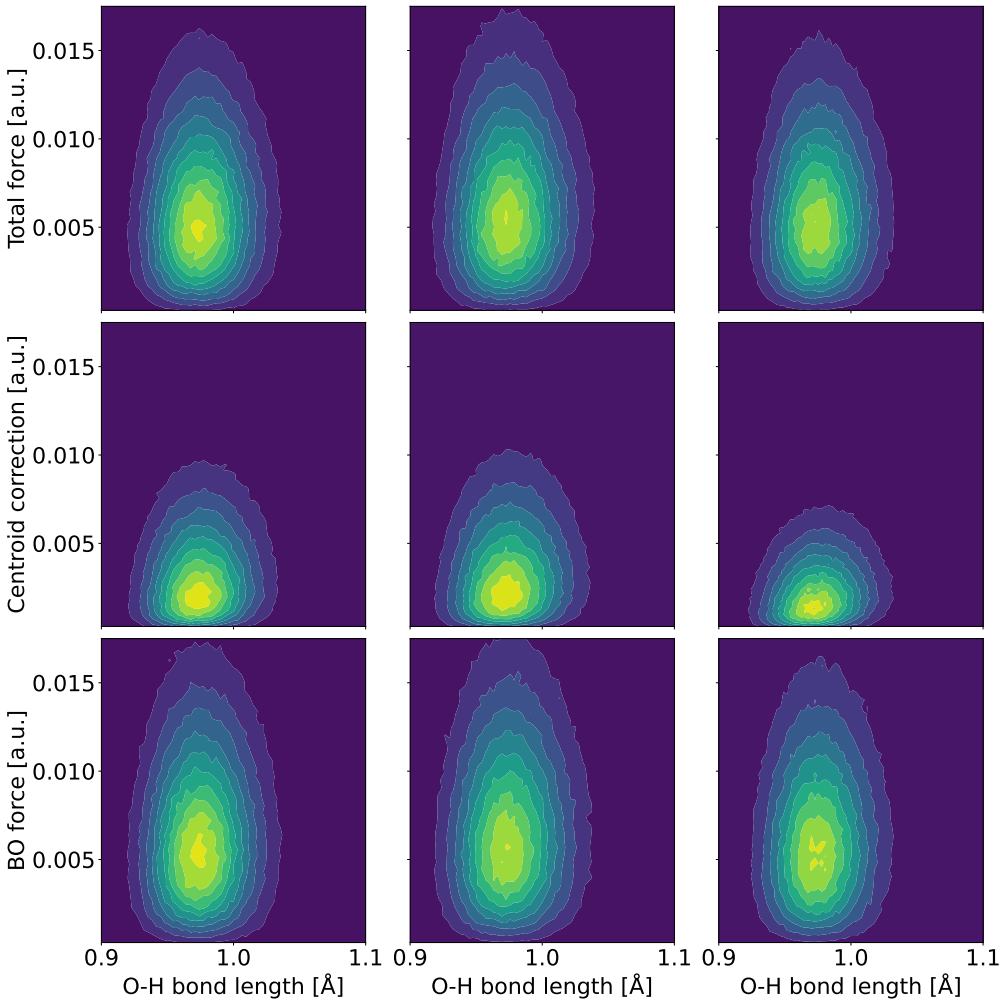

Figure 27: Joint distributions of the magnitude of the components of the forces on the H atoms orthogonal on the O-H bond, and the O-H bond length for the three different systems studied. The left-most column corresponds to the H_2O 300 K system, the middle one to the H_2O 350 k system, and the right-most one to the D_2O 300 K system.

Another property of the centroid correction which was investigated was its "locality". As was stated previously in subsection 3.3, the machine learning tools used for the construction of the centroid correction "potential" generally perform better for potentials that are more local. In this context, locality means that the potential (or force) on a given particle significantly depends only on the positions of other particles in the immediate surrounding of the particle. The sensitivity of the force on a given particle i on the position of some other particle j is contained in the matrix element of the Hessian matrix $\frac{\partial^2 U(x)}{\partial \vec{x}_i \partial \vec{x}_j}$. To evaluate the locality of the centroid correction, a plot of the

dependence of the average absolute value (the averaging being done over all particles and spatial directions) of these matrix elements on the inter-particle distance $\left|\vec{x}_i - \vec{x}_j\right|$ has been constructed. This analysis has been done on 30 structures from the CMD simulation of liquid H_2O at 300 K, each separated by 100 fs. An identical plot was also done for the B–O potential for comparison. The results can be seen on figure 28. As can be seen, the average absolute value of the Hessian matrix element is about an order of magnitude lower for the CMD correction than for the B–O potential. The dependence of this quantity on the interparticle distance is pretty much the same for both "potentials" up to a distance of about 3 Å. After this point the centroid correction decays faster than the B–O potential. The centroid correction thus is in some sense more local than the B–O potential, however the difference is not as stark as was thought. However, it can be concluded that the centroid correction should not be expected to be harder to describe using the machine learning models than the B–O potential.

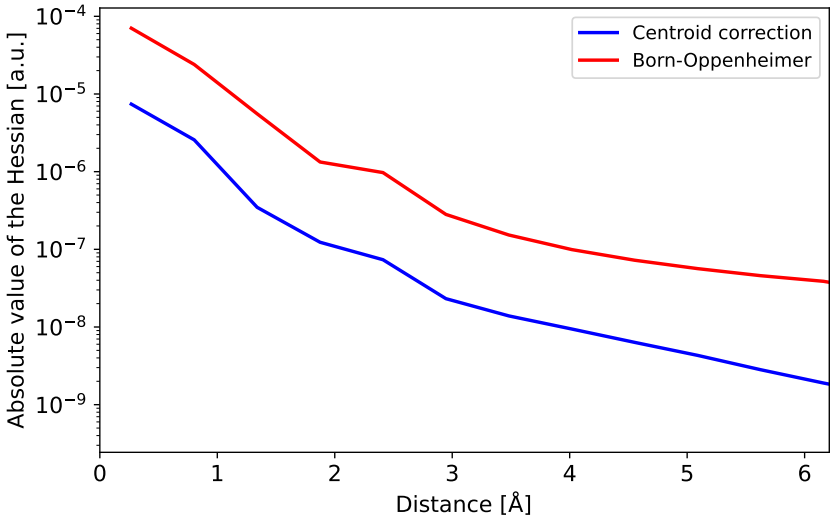

Figure 28: The dependence of the average absolute value of the elements of the Hessian matrix on the interparticle distance. The two curves correspond to the Hessian matrix calculated from the centroid correction "potential", calculated for the liquid H_2O system at 300 K, and the Born–Oppenheimer potential. The y-axis is in logarithmic scale.

To conclude this section, the computational cost will be briefly discussed. As was mentioned, the time step which was found to be necessary for the A-CMD simulations was 0.05 fs. This timestep is 10 times shorter than the time step of 0.5 fs which is commonly used for classical simulations of aqueous systems, which could thus also be safely used for the CMD simulations performed according to the newly proposed "interpolated" scheme. Further-

more, since the A-CMD simulations are, in the end, path-integral molecular dynamics simulations, the force needs to be evaluated P times per step in these simulations (in this work $P = 32$), whereas for the interpolated implementation just one evaluation of the B–O force and one evaluation of the centroid correction are needed. Thus, not only do A-CMD simulations need more simulation steps to reach equal sampling, but they can also be expected to be more expensive per simulation step. Of course, a large chunk of the computational expenses associated with the "interpolated" scheme is "hidden" in the preparation of the training data. This preparation requires running a number (in this work 111) of short path-integral molecular dynamics simulations. These also require P force evaluations per step, but since the purpose of these simulations is just to sample static equilibrium properties (with the centroid fixed at a certain position), there is far more freedom in the choice of the simulation parameters than in the A-CMD simulations. The only two criteria on the choice of the simulation parameters are, that the numerical integration is of reasonable accuracy and that enough relevant samples are obtained. For instance, in this work, the simulations to obtain the training data had a time step of 0.2 fs and a length of 10 ps, coming in at 50 000 steps of path-integral molecular dynamics simulations per point of training data. To obtain an 80 ps simulation, the total amount of path-integral molecular dynamics steps used was thus still higher $50000 \times 111 = 5550000$ in the machine-learning implementation than in the adiabatic one $80000/0.05 = 1600000$. However, for the proposed implementation, this number is fixed, and thus as the required simulation length would increase, this comparison would change in its favor. At the same time, the simulations to obtain the training data can be "embarrassingly" parallelized — all the training simulations could be run simultaneously, if the hardware is available, while the single A-CMD simulation needs to be run strictly in sequence. Furthermore, the adiabatic implementation requires further benchmarking and convergence tests to even arrive at the starting parameters for the simulations, while the "interpolated" implementation mostly does not require these. The only thing really requiring such tests being the set of configurations considered in the training set. However, here, an advantage may be taken of methods already described in the literature, which were developed for the efficient construction of training sets for the construction of machine learning models of the B–O potential energy surfaces for molecular systems.

5 Conclusions

In this work, new methodology for performing CMD simulations with the help of machine learning has been proposed and tested. This new approach was shown to yield results (position autocorrelation functions and VDOSes) that are largely consistent with the previous adiabatic realization of CMD for both low-dimensional model systems and high-dimensional molecular systems while being computationally cheaper, and easier to run. Despite this, some slight differences were still found — most noticeably the position of the O-H stretching peak or the resolution of the overtone peaks in the H-atom VDOS of liquid water. Even though the nature of the former remains unknown, in the latter case the results obtained from the newly proposed approach were deemed to be the more physical and accurate ones. The potential usage of the newly proposed methodology in the solution of the curvature problem, present in CMD simulation were also briefly discussed. A further advantage of the newly proposed approach is the fact that the centroid potential and forces, used in CMD, are readily available. Thanks to this, their properties were directly studied as well.

In the future, the proposed methodology can immediately be used to calculate other dynamical properties of the studied systems — such as reorientation time correlation functions or diffusion coefficients. The proposed approach is especially well suited for properties that require simulations with a large amount of sampling — typically diffusion coefficients or decomposed time correlation functions or vibrational spectra. The methodology can also be used to run efficient CMD simulations for different, possibly quite complicated systems. In particular, using the new approach, CMD simulations could be used to answer questions about various systems that were previously unobtainable using the adiabatic implementation due to the high costs and technical difficulties.

A further interesting application of the current methodology could be called a “return to the original idea of Feynman and Kleinert” [13]. The original idea of Feynman and Kleinert was to use the concept of the centroid and centroid potential to simplify the calculations of quantum equilibrium properties by effectively reducing the problem of quantum statistics to a problem of classical statistics of the same dimensionality. Using the currently proposed approach, not only could the idea be expanded for realistic, and practically useful, highly-dimensional systems, but also some of the original approximations made by Feynman and Kleinert, which were introduced to make the problem tractable for computers in their time, could be eliminated.

References

[1] Mark Tuckerman. *Statistical mechanics: theory and molecular simulation*. Oxford university press, 2010.

[2] Dominik Marx and Jurg Hutter. "Ab initio molecular dynamics: Theory and implementation". In: *Modern methods and algorithms of quantum chemistry* 1.301-449 (2000), p. 141.

[3] Richard P Feynman, Albert R Hibbs, and Daniel F Styer. *Quantum mechanics and path integrals*. Courier Corporation, 2010.

[4] Thomas E Markland and Michele Ceriotti. "Nuclear quantum effects enter the mainstream". In: *Nature Reviews Chemistry* 2.3 (2018), p. 0109.

[5] Matthias Troyer and Uwe-Jens Wiese. "Computational complexity and fundamental limitations to fermionic quantum Monte Carlo simulations". In: *Physical review letters* 94.17 (2005), p. 170201.

[6] Michael F Herman and Edward Kluk. "A semiclasical justification for the use of non-spreading wavepackets in dynamics calculations". In: *Chemical Physics* 91.1 (1984), pp. 27–34.

[7] William H Miller. "Quantum dynamics of complex molecular systems". In: *Proceedings of the National Academy of Sciences* 102.19 (2005), pp. 6660–6664.

[8] Ian R Craig and David E Manolopoulos. "Quantum statistics and classical mechanics: Real time correlation functions from ring polymer molecular dynamics". In: *The Journal of chemical physics* 121.8 (2004), pp. 3368–3373.

[9] Mariana Rossi, Michele Ceriotti, and David E Manolopoulos. "How to remove the spurious resonances from ring polymer molecular dynamics". In: *The Journal of chemical physics* 140.23 (2014), p. 234116.

[10] Jianshu Cao and Gregory A Voth. "The formulation of quantum statistical mechanics based on the Feynman path centroid density. II. Dynamical properties". In: *The Journal of chemical physics* 100.7 (1994), pp. 5106–5117.

[11] Timothy JH Hele et al. "Communication: Relation of centroid molecular dynamics and ring-polymer molecular dynamics to exact quantum dynamics". In: *The Journal of Chemical Physics* 142.19 (2015), p. 191101.

[12] Timothy JH Hele et al. "Boltzmann-conserving classical dynamics in quantum time-correlation functions:"Matsubara dynamics"". In: *The Journal of Chemical Physics* 142.13 (2015), p. 134103.

[13] RP Feynman and H Kleinert. "Effective classical partition functions". In: *Physical Review A* 34.6 (1986), p. 5080.

[14] Jianshu Cao and Gregory A Voth. “The formulation of quantum statistical mechanics based on the Feynman path centroid density. IV. Algorithms for centroid molecular dynamics”. In: *The Journal of chemical physics* 101.7 (1994), pp. 6168–6183.

[15] Tyler D Hone, Peter J Rossky, and Gregory A Voth. “A comparative study of imaginary time path integral based methods for quantum dynamics”. In: *The Journal of chemical physics* 124.15 (2006), p. 154103.

[16] Jörg Behler. “Four generations of high-dimensional neural network potentials”. In: *Chemical Reviews* 121.16 (2021), pp. 10037–10072.

[17] Max Born, Kun Huang, and M Lax. “Dynamical theory of crystal lattices”. In: *American Journal of Physics* 23.7 (1955), pp. 474–474.

[18] Max Born and Robert Oppenheimer. “On the quantum theory of molecules”. In: *Quantum Chemistry: Classic Scientific Papers*. World Scientific, 2000, pp. 1–24.

[19] Richard Phillips Feynman. “Space-time approach to non-relativistic quantum mechanics”. In: *Reviews of modern physics* 20.2 (1948), p. 367.

[20] Paul Adrien M Dirac. “The Lagrangian in quantum mechanics”. In: *Physikalische Zeitschirift der Sowjetunion* 3 (1933), pp. 312–320.

[21] Hale F Trotter. “On the product of semi-groups of operators”. In: *Proceedings of the American Mathematical Society* 10.4 (1959), pp. 545–551.

[22] James Glimm and Arthur Jaffe. *Quantum physics: a functional integral point of view*. Springer Science & Business Media, 2012.

[23] Ryogo Kubo, Morikazu Toda, and Natsuki Hashitsume. *Statistical physics II: nonequilibrium statistical mechanics*. Vol. 31. Springer Science & Business Media, 2012.

[24] Ryogo Kubo. “Statistical-mechanical theory of irreversible processes. I. General theory and simple applications to magnetic and conduction problems”. In: *Journal of the Physical Society of Japan* 12.6 (1957), pp. 570–586.

[25] Paul Langevin. “Sur la théorie du mouvement brownien”. In: *Compt. Rendus* 146 (1908), pp. 530–533.

[26] Gerald Mathias and Marcel D Baer. “Generalized normal coordinates for the vibrational analysis of molecular dynamics simulations”. In: *Journal of Chemical Theory and Computation* 7.7 (2011), pp. 2028–2039.

[27] Nicholas Metropolis et al. “Equation of state calculations by fast computing machines”. In: *The journal of chemical physics* 21.6 (1953), pp. 1087–1092.

[28] William C Swope et al. “A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters”. In: *The Journal of chemical physics* 76.1 (1982), pp. 637–649.

[29] MBBJM Tuckerman, Bruce J Berne, and Glenn J Martyna. “Reversible multiple time scale molecular dynamics”. In: *The Journal of chemical physics* 97.3 (1992), pp. 1990–2001.

[30] Michele Ceriotti et al. “Efficient stochastic thermostatting of path integral molecular dynamics”. In: *The Journal of chemical physics* 133.12 (2010), p. 124104.

[31] Benedict Leimkuhler and Charles Matthews. “Rational construction of stochastic numerical methods for molecular sampling”. In: *Applied Mathematics Research eXpress* 2013.1 (2013), pp. 34–56.

[32] Giovanni Bussi and Michele Parrinello. “Stochastic thermostats: comparison of local and global schemes”. In: *Computer Physics Communications* 179.1-3 (2008), pp. 26–29.

[33] David Chandler and Peter G Wolynes. “Exploiting the isomorphism between quantum theory and classical statistical mechanics of polyatomic fluids”. In: *The Journal of Chemical Physics* 74.7 (1981), pp. 4078–4095.

[34] MF Herman, EJ Bruskin, and BJ Berne. “On path integral Monte Carlo simulations”. In: *The Journal of Chemical Physics* 76.10 (1982), pp. 5150–5155.

[35] Jianshu Cao and Gregory A Voth. “A new perspective on quantum time correlation functions”. In: *The Journal of chemical physics* 99.12 (1993), pp. 10070–10073.

[36] GJ Martyna. “Adiabatic path integral molecular dynamics methods. I. Theory”. In: *The Journal of chemical physics* 104.5 (1996), pp. 2018–2027.

[37] Bastiaan J Braams and David E Manolopoulos. “On the short-time limit of ring polymer molecular dynamics”. In: *The Journal of chemical physics* 125.12 (2006), p. 124105.

[38] Randall W Hall and Bruce J Berne. “Nonergodicity in path integral molecular dynamics”. In: *The Journal of chemical physics* 81.8 (1984), pp. 3641–3643.

[39] Mark E Tuckerman et al. “Efficient molecular dynamics and hybrid Monte Carlo algorithms for path integrals”. In: *The Journal of Chemical Physics* 99.4 (1993), pp. 2796–2808.

[40] Venkat Kapil et al. “i-PI 2.0: A universal force engine for advanced molecular simulations”. In: *Computer Physics Communications* 236 (2019), pp. 214–223.

[41] Thomas D Kühne et al. “CP2K: An electronic structure and molecular dynamics software package-Quickstep: Efficient and accurate electronic structure calculations”. In: *The Journal of Chemical Physics* 152.19 (2020), p. 194103.

[42] Christoph Schran, Krystof Brezina, and Ondrej Marsalek. “Committee neural network potentials control generalization errors and enable active learning”. In: *The Journal of Chemical Physics* 153.10 (2020), p. 104105.

[43] Matteo Frigo and Steven G Johnson. “FFTW: An adaptive software architecture for the FFT”. In: *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*. Vol. 3. IEEE. 1998, pp. 1381–1384.

[44] J Cao and GJ Martyna. “Adiabatic path integral molecular dynamics methods. II. Algorithms”. In: *The Journal of chemical physics* 104.5 (1996), pp. 2028–2035.

[45] Mariana Rossi et al. “Communication: On the consistency of approximate quantum dynamics simulation methods for vibrational spectra in the condensed phase”. In: *The Journal of Chemical Physics* 141.18 (2014), p. 181101.

[46] Scott Habershon, Thomas E Markland, and David E Manolopoulos. “Competing quantum effects in the dynamics of a flexible water model”. In: *The journal of chemical physics* 131.2 (2009), p. 024501.

[47] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. “Multilayer feedforward networks are universal approximators”. In: *Neural networks* 2.5 (1989), pp. 359–366.

[48] Jörg Behler and Michele Parrinello. “Generalized neural-network representation of high-dimensional potential-energy surfaces”. In: *Physical review letters* 98.14 (2007), p. 146401.

[49] Jörg Behler. “Atom-centered symmetry functions for constructing high-dimensional neural network potentials”. In: *The Journal of chemical physics* 134.7 (2011), p. 074106.

[50] Simon Batzner et al. “E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials”. In: *Nature communications* 13.1 (2022), p. 2453.

[51] A. P. Thompson et al. “LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales”. In: *Comp. Phys. Comm.* 271 (2022), p. 108171. DOI: 10.1016/j.cpc.2021.108171.

[52] Carl De Boor and Carl De Boor. *A practical guide to splines*. Vol. 27. springer-verlag New York, 1978.

[53] Pauli Virtanen et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[54] Claude E Shannon. “Communication in the presence of noise”. In: *Proceedings of the IRE* 37.1 (1949), pp. 10–21.

[55] Charles R. Harris et al. “Array programming with NumPy”. In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2.

[56] Jiří Eliášek and Ondřej Maršálek. *Fourier-DVR-1D*. https://github.com/OndrejMarsalek/Fourier-DVR-1D.

[57] George Trenins and Stuart C Althorpe. “Mean-field Matsubara dynamics: Analysis of path-integral curvature effects in rovibrational spectra”. In: *The Journal of Chemical Physics* 149.1 (2018).

[58] Alejandro Pérez, Mark E Tuckerman, and Martin H Müser. “A comparative study of the centroid and ring-polymer molecular dynamics methods for approximating quantum time correlation functions from path integrals”. In: *The Journal of chemical physics* 130.18 (2009).

[59] Jeremy O Richardson and Stuart C Althorpe. “Ring-polymer instanton method for calculating tunneling splittings”. In: *The Journal of chemical physics* 134.5 (2011).

[60] Sergei D Ivanov et al. “Communications: On artificial frequency shifts in infrared spectra obtained from centroid molecular dynamics: Quantum liquid water”. In: *The Journal of chemical physics* 132.3 (2010).

[61] George Trenins, Michael J Willatt, and Stuart C Althorpe. “Path-integral dynamics of water using curvilinear centroids”. In: *The Journal of Chemical Physics* 151.5 (2019).

[62] Félix Musil et al. “Quantum dynamics using path integral coarse-graining”. In: *The Journal of Chemical Physics* 157.18 (2022).

[63] Michele Ceriotti et al. “Nuclear quantum effects in water and aqueous systems: Experiment, theory, and current challenges”. In: *Chemical reviews* 116.13 (2016), pp. 7529–7550.

[64] RB Blackman and JW Tukey. “The measurement of power spectra Dover Publications”. In: *Inc, New York* (1958).

[65] Jun Ye, Long-Sheng Ma, and John L Hall. “Ultrasensitive detections in atomic and molecular physics: demonstration in molecular overtone spectroscopy”. In: *JOSA B* 15.1 (1998), pp. 6–15.

[66] Attila Szabo and Neil S Ostlund. *Modern quantum chemistry: introduction to advanced electronic structure theory*. Courier Corporation, 2012.

[67] Robert G. Parr and Weitao Yang. “Density-Functional Theory of Atoms and Molecules”. In: Oxford University Press, 1989.

[68] Douglas R Hartree. “The wave mechanics of an atom with a non-Coulomb central field. Part I. Theory and methods”. In: *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 24. 1. Cambridge university press. 1928, pp. 89–110.

[69] Vladimir Fock. “Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems”. In: *Zeitschrift für Physik* 61 (1930), pp. 126–148.

[70] Chr Møller and Milton S Plesset. “Note on an approximation treatment for many-electron systems”. In: *Physical review* 46.7 (1934), p. 618.

[71] John C Slater. “The theory of complex spectra”. In: *Physical Review* 34.10 (1929), p. 1293.

[72] EU Condon. “The theory of complex spectra”. In: *Physical Review* 36.7 (1930), p. 1121.

[73] Jiří Čížek. “On the correlation problem in atomic and molecular systems. Calculation of wavefunction components in Ursell-type expansion using quantum-field theoretical methods”. In: *The Journal of Chemical Physics* 45.11 (1966), pp. 4256–4266.

[74] Pierre Hohenberg and Walter Kohn. “Inhomogeneous electron gas”. In: *Physical review* 136.3B (1964), B864.

[75] Walter Kohn and Lu Jeu Sham. “Self-consistent equations including exchange and correlation effects”. In: *Physical review* 140.4A (1965), A1133.

[76] David Bohm. *Quantum theory*. Courier Corporation, 2012.

[77] L. D. Landau and E. M. Lifshitz. *Mechanics, Third Edition: Volume 1 (Course of Theoretical Physics)*. Butterworth-Heinemann, 1976.

Appendix A Electronic structure problem

This appendix will briefly discuss the main ideas and methods used to find approximate solutions to the *electronic structure problem*. This appendix will not go into much detail, since a plethora of literature exists treating this topic in great detail (see for instance ref. 66 or 67). The problem rests in finding the ground state wave function and energy of a system of N electrons and M fixed nuclei (entering the problem only as external point charges), all interacting via the Coulomb potential. The Hamiltonian for this problem is in atomic units:

$$\hat{H}_{\text{elec}} = -\sum_{i=1}^{N} \frac{1}{2}\nabla_{\vec{x}_i}^2 + \sum_{i=1}^{N}\sum_{j>i}^{N} \frac{1}{|\vec{x}_i - \vec{x}_j|} + \sum_{i=1}^{M}\sum_{j>i}^{M} \frac{Z_i Z_j}{|\vec{X}_i - \vec{X}_j|} - \sum_{i=1}^{N}\sum_{j=1}^{M} \frac{1}{|\vec{x}_i - \vec{X}_j|} \tag{A.1}$$

The most common starting point for finding the ground state is to employ a mean-field ansatz (of proper exchange symmetry) for the wave function and the Ritz variational principle to search for the lowest energy among the wave functions in this subspace. For fermions, this method is referred to as the *Hartree–Fock method* [68, 69]. As foreshadowed previously, this approximation uses the following ansatz for the many-electron wave function:

$$\psi(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N) = \hat{\mathscr{A}}(\phi_1(\vec{x}_1)\phi_2(\vec{x}_2)\ldots\phi_N(\vec{x}_N)), \tag{A.2}$$

where $\hat{\mathscr{A}}$ is the antisymmetrization operator. It is thus assumed that the total wave function can be written in an antisymmetrized product form. This corresponds to the assumption, that there is no correlation between the electrons, apart from the correlation caused by the antisymmetrization. Within the Ritz variational principle the individual one-electron wavefunctions ϕ_i (also referred to as "orbitals") are then subject to variation, under the constraint that they remain orthonormal:

$$\delta\left[E[\phi_i] - \epsilon_{ij}\left(\langle\phi_i|\phi_j\rangle - \delta_{ij}\right)\right] = 0, \tag{A.3}$$

ϵ_{ij} being Lagrange multipliers. Skipping the derivation, this variation leads to the *Hartree–Fock equations*:

$$\begin{aligned}\hat{h}(\vec{x}_1)|\phi_i(\vec{x}_1)\rangle + \sum_{j=1}^{N} \langle\phi_j(\vec{x}_2)|\,\hat{v}(\vec{x}_1,\vec{x}_2)\left(|\phi_i(\vec{x}_1)\rangle|\phi_j(\vec{x}_2)\rangle - |\phi_j(\vec{x}_1)\rangle|\phi_i(\vec{x}_2)\rangle\right) \\ = \epsilon_{ii}|\phi_i(\vec{x}_i)\rangle,\end{aligned} \tag{A.4}$$

where

$$\hat{h}(\vec{x}_1) = -\frac{1}{2}\nabla_{\vec{x}_1}^2 + \sum_{k=1}^{M} \frac{1}{|\vec{x}_1 - \vec{X}_k|} \tag{A.5}$$

and

$$\hat{v}(\vec{x}_1, \vec{x}_2) = \frac{1}{|\vec{x}_1 - \vec{x}_2|}. \tag{A.6}$$

These equations for the functions ϕ_i are then usually solved in a basis. Note that the second term on the left-hand side of these equations contains the average electrostatic potential generated by all the other electrons (the "mean field"), as well as a non-local *exchange interaction*. The left-hand side thus depends on the functions ϕ_i for which the equations are solved, and as such needs to be solved self-consistently. After the equations are solved and the functions ϕ_i obtained, the ground state energy (from the point of view of this work, the main quantity of interest) can be obtained as

$$\begin{aligned} E_0 = E[\phi_{0i}] &= \sum_{i=1}^{N} \epsilon_{0ii} \\ &- \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left[\left\langle \phi_{0j}(\vec{x}_2) \right| \hat{v}(\vec{x}_1, \vec{x}_2) \left(\left| \phi_{0i}(\vec{x}_1) \right\rangle \left| \phi_{0j}(\vec{x}_2) \right\rangle - \left| \phi_{0j}(\vec{x}_1) \right\rangle \left| \phi_{0i}(\vec{x}_2) \right\rangle \right) \right], \end{aligned} \tag{A.7}$$

where the zeroes denote quantities that solve the variational equation (A.3). The slowest step of the whole calculation is the evaluation of "two-body integrals" like $\left\langle \phi_i(\vec{x}_1) \right| \left\langle \phi_j(\vec{x}_2) \right| \hat{v}(\vec{x}_1, \vec{x}_2) \left| \phi_k(\vec{x}_1) \right\rangle \left| \phi_l(\vec{x}_2) \right\rangle$ and as such, the algorithm scales as $O(N^4)$. Note, however, that this scaling can be improved by employing more advanced methods of evaluation of the interaction and exchange terms. The Hartree–Fock approximation is usually too drastic to yield quantitatively useful results. The results obtained from the method are thus usually further improved by including the electron correlation either perturbatively [70], variationally [71, 72], or by other correction schemes [73]. Such improvements are however quite costly and scale quite unfavorably with the number of electrons. An alternative approach to achieve more accurate results is thus often needed, when solving the problem for a system of a comparatively high number of electrons, or when the calculations are needed to be done quickly. The most often used such alternative is *Density functional theory*. The theory itself is based on the Hohenberg–Kohn theorems [74], which, in short, state that all properties of the many-electron system can be deduced from the ground state electron density:

$$\rho_0(\vec{x}_1) = N \int_{\mathrm{d}} \vec{x}_2 \mathrm{d}\vec{x}_3 \ldots \mathrm{d}\vec{x}_N |\psi_0(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N)|^2, \tag{A.8}$$

and this density in turn can be found by minimizing the ground state energy functional $E_0[\rho]$ with respect to the trial ground state density. The theory thus presents a possibility for a considerable reduction in the dimensionality of the problem — from the original $3N$ dimensions to only 3 dimensions

in total. To solve the electronic problem in such a way, however, would require the functional $E_0[\rho]$ to be known, which is not the case. Furthermore, no systematic way of finding such a functional exists. Particularly problematic is the kinetic energy contribution to the functional $T[\rho]$. The most often used Kohn–Sham [75] approach to density functional theory takes advantage of the fact that the ground state wave function of the form A.2 provides a reasonable approximation to the quantum kinetic energy (calculated from this wave function in the usual way using the Laplace operator). The approach thus seeks to find such a wave function as a ground state of a Hamiltonian with the potential energy chosen in such a way that the resulting electron density exactly matches the exact ground state density of the original problem. The problem thus reduces again to the solution of a set of equations (the Kohn–Sham equations) which look like the original Hartree–Fock equations, with the two-body terms replaced by the effective potential:

$$\left(-\frac{1}{2}\nabla^2_{\vec{x}_1} + v_{\text{eff}}(\vec{x}_1)\right)\left|\phi_i(\vec{x}_1)\right\rangle = \epsilon_i \left|\phi_i(\vec{x}_1)\right\rangle. \tag{A.9}$$

The effective potential can nowadays be reasonably approximated and the equations solved in a basis. The scaling of the method with the number of electrons tends to be of the order $O(N^3)$ or lower.

Appendix B Classical approximation

This appendix will briefly deal with the motivation and applicability of replacing the quantum description of a system of particles with the corresponding classical description. Consider a single particle described by a wave function, which solves the time-dependent Schrödinger equation:

$$\frac{\partial \psi(\vec{x},t)}{\partial t} = \hat{H}\psi(\vec{x},t) \tag{B.1}$$

$$\hat{H} = -\frac{\hbar^2}{2m}\nabla^2 + V(\vec{x}). \tag{B.2}$$

This wave function can be written using the ansatz:

$$\psi(\vec{x},t) = A(\vec{x},t)\exp\left(i\frac{S(\vec{x},t)}{\hbar}\right), \tag{B.3}$$

where A and S are both taken to be real, and the global phase can be fixed such, that $A > 0$. Substituting the wave function in such a form into the original time-dependent Schrödinger equation and solving the real and imaginary parts separately yields the following two equations:

$$\frac{\partial S}{\partial t} + \frac{(\nabla S)^2}{2m} + V = \frac{\hbar^2\nabla^2 A}{2mA} \tag{B.4}$$

$$\frac{\partial A}{\partial t} + \frac{(\nabla A)\cdot(\nabla S)}{m} + \frac{A\left(\nabla^2 S\right)}{2m} = 0. \tag{B.5}$$

Note that this reformulation of the time-dependent Schrödinger equation in terms of the quantities A and S also plays a central role in the De Broglie–Bohm interpretation of quantum mechanics [76]. The second of these equations can be rewritten in the form:

$$\frac{\partial A^2}{\partial t} + \frac{\nabla\left(A^2\nabla S\right)}{m} = 0, \tag{B.6}$$

which is just the continuity equation for the probability density:

$$\frac{\partial \rho}{\partial t} + \nabla \boldsymbol{j}_\rho = 0. \tag{B.7}$$

Taking the classical limit $\hbar \to 0$, equation B.4 becomes:

$$\frac{\partial S}{\partial t} + \frac{(\nabla S)^2}{2m} + V = 0, \tag{B.8}$$

in which the Hamilton-Jacobi equation of motion of classical mechanics [77] can be recognized:

$$\frac{\partial S}{\partial t} + H(\vec{x}, \nabla S) = 0. \tag{B.9}$$

The connection between quantum and classical mechanics can thus be established by identifying the phase function $S(x,t)$ in the classical limit as the classical action (as a function of the trajectory endpoint position and time).

Appendix C Exchange symmetry and path integrals

To account for the indistinguishability of particles, the expressions obtained in subsection 2.2 need to be symmetrized, in the case of bosonic particles, or anti-symmetrized in the case of fermions:

$$Z_{\pm} = \frac{1}{N!}\sum_{P_N}(\pm 1)^{\Pi(P_N)}\int \mathrm{d}^n x \left\langle P_N(x)\right| \exp\left(-\beta\hat{H}\right) \left|x\right\rangle, \tag{C.1}$$

where the sum is understood to go over all permutations of the particles, $\Pi(P_N)$ is the parity of the permutation, and $P_N(x)$ are the permuted coordinates of the particles. Performing a permutation does not change the form of the path integral expression (2.33), however, it changes the condition $x_P = x_0$ to $x_P = P_N(x_0)$. To put this into words, using again the "integral over paths" interpretation: each of the terms in equation C.1 represents an integral of $\exp(-S_\mathrm{E})$ over all paths, where after the imaginary time interval β, each of the particles ends at the starting point of another particle (determined by the particular P_N). If the individual particles are assumed to be well localized (as is the case for atomic nuclei in molecular systems in the temperature ranges considered), the terms where the particles are permuted can be expected to be suppressed by a factor of the order of magnitude of:

$$\exp\left(-\frac{mD^2}{2\hbar^2\beta}\right), \tag{C.2}$$

D being the mean distance between the particles in their well-localized positions. The contributions are thus strongly suppressed with increasing interparticle distance. Using conservative estimates of these parameters for a molecular system, $m = m_P$, T=300 K, D=1 Å, this order of magnitude can be estimated to be $\approx 10^{-2}$ or less. Based on this, it can be assumed that the individual nuclei are safe to be approximated as distinguishable particles, and thus equations 2.29 and 2.33 can be used to describe the statistics of the atomic nuclei in the systems considered.

Appendix D A global version of the Langevin thermostat

This appendix will give a brief explanation of how one can obtain the equations defining the CSVR thermostat 2.140. The explanation is based on ref. 32 where the topic is covered in full detail.

The formal solution of the Langevin equation 2.83 for a free particle in one dimension will serve as a starting point. Setting $t = \epsilon$ and $c = \exp\left(-2\gamma\epsilon\right)$, this equation reads:

$$p(\epsilon) = \sqrt{c}p(0) + \sqrt{\left[\frac{m}{\beta}(1-c)\right]}\zeta. \tag{D.1}$$

Considering the system to be composed of N copies of this particle (this is done for simplicity and can be equally thought of as $N/3$ free Brownian particles in three dimensions), denoting the momentum of each one p_i, the change of the kinetic energy

$$K = \sum_{i=1}^{N} \frac{p_i^2}{2m} \tag{D.2}$$

going from $t = 0$ to $t = \epsilon$ is equal to:

$$K(\epsilon) - K(0) = \sum_{i=1}^{N}\left[\frac{(c-1)p_i^2(0)}{2m} + \sqrt{\frac{c(1-c)}{m\beta}}p_i(0)\zeta_i + \frac{(1-c)\zeta_i^2}{2m\beta}\right], \tag{D.3}$$

where ζ_i is a vector of statistically independent normally distributed random numbers with unit variance.

The goal is to obtain an algorithm that enforces the same variation and mean value ($K = N/(2\beta)$) of the kinetic energy by minimally perturbing the momenta. Defining the magnitude of the perturbation to be

$$\delta = \frac{\sum_{i=1}^{N}\left(p_i(\epsilon) - p_i(0)\right)^2}{m}, \tag{D.4}$$

then the perturbation to the momenta that minimizes this magnitude while keeping the chosen increment of kinetic energy constant is

$$p_i(\epsilon) = \alpha(0)\,p_i(0). \tag{D.5}$$

For an instructive illustration of this fact see again ref. 32. This is the third of the final CSVR equations 2.140. The increment of the kinetic energy is equal to:

$$K(\epsilon) - K(0) = \sum_{i=1}^{N}\frac{\left(\alpha^2 - 1\right)}{2m}p_i^2(0). \tag{D.6}$$

Setting this increment equal to the change of the kinetic energy according to the equation D.3 yields the first of the CSVR equations 2.140. The only non-trivial step involves using the fact that the sum of two normally distributed numbers is also a normally distributed number:

$$\mathcal{N}(\mu_a, \sigma_a^2) + \mathcal{N}(\mu_b, \sigma_b^2) = \mathcal{N}(\mu_a + \mu_b, \sigma_a^2 + \sigma_b^2), \tag{D.7}$$

and the property that a number $x = \sigma\zeta + \mu$ is normally distributed with standard deviation σ and mean μ. Together these imply:

$$\sum_{i=1}^{N} p_i \zeta_i = \zeta \sqrt{\sum_{i=1}^{N} p_i^2}. \tag{D.8}$$

The remaining single Gaussian distributed number ζ needs to be chosen to be the same as one of the numbers ζ_i to ensure proper correlation between the individual terms in the resulting equation.

The second and final CSVR equation, eq. 2.140, which determines the sign of α, comes from a more detailed analysis. The main motivation for choosing the sign in this fashion comes from the requirement to keep the probability to observe a flip of the momenta finite while yielding correct limits in the case of large and small N. For large N, this probability should be negligible, while for $N = 1$, this probability should be the same as given by the Langevin equation.