JOHANNES GUTENBERG UNIVERSITÄT MAINZ

JG|U

Novel Algorithms for Electronic Structure Based Molecular Dynamics with Linear System-Size Scaling

Dissertation zur Erlangung des Grades eines
„Doktor rerum naturalium (Dr. rer. nat.)“ der Fachbereiche:
08 - Physik, Mathematik und Informatik,
09 - Chemie, Pharmazie und Geowissenschaften,
10 - Biologie,
04 - Universitätsmedizin
der Johannes Gutenberg-Universität Mainz

von
Dorothee Richters
aus
Wiesbaden

Max Planck Graduate Center
mit der Johannes Gutenberg-Universität Mainz

Juni 2014

Datum der Einreichung: 26.06.2014

Referent:

Koreferent:

Tag der mündlichen Prüfung: 19.09.2014

Abstract

This thesis addresses the development and improvement of linear scaling algorithms for electronic structure based molecular dynamics.
Molecular dynamics is a method for computer simulation of the complex interplay between atoms and molecules at finite temperature. Important advantages of this method are its high accuracy and predictive power. But the computational effort, which generally scales cubically with the number of atoms, hinders its applications to large systems and long time scales.
With a new formalism, based on the grand-canonical potential and a factorization of the density matrix, the diagonalization of the corresponding Hamiltonian matrix is avoided. It exploits the fact that the Hamiltonian and the density matrix are sparse due to localization. This reduces the complexity of the calculations, so that linear scaling with respect to the system's size is achieved.
To demonstrate its efficiency, the resulting algorithm is applied to a system of liquid methane, exposed to extreme pressure (around 100 GPa) and temperature (2000 - 8000 K). In the simulations, methane dissociates at temperatures exceeding 4000 K. The formation of sp^2-bonded polymeric carbon is observed. The simulations provide no evidence for the formation of diamond and therefore have an impact on the hitherto planetary models of Neptune and Uranus.
As the circumvention of the diagonalization of the Hamiltonian entails the inversion of matrices, the problem of calculating the (inverse) p-th root of a given matrix is further addressed. It results in a new formula for symmetric positive definite matrices that generalizes the Newton-Schulz iteration, Altman's scheme for bounded and non-singular operators, and Newton's method for finding roots of functions. The proof is furnished that the order of convergence is always at least quadratic and that adaptively adjusting a parameter q leads in all cases to a better performance.

Zusammenfassung

Die vorliegende Arbeit behandelt die Entwicklung und Verbesserung von linear skalierenden Algorithmen für Elektronenstruktur basierte Molekulardynamik. Molekulardynamik ist eine Methode zur Computersimulation des komplexen Zusammenspiels zwischen Atomen und Molekülen bei endlicher Temperatur. Ein entscheidender Vorteil dieser Methode ist ihre hohe Genauigkeit und Vorhersagekraft. Allerdings verhindert der Rechenaufwand, welcher grundsätzlich kubisch mit der Anzahl der Atome skaliert, die Anwendung auf große Systeme und lange Zeitskalen.

Ausgehend von einem neuen Formalismus, basierend auf dem großkanonischen Potential und einer Faktorisierung der Dichtematrix, wird die Diagonalisierung der entsprechenden Hamiltonmatrix vermieden. Dieser nutzt aus, dass die Hamilton- und die Dichtematrix aufgrund von Lokalisierung dünn besetzt sind. Das reduziert den Rechenaufwand so, dass er linear mit der Systemgröße skaliert.

Um seine Effizienz zu demonstrieren, wird der daraus entstehende Algorithmus auf ein System mit flüssigem Methan angewandt, das extremem Druck (etwa 100 GPa) und extremer Temperatur (2000 - 8000 K) ausgesetzt ist. In der Simulation dissoziiert Methan bei Temperaturen oberhalb von 4000 K. Die Bildung von sp^2-gebundenem polymerischen Kohlenstoff wird beobachtet. Die Simulationen liefern keinen Hinweis auf die Entstehung von Diamant und wirken sich daher auf die bisherigen Planetenmodelle von Neptun und Uranus aus.

Da das Umgehen der Diagonalisierung der Hamiltonmatrix die Inversion von Matrizen mit sich bringt, wird zusätzlich das Problem behandelt, eine (inverse) p-te Wurzel einer gegebenen Matrix zu berechnen. Dies resultiert in einer neuen Formel für symmetrisch positiv definite Matrizen. Sie verallgemeinert die Newton-Schulz Iteration, Altmans Formel für beschränkte und nicht singuläre Operatoren und Newtons Methode zur Berechnung von Nullstellen von Funktionen. Der Nachweis wird erbracht, dass die Konvergenzordnung immer mindestens quadratisch ist und adaptives Anpassen eines Parameters q in allen Fällen zu besseren Ergebnissen führt.

Contents

List of Abbreviations

AIMD	Ab Initio Molecular Dynamics
BOMD	Born-Oppenheimer Molecular Dynamics
CRS	Compressed Row Storage
DFT	Density Functional Theory
DMM	Density Matrix Minimization
FOE	Fermi Operator Expansion
GCP	Grand-Canonical Potential
HF	Hartree-Fock
HK	Hohenberg-Kohn
KS	Kohn-Sham
LCAO	Linear Combination of Atomic Orbitals
LCN	Local Charge Neutrality
MD	Molecular Dynamics
PCF	Pair Correlation Function
SCF	Self-Consistent Field
SCTB	Self-Consistent Tight-Binding
TB	Tight-Binding

Chapter 1

Introduction

Ab initio molecular dynamics is a method for computer simulation of the complex interplay between electrons, atoms, and molecules at finite temperature. To achieve this, the time evolution of atoms is calculated by solving Newton's equations of motion. In contrast to molecular dynamics simulations with empirical interaction potentials, the forces are obtained "on the fly" with the aid of parameter-free electronic structure calculations. Two key advantages of this method are transferability and high accuracy, which are important, if aiming to make predictions for unknown systems under conditions where experimental information is not yet available.

The main challenge, when performing *ab initio* molecular dynamics calculations, is that a huge Hamilton matrix has to be diagonalized in every time step in order to extract the energy eigenvalues. The computational effort is generally of order $\mathcal{O}(N^3)$, where N denotes the number of atoms. Developing new methods that scale linearly with respect to the size of the system is therefore a desirable aim. The central question in this work is to circumvent this computationally expensive operation of diagonalizing, and to demonstrate its capability in selected applications.

To achieve this, we use an alternative field-theoretic method suitable for linear scaling molecular dynamics simulations using forces from self-consistent electronic structure calculations. It is based on an exact decomposition of the grand-canonical potential for independent fermions and does not rely on either the ability to localize the orbitals or whether the Hamilton operator is well-conditioned. We compute the finite-temperature density matrix, or

Fermi matrix, by a hybrid approach. Taking inspiration from the Fermi operator expansion method, the Fermi operator is decomposed into a sum of matrices, which have to be inverted. The inversion of these matrices is done by Chebyshev polynomial expansion for the large number of well-conditioned matrices and Newton-Schulz iteration for the few remaining ill-conditioned matrices. Our method enables highly accurate all-electron linear scaling calculations, including for metallic systems. The problem of inherent energy drift of Born-Oppenheimer molecular dynamics simulations, arising from an incomplete convergence of the self-consistent field cycle, is solved by using a properly modified Langevin equation. In this way, we reduce the complexity of the calculations so that linear scaling with respect to the system size is achieved.

We illustrate the predictive power of this approach using the example of liquid methane at planetary conditions. Methane occurs in the middle ice layer of the giant gas planets Uranus and Neptune. In this layer, at a depth of one-third of the planetary radius, pressure and temperature range from 20 GPa and 2000 K to 600 GPa and 6000 K, which we simulate by means of large-scale electronic structure based molecular dynamics using our method. We address the controversy of whether or not the interior of Uranus and Neptune consists of diamond. We find no evidence for the formation of diamond, but rather carbon chains and sp^2-bonded polymeric carbon. Furthermore, we predict that at high temperature hydrogen may exist in its mono-atomic and metallic state.

A time consuming step in our linear scaling scheme is the part, where ill-conditioned matrices have to be inverted. We present here a new iteration scheme to calculate the inverse p-th root of symmetric positive definite matrices. We show that the order of convergence is always quadratic and that in the case $p = 1$ we have an arbitrary order of convergence, contingent upon a parameter q. By choosing q adaptively, better results than with before known formulas of this type can be achieved as less iterations and matrix-matrix multiplications are required. This iteration scheme emerges as a generalization of the Newton-Schulz iteration, Altman's method, and Newton's method for finding roots of functions. Its performance is evaluated by a MATLAB code using random matrices with different spectral radii.

The present work is organized as follows. In Chapter 2, we introduce the principles of electronic structure calculations and explain two methods that are important in this work, i.e. density functional theory and tight-binding. An introduction to linear scaling methods is presented in Chapter 3. Chapter 4 explains the field-theoretic approach to linear scaling and its implementation. The application to liquid methane at planetary pressure and temperature conditions as well as the corresponding results and discussion can be found in Chapter 5. Chapter 6 is dedicated to the determination of the inverse p-th root of a symmetric positive definite matrix in an efficient way. The last chapter contains the conclusion of our work and gives an outlook to possible future projects.

Publications

The following results of this work have been published in peer reviewed journals.

Chapter 4: D. Richters and T. D. Kühne. "Self-consistent field theory based molecular dynamics with linear system-size scaling", J. Chem. Phys., 140(13):134109, 2014.

Chapter 5: D. Richters and T. D. Kühne. "Liquid methane at extreme temperature and pressure: Implications for models of Uranus and Neptune", JETP Lett., 97(4):184-187, 2013.

The following result is expected to be published soon.

Chapter 6: D. Richters and T. D. Kühne. "An algorithm to calculate the inverse principal p-th root of symmetric positive definite matrices"

Chapter 2

Density Functional Theory and Tight-Binding

In this chapter, we describe the essential theory for quantum mechanical calculations that will be employed in this work.
The main objective in quantum mechanics is to solve the Schrödinger equation

$$\hat{H}\Psi = E\Psi \tag{2.1}$$

for the system of interest. When dealing with a complex system, a series of approximations becomes essential. This is due to the fact that solving this equation (2.1) for a given many-electron wavefunction Ψ is so complex that it is analytically impossible and even numerically intractable. With these approximations, algorithms can be designed in order to perform numerical calculations. For electronic structure calculations, density functional theory (DFT) is a method that provides a good balance between efficiency and accuracy, making it possible to handle more than just a handful of electrons, while still providing reasonable accuracy.

If a system is too large for DFT calculations, one can use the so-called tight-binding (TB) method, which is a less costly method that enables the modelling of larger systems. It can be derived from DFT through the the Harris-Foulkes functional [34, 48], which allows to get an estimate for the energy of a molecule from the electronic structure of its atoms with computational cost less than DFT. In the TB method, the eigenstates of the Hamiltonian are written in an

atomic-like basis and the true Hamiltonian is replaced by a parametrized and simplified one [44]. These two techniques, DFT and TB, are described in the following subsections.

2.1 The Born-Oppenheimer Approximation

The Hamiltonian $\hat{H}$ of a system for representing the interaction of nuclei and electrons can be written in the following way

$$\hat{H} = \hat{T}_N + \hat{T}_e + \hat{V}_{NN} + \hat{V}_{ee} + \hat{V}_{Ne}, \tag{2.2}$$

where $\hat{T}_N$ and $\hat{T}_e$ are the kinetic energies of the nuclei (N) and the electrons (e) and $\hat{V}_{NN}, \hat{V}_{Ne}$ and $\hat{V}_{Ne}$ are the potential energies of the repulsion of the nuclei (NN) and the electrons (ee) and the attraction between the nuclei and the electrons (Ne).
The Born-Oppenheimer approximation [14] is based on the observation that the nuclei are around 2000 times heavier than the electrons. In a classical picture, this means that the time scales on which electrons and nuclei move are significantly different and the velocities of the electrons are much larger than those of the nuclei. The nuclei can be treated as fixed while solving the electronic problem. This leads to the possibility to factorize the wavefunction $\Psi(R_1, \ldots, R_I, r_1, \ldots, r_n)$ as a product of a nuclear wavefunction $\chi(R_1, \ldots R_I)$, depending only on the positions R_J of the I nuclei and an electronic wavefunction. The electronic wavefunction $\Phi(r_1, \cdots r_n; R_1, \cdots R_I)$ is a function of the positions r_i of the n electrons whereas the positions of the nuclei are parameters. So we get

$$\Psi(\{R_J\}, \{r_i\}) = \chi(\{R_J\})\Phi(\{r_i\}; \{R_J\}). \tag{2.3}$$

2.2 Density Functional Theory

DFT is a quantum mechanical method for correlated many-body systems. The properties of a such a system can be determined by using a functional of the electron density [56]. DFT emerged as an important tool for quantitative studies of molecules and the calculation of electronic structure in condensed

matter physics. In the following subsection, we describe the two theorems of Hohenberg and Kohn (HK) that build the foundation of DFT and allow us to find the ground state properties of a given system by just dealing with the ground state density. Thereafter, we introduce the Kohn-Sham (KS) equations that provide a practical approach to DFT. They open up the possibility to accurately determine the ground-state density of particles interacting with each other by calculating the ground-state density of an auxiliary system of non-interacting particles.

2.2.1 The Hohenberg-Kohn Theorems

The principle that makes DFT so useful is that the properties of a many-electron system can be determined by evaluating a functional of the ground state density $\rho_0(r)$ [101]. The uniqueness of such a functional and therefore the concept of DFT is justified by the two theorems of Hohenberg and Kohn [56]. They were the first to describe DFT as a formal exact theory that can be applied to any system of particles interacting in an external potential [101].
The first HK theorem states that the density is up to a trivial constant uniquely determined for any system of interacting particles in an external potential $v_{\text{ext}}(r)$. This is the basis for reducing the N-electronic many-body problem with $3N$ spatial coordinates to the three spatial coordinates that the electron density depends on.
The second HK theorem affirms that one can define a universal functional of the density $\rho(r)$ for the energy $E[\rho(r)]$ and that its global minimum value matches with the exact ground state. So for an arbitrary density $\rho'(r)$ with $\rho'(r) > 0$ for all position vectors r and $\int \rho'(r)d^3(r) = N$, we have always $E_0 \leq E[\rho'(r)]$ [56]. While the uniqueness of such a functional of the ground state density has been proven, the exact construction is only solved for one-electron systems.
We also have the problem of representability in DFT. If we have N particles, we want to write the particle density as follows

$$\rho(r) = N \iint \dots \int \Psi^*(r, r_2, \dots, r_N)\Psi(r, r_2, \dots, r_N)\mathrm{d}^3r_2\mathrm{d}^3r_3 \dots \mathrm{d}^3r_N. \quad (2.4)$$

The question whether one can write an arbitrary density in this above presented way, meaning that this density arises from an antisymmetric N-body wavefunction, is called N-representability problem. The question if a density

written in the above manner is corresponds to the ground-state density of a local external potential $v_{\text{ext}}(r)$ is called v-representability problem. While the question of N-representability has received a favourable answer, there is no known solution to the problem of v-representability [20, 101]. We can deduce from the first HK theorem that the potential for any density is unique but we have no evidence of its existence. However, the constrained search algorithm of Levy [89] and Lieb [93] shows that this is not necessary in the proof of the HK theorem[20].
In principle, we can summarize the concept of DFT as follows. The knowledge of the density entails the knowledge of the wavefunction $\Psi(r_1, \dots r_N)$ and the potential. Thus we can calculate all other observables [20]. From this, it follows that we can write the total energy as

$$E[\rho(r)] = \int \rho(r) v_{\text{ext}}(r) \mathrm{d}r + F[\rho(r)], \tag{2.5}$$

where

$$F[\rho(r)] = T[\rho(r)] + V_{ee}[\rho(r)] \tag{2.6}$$

is a universal functional which does not explicitly depend on the external potential. The two HK theorems yield a framework for obtaining the ground state properties. Hereby, an approximation for $F[\rho(r)]$ has to be found.

2.2.2 The Kohn-Sham Equations

In this subsection, we explain the method of Kohn and Sham [75], presented in 1965. Their idea was to construct a fictitious non-interacting system that has exactly the same electron density $\rho(r)$ as the system of interacting electrons. As the kinetic energy of a non-interacting system $T_s[\rho(r)]$ is known

$$T_s[\rho(r)] = -\frac{1}{2} \sum_{i=1}^{N_e} \int \psi_i^*(r) \nabla^2 \psi_i(r) \mathrm{d}r = T_s[\{\psi_i[\rho(r)]\}], \tag{2.7}$$

where ψ_i are the fictitious single-particle wavefunctions or KS orbitals, more accurate DFT calculations can be performed. The challenge hereby is to find this fictitious system. Then, the wavefunction of the non-interacting electrons can be computed and the ground-state density $\rho_0(r)$ can be built from the resulting molecular orbitals. With that, the ground-state energy $E[\rho_0(r)]$ can

be computed [83].

The bright idea of Kohn and Sham was to write the energy functional of the electronic problem

$$E[\rho(r)] = T[\rho(r)] + V_{Ne}[\rho(r)] + V_{ee}[\rho(r)] \tag{2.8}$$

as

$$\begin{aligned} E_{\mathrm{KS}}[\rho(r)] &= E_{\mathrm{KS}}[\{\psi_i[\rho(r)]\}] \\ &= T_s[\{\psi_i[\rho(r)]\}] + U_H[\rho(r)] + V_{Ne}[\rho(r)] + E_{\mathrm{xc}}[\rho(r)], \end{aligned} \tag{2.9}$$

where

$$U_H[\rho(r)] = \frac{1}{2}\iint \frac{\rho(r)\rho'(r)}{|r-r'|}\mathrm{d}r'\mathrm{d}r \tag{2.10}$$

is the Hartree energy, and

$$E_{\mathrm{xc}}[\rho(r)] = T[\rho(r)] - T_s[\{\psi_i[\rho(r)]\}] + V_{ee}[\rho(r)] - U_H[\rho(r)] \tag{2.11}$$

is the exchange-correlation energy functional. It is important to notice that this results in a formally exact theory, provided that $E_{\mathrm{xc}}[\rho(r)]$ is known.
Now, the aim is to minimize equation (2.9). As this is not straightforward, we make it stationary with the help of an Euler-Lagrange equation [83]

$$0 = \frac{\delta E_{\mathrm{KS}}[\rho(r)]}{\delta\rho(r)} = \frac{\delta E_{\mathrm{KS}}[\{\psi_i[\rho(r)]\}]}{\delta\rho(r)} \tag{2.12}$$

$$= \frac{\delta T_s[\{\psi_i[\rho(r)]\}]}{\delta\rho(r)} + \frac{\delta U_H[\rho(r)]}{\delta\rho(r)} + \frac{\delta V_{Ne}[\rho(r)]}{\delta\rho(r)} + \frac{\delta E_{\mathrm{xc}}[\rho(r)]}{\delta\rho(r)} \tag{2.13}$$

$$= \frac{\delta T_s[\{\psi_i[\rho(r)]\}]}{\delta\rho(r)} + v_H(r) + v_{\mathrm{ext}}(r) + v_{\mathrm{xc}}(r), \tag{2.14}$$

where $\frac{\delta\bullet}{\delta\rho(r)}$ denotes the functional derivative with respect to the electron density $\rho(r)$. Furthermore,

$$v_H(r) = \frac{1}{2}\int \frac{\rho(r')}{|r-r'|}\mathrm{d}r' \tag{2.15}$$

is the Hartree potential, $v_{\mathrm{ext}}(r)$ is the external potential and $v_{\mathrm{xc}}(r)$ is the exchange-correlation potential. If we consider a non-interacting system, the

Euler-Lagrange equation is simplified as the derivatives with respect to $\rho(r)$ of the Hartree energy and the electron-ion interaction energy vanish. We get

$$0 = \frac{\delta T_s[\{\psi_i[\rho(r)]\}]}{\delta\rho(r)} + v_s(r), \tag{2.16}$$

where $v_s(r)$ is the effective potential of the non-interacting system. By solving the latter equation, we solve the non-interacting one particle system to get the solution of the interacting many-body system. We take a fictitious single-particle sytem, called the KS system, with an effective potential

$$v_{\mathrm{KS}}(r) = v_H(r) + v_{\mathrm{ext}}(r) + v_{\mathrm{xc}}(r), \tag{2.17}$$

called KS potential. The KS potential is chosen such that its density is the same as the density of the interacting system. The equations that result from this approach are known as the KS equations

$$\left(-\frac{1}{2}\nabla^2 + v_{\mathrm{KS}}(r)\right)\psi_i(r) = \varepsilon_i\psi_i(r) \tag{2.18}$$

$$\rho_s(r) = \sum_{i=1}^{N_{occ}} f_i\psi_i(r)\psi_i^*(r) = \rho(r), \tag{2.19}$$

where N_{occ} is the number of occupied orbitals, and $\{f_i\}_{i=1}^{N_{occ}}$ are the corresponding occupation numbers fulfilling

$$\sum_{i=1}^{N_{occ}} f_i = N_e, \tag{2.20}$$

and ε_i are the KS eigenvalues. Then, the total energy can be calculated from the energy functional

$$E_{\mathrm{KS}}[\rho(r)] = \sum_{i=1}^{N_{occ}} f_i\varepsilon_i - \frac{1}{2}\iint\frac{\rho(r)\rho'(r)}{|r-r'|}\mathrm{d}r'\mathrm{d}r - \int v_{\mathrm{xc}}(r)\rho(r)\mathrm{d}r + E_{\mathrm{xc}}[\rho(r)]. \tag{2.21}$$

The direct calculation of the energy by the solution of the Schrödinger equation for the non-interacting KS orbitals ψ_i scales cubically with the size of the system. However, the KS equations provide a method for finding the ground state energy of an interacting system. It is an exact method, assumed that $E_{\mathrm{xc}}[\rho(r)]$ is known, which is in general not the case. However, very useful approximations do exist [126]. In electronic structure calculations, the local density approximation or the generalized-gradient approximation are often used to get an approximate exchange-correlation functional.

2.3 Tight-Binding

In this section, we describe the TB bond model, which is the basis for the Hamiltonian set-up in our calculations. TB is, in comparison with *ab initio* methods, computationally less expensive, but comes along with a decrease in transferability. In contrast to pure empirical methods, TB preserves the quantum mechanical structure of bonding, but is computationally less efficient [44]. TB can be derived from DFT through the Harris-Foulkes functional [34, 48], which yields an approximation for the energy without solving the KS equations. Hereby, the electronic density is written as a superposition of atomic charge densities and the energy is only calculated in terms of this density.
The cubic scaling of the diagonalization of the Hamiltonian is the bottleneck in electronic structure calculations, and it is therefore desirable to reduce the computational effort. TB is a simplified approach to the electronic problem and a central advantage is the fact that the diagonalization can be effectuated relatively easy, so that linear scaling can be achieved. The first attempt to reduce the complexity is to neglect the core electrons which play only a minor role in chemical bonding and replace them, together with the nucleus, with a pseudopotential or by an effective charge distribution. Additionally, one can use a minimal basis set. This leads to a reduction of the size of the Hamiltonian, which considerably lowers the computational cost of the diagonalization [72]. In the second step, the construction of the Hamiltonian can be simplified. Hereby, the exact many-body Hamiltonian is replaced by a Hamiltonian that depends only parametrically on the nuclear positions. In this approach, one deals with an atomic-like basis set that has the same symmetry properties as the atomic orbitals.

2.3.1 The Tight-Binding Hamiltonian

As its name already indicates, the electrons in the TB model are supposed to be tightly bound to the atom to which they belong. The interaction of the electrons with neighbouring atoms is limited. This entails that the wavefunction of the electron resembles the atomic orbital of the atom.

We now want to find the electronic band-structrure of the system by solving the Schrödinger equation

$$\hat{H}\Psi_k(r) = E(k)\Psi_k(r), \tag{2.22}$$

where $\hat{H}$ is the Hamiltonian, $\Psi_k(r)$ is the wavefunction and k the wave vector [134]. To solve this, one can write $\Psi_k(r)$ according to Bloch's theorem [13] as

$$\Psi_k(r) = \frac{1}{\sqrt{N}} \sum_R e^{ik\cdot R}\Phi(r-R), \tag{2.23}$$

where N is the number of unit cells and R is a lattice vector. This wavefunction has the periodicity of the lattice, and therefore it fulfils for an arbitrary R

$$\Psi_k(r) = \Psi_k(r+R). \tag{2.24}$$

In equation (2.23), $\Phi(r)$ is the wavefunction on a unit cell of the system, which can be constructed as a sum of a set of selected orbitals of the atoms in the unit cell

$$\Phi(r) = \sum_{i=1}^{n}\sum_{\alpha=1}^{m} c_i^\alpha \phi_i^\alpha(r). \tag{2.25}$$

Thereby, $\phi_i^\alpha(r)$ refers to the α-th orbital of atom i in the unit cell, and c_i^α are the corresponding coefficients. This is known as linear combination of atomic orbitals (LCAO). If only one orbital per atom is used, equation (2.25) reduces to

$$\Phi(r) = \sum_{i=1}^{n} c_i \phi_i(r). \tag{2.26}$$

Now, we consider the Hamiltonian itself

$$\hat{H} = -\frac{1}{2}\nabla^2 + \sum_R U_{\text{at}}(r-R), \tag{2.27}$$

where $U_{\text{at}}(\cdot)$ is a rotation symmetric atomic potential. If we apply this Hamiltonian to the atomic orbital ϕ_i^α, we get

$$\hat{H}\phi_i^\alpha = \varepsilon_i^\alpha \phi_i^\alpha + \left[\sum_{R\neq 0} U_{\text{at}}(r-R)\right]\phi_i^\alpha, \tag{2.28}$$

where ε_i^α is the solution of the atomic Hamiltonian

$$\hat{H}_{\mathrm{at}}\phi_i^\alpha(r) = \left[-\frac{1}{2}\nabla^2 + U_{\mathrm{at}}(r)\right]\phi_i^\alpha = \varepsilon_i^\alpha\phi_i^\alpha. \tag{2.29}$$

By multiplying equation (2.28) on the left by $\langle\phi_j^\beta\mid$, we can calculate the matrix elements of the Hamiltonian as [100]

$$\hat{H}_{ji}^{\beta\alpha} = \langle\phi_j^\beta \mid \hat{H} \mid \phi_i^\alpha\rangle, \tag{2.30}$$

and the overlap matrix S as

$$S_{ji}^{\beta\alpha} = \langle\phi_j^\beta \mid \phi_i^\alpha\rangle. \tag{2.31}$$

The energy of the electron described by $\Psi_k(r)$ can be calculated through

$$E(k) = \langle\Psi_k \mid \hat{H} \mid \Psi_k\rangle. \tag{2.32}$$

The breakthrough for the TB method was provided by Slater and Koster in 1954 [143]. They proposed a modification of the LCAO method, and showed that the large number of integrals that have to be evaluated to calculate the Hamiltonian matrix elements, can be replaced by a parametrized form depending only on the internuclear distances and the symmetries of the orbitals.
By using Löwdins method [98], we replace the orbitals ϕ_i^α by wavefunctions χ_i^α that have the same symmetry properties but are mutually orthogonal [44]

$$\chi_i^\alpha = \sum_{i'\alpha'}(S^{-1/2})_{ii'}^{\alpha\alpha'}\phi_{i'}^{\alpha'}. \tag{2.33}$$

As described in the previous section, we can describe the system by a set of non-interacting single-particle wavefunctions, which, according to Bloch's theorem, can be written as a Bloch sum [124]

$$(\mathcal{X}_i^\alpha)_k(r) = \frac{1}{\sqrt{N}}\sum_R e^{ik\cdot R}\chi_i^\alpha(r - R - R_i), \tag{2.34}$$

where R_i are the positions of the atoms within the unit cell. The sum runs over all periodic images of the orbital. Then, the Hamiltonian elements can be evaluated as a function of k as

$$H_{ij}^{\alpha\beta}(k) = \sum_R e^{ik\cdot R}\int(\chi_i^\alpha)^*(r - R - R_i)\hat{H}\chi_j^\beta(r - R_j)\mathrm{d}r. \tag{2.35}$$

In the TB model, the matrix elements between two atomic orbitals on neighbouring atoms play an important role. The two-center approximation presented by Slater and Koster replaces the integrals in (2.35) by a parameter that depends only on the displacement $\mid R_i - R_j \mid$ between the two atoms.

2.3.2 The Tight-Binding Bond Model

In this subsection, we describe the TB model developed by Sutton, Finnis, Pettifor, and Ohta [146]. Their work is based on the idea of Harris [48] and Foulkes and Haydock [34], where they made use of DFT's variational principle to get an expression for the total energy that is correct up to second order with respect to the error in the charge density.
Let ρ^{f} be a trial charge density that is a sufficiently good approximation to the exact charge density ρ^{sc}. Then, we write the effective single-particle potential, or KS potential, of the KS equations as

$$\tilde{v}_{\mathrm{KS}}(r) = v(r) + v_H^{\mathrm{f}}(r) + v_{\mathrm{xc}}^{\mathrm{f}}(r), \tag{2.36}$$

where $v(r)$ is the total ionic potential, $v_H^{\mathrm{f}}(r)$ is the Hartree potential and $v_{\mathrm{xc}}^{\mathrm{f}}(r)$ is the exchange and correlation potential, all expressed as a functional of ρ^{f}. With $\tilde{v}_{\mathrm{KS}}(r)$, one constructs an approximative single-particle Hamiltonian

$$\tilde{H} = -\frac{1}{2}\nabla^2 + \tilde{v}_{\mathrm{KS}}(r) = T + \tilde{v}_{\mathrm{KS}}(r). \tag{2.37}$$

We define the output charge density ρ^{out} as the density that is formed from the eigenstates solving the Schrödinger equation for $\tilde{H}$ once. By the work of Harris and Foulkes, we get an approximate expression for the total energy as a functional of ρ^{f} and ρ^{out}

$$\begin{aligned} E \approx & \sum_n a_n \tilde{\varepsilon}_n - \int \rho^{\mathrm{f}}(r) \left(\frac{1}{2} v_H^{\mathrm{f}}(r) + v_{\mathrm{xc}}^{\mathrm{f}}(r) \right) \mathrm{d}r + E_{\mathrm{xc}}[\rho^{\mathrm{f}}] \\ & + E_{\mathrm{ii}} + \mathcal{O}(\rho^{\mathrm{sc}} - \rho^{\mathrm{f}})^2 + \mathcal{O}(\rho^{\mathrm{sc}} - \rho^{\mathrm{out}})^2. \end{aligned} \tag{2.38}$$

Here $\tilde{\varepsilon}_n$ are the eigenvalues of $\tilde{H}$, a_n the corresponding occupation numbers, and E_{ii} is the internuclear interaction. From the variational principle follows

that the error in the total energy is second order with respect to the error in the charge density ρ^{sc}. The output charge density can be constructed from the eigenstates $|n\rangle$ of the Hamiltonian $\tilde{H}$

$$\rho^{\mathrm{out}} = \sum_n a_n |n\rangle\langle n|, \tag{2.39}$$

and we have furthermore

$$\sum_n a_n \tilde{\varepsilon}_n = \mathrm{tr}[\rho^{\mathrm{out}} \tilde{H}]. \tag{2.40}$$

This leads to a basis-independent form of equation (2.38)

$$E \approx \mathrm{tr}[\rho^{\mathrm{out}} \tilde{H}] - \mathrm{tr}[\rho^{\mathrm{f}}(v_H^{\mathrm{f}}/2 + v_{\mathrm{xc}}^{\mathrm{f}})] + E_{\mathrm{xc}}[\rho^{\mathrm{f}}] + E_{\mathrm{ii}}, \tag{2.41}$$

which is correct to the second order in the difference between the exact charge density and the trial charge density. According to Harris [48], we express the approximate charge density ρ^{f} as a superposition of atomic charge densities ρ_i

$$\rho^{\mathrm{f}} = \sum_i \rho_i. \tag{2.42}$$

Furthermore, we can write the full Hartree potential as a superposition of the Hartree potentials $(v_H^{\mathrm{f}})_i$ of the non-interacting atomic charge densities at site i

$$v_H^{\mathrm{f}} = \sum_i (v_H^{\mathrm{f}})_i. \tag{2.43}$$

We recall the definition of the approximative single particle Hamiltonian (2.37)

$$\tilde{H} = T + v_H^{\mathrm{f}} + v_{\mathrm{xc}}^{\mathrm{f}} + v.$$

Using equations (2.42) and (2.43), and adding and subtracting $\mathrm{tr}[\rho^{\mathrm{f}} \tilde{H}]$ from equation (2.41), leads to the following expression for the total energy

$$\begin{aligned} E \approx & \mathrm{tr}\left((\rho^{\mathrm{out}} - \rho^{\mathrm{f}})\tilde{H}\right) + \mathrm{tr}\left(\sum_i \rho_i \left(\sum_{j\neq i} (v_H^{\mathrm{f}})_j/2 + v_j\right)\right) + E_{\mathrm{ii}} + E_{\mathrm{xc}}[\rho^{\mathrm{f}}] \\ & - \sum_i E_{\mathrm{xc}}[\rho_i] + \sum_i \left(\mathrm{tr}\left(\rho_i(T + (v_H^{\mathrm{f}})_i/2 + v_i) + E_{\mathrm{xc}}[\rho_i]\right)\right), \end{aligned} \tag{2.44}$$

which is also correct to the second order. Here, v_j is the pseudopotential of the ionic core at site j.
Finally, we get the binding energy E_B by subtracting from (2.44) the total energy of the isolated atoms

$$E_B \approx \mathrm{tr}\left((\rho^{\mathrm{out}} - \rho^{\mathrm{f}})\tilde{H}\right) + \Delta E_{\mathrm{es}}[\rho^{\mathrm{f}}] + \Delta E_{\mathrm{xc}}[\rho^{\mathrm{f}}]. \tag{2.45}$$

Each of the terms in the sum has a clear physical meaning. Hereby, $\Delta E_{\mathrm{es}}[\rho^f]$ is the change in the electrostatic energy and $\Delta E_{\mathrm{xc}}[\rho^f]$ is the change in the exchange-correlation energy. Both are expressed as functionals of a superposition of the atomic charge densities ρ_i. The term $\mathrm{tr}(\rho^{\mathrm{out}} - \rho^{\mathrm{f}})\tilde{H}$ is the sum of the occupied eigenstates in the system minus the sum of occupied eigenstates in the free atoms and comes from the formation of bonds and the resulting charge redistribution. By further approximations, one can write the total energy as the sum of the eigenstates of a TB Hamiltonian and a sum of pair terms. This leads to the TB bond model and a physically clear form for the TB energy (2.45), which is correct to the second order [44, 146].

2.3.3 A Self-Consistent Tight-Binding Model for Hydrocarbons

In their work [57], Horsfield, Godwin, Pettifor, and Sutton presented a TB scheme for hydrocarbons, which turned out to agree very well with experiments as well as related calculations for hydrocarbons and the hydrogenated surface of diamond. Their model follows the work by Sutton et al. [146], with the difference that the repulsive energy is not given by a pair potential but by a pair functional of the charge density. One relevant feature of Sutton's model is that it brings along a form of self-consistency. Without self-consistency, TB suffers from large errors as the is no charge conservation and no considerations regarding the potential variations due to charge distribution. The easiest approach to self-consistency is local charge neutrality (LCN).
The concept of LCN is perfectly suitable for hydrocarbons, as they come along with only little charge transfer. LCN is hereby achieved by shifting the diagonal elements of the Hamiltonian until the number of electrons of every atom is equal to the number of its valence electrons. The off-diagonal elements, however, are calculated as the product of the angular factors and the two-center

integrals by Slater and Koster [143].
Details can be found in section 4.4, where we explain explicitly how LCN can be implemented.

Chapter 3

Linear Scaling Methods

The calculation of the electronic structure of materials is a task that is computationally very expensive. This results from the fact that solving the Schrödinger equation typically scales cubically with the size of the system as the diagonalization of the Hamiltonian corresponds to the solution of a high-dimensional eigenvalue problem. Thus, methods that scale linearly with the size of the system are desirable as they provide the possibility to treat considerably larger systems. However, for developing a computational chemistry method that scales linearly with the size of a system, one is in the need of approximations.

Several linear scaling methods have been developed [9, 16, 17, 35, 40, 70, 73, 90, 105, 120–122, 153], and they are all based on the concept of locality or nearsightedness which comprehends that a small disturbance in one part of the system has only a local effect on the electron density. It is based on the observation that the interaction between electrons occurs only within a finite distance from the nuclei. For insulators, we have an exponential decay of the density matrix. This enables the introduction of a localization region. The interaction is evaluated only in this region and is assumed to be zero outside. Like this, we have more vanishing matrix elements, a necessity to achieve linear scaling [40], and the matrices become sparse at last.

In traditional TB, the Schrödinger equation is solved in the reciprocal space by diagonalization of the Hamiltonian matrix. This entails a cubic scaling. In order to achieve linear scaling, the sparsity of the Hamiltonian and overlap ma-

trix is an important aspect. With the above described approximation, where only the local environment influences the bonding [15], both matrices are assumed to have non-zero elements only within a finite range.
Linear scaling methods, which have been developed in the last decades, calculate the band energy in real space. A sparse Hamiltonian can be constructed if the basis functions are localized to regions that are significantly smaller than the size of the system. The most intuitive basis functions are hereby local orbitals, on which almost all linear scaling methods rely on.

Another important criterion for the efficiency of a linear scaling algorithm is its parallelizability. Intrinsically parallel algorithms that do not need much all-to-all communication are suitable for achieving a favourable scalability.

The fundamental quantity in linear scaling techniques is the density matrix and the property that is to be exploited is its sparsity. A suitable localized basis set is needed to get a sparse representation. There are different approaches to find the density matrix, among them are minimization methods and recursive or stochastic approaches.

3.1 Minimization Methods

One approach towards linear scaling is based on efficient minimization methods. They can be direct or iterative, but they all have the same goal to minimize the total energy with respect to the density matrix so that the ground state is found. A small selection of minimization methods are presented in the following subsections.

3.1.1 Orbital Minimization

The concept of orbital minimization has been developed by different research groups [70, 105, 120, 121]. The idea is to circumvent the computationally expensive orthonormalization of the orbitals. This can be done by not directly calculating the inverse overlap matrix but to replace it with its Neumann series up to a predefined order M. For a system with N electrons and $N/2$ orbitals

$\{|\psi_i\rangle\}$, we can calculate the energy by minimizing

$$E = 2\,\mathrm{tr}(QH) - \eta\,(2\,\mathrm{tr}(QS) - N)\,, \tag{3.1}$$

where

$$Q = \sum_{n=0}^{M}(I - S)^n \approx S^{-1}. \tag{3.2}$$

and η is a constant [17, 121]. The orbitals are expressed in a basis $\{|\phi_\mu\rangle\}$

$$|\psi_i\rangle = \sum_{\mu} C_{i\mu}|\phi_\mu\rangle, \tag{3.3}$$

and minimized with respect to the coefficients $C_{i\mu}$. If the orbitals are localized, linear scaling can be achieved. An issue of this method is that it suffers from local minima if a too small set of orbitals is used. Furthermore, the method suffers from fluctuations in the total energy when it is used in the context of self-consistent calculations [104]. There are two attempts to parallelize the orbital minimization method [19, 65], but both require a considerable amount of communication between the processors [41].

3.1.2 Penalty Method

Another method to find the ground-state density is based on the idea of using penalty functionals in total energy calculations. It has been developed by Walter Kohn [73]. He described the method in the context of the nearsightedness principle, which applies to the one-particle density matrix

$$\rho(r, r') = \sum_{j=1}^{N} \psi_j^*(r)\psi_j(r'), \tag{3.4}$$

where $\{\psi_j\}$ are the KS orbitals. The idea is to use penalty functionals in total energy calculations. One adds a functional $P[\cdot]$ of Hermitian trial functions $\tilde{\rho}(r, r')$ to the energy functional. The functional

$$P[\tilde{\rho}(r, r')] = \sqrt{\int \tilde{\rho}^2(r, r')(1 - \tilde{\rho}(r, r'))^2 \mathrm{d}r} > 0 \tag{3.5}$$

is composed in such a way that, when inserting the density matrix $\mathcal{P}$, that fulfils the idempotency condition $\mathcal{P}^2(r, r') = \mathcal{P}(r, r')$, it is equal to zero. Thus, $P[\cdot]$ is a measure for the idempotency of $\mathcal{P}$. We define the functional $Q[\tilde{\rho}(r, r')]$

for the ground state search by the help of the grand-canonical potential (GCP) $\Omega = E[\bar{\rho}(r)] - \mu N[\bar{\rho}(r)]$

$$Q[\tilde{\rho}(r,r')] = E[\bar{\rho}(r)] - \mu N[\bar{\rho}(r)] + \alpha P[\tilde{\rho}(r,r')]. \tag{3.6}$$

Thereby, $E[\bar{\rho}(r)]$ is the KS energy functional, μ the chemical potential, $N[\bar{\rho}(r)] = \int \bar{\rho}(r)\mathrm{d}r = \int \tilde{\rho}(r,r'))\mathrm{d}r$ corresponds to the number of electrons and $\alpha > 0$ is a parameter. The matrix $\bar{\rho}(r) = \tilde{\rho}^2(r,r')$ is non-negative with eigenvalues greater than zero. As the penalty functional $P[\cdot]$ is a summand in equation (3.6), a trial density, that is not the (idempotent) ground state density leads to a larger value of $Q[\cdot]$. The ground state can consequently be found by varying the density and seeking the lowest value of $Q[\cdot]$. Hereby, it is important to notice that $Q[\cdot]$ can, in contrast to the GCP, be minimized without constraints.
The determination of the best value of the parameter α is challenging, but if it is found, we can only get as a result the correct ground state density, and idempotency is achieved automatically without imposing the constraint in advance. The optimal α is chosen such that it is larger than a critical value α_c which can unfortunately not be predicted exactly. If $\alpha < \alpha_c$, one might fall into a local minimum, but if α is chosen too large, this leads to slow convergence.
One issue of the penalty functional method is the branch point of the square root, which is attained at its minimum. The functional is not analytic, and therefore the variational principle can not be exploited. Thus, this method can not be used in practice [49].
Instead, one can use the functional $P^2[\cdot]$, see [50]. Here, the idempotency condition is not exactly imposed, so that the energy has to be corrected. This is simple for occupied bands but challenging for unoccupied ones [16].

3.1.3 Density Minimization Method

In their work [90], Li, Nunes, and Vanderbilt introduced the idea of density matrix minimization (DMM). This is a linear scaling method that is based on a variational solution for the density matrix. The off-diagonal elements of the density matrix are truncated with respect to a cut-off parameter R_c. For $R_c \to \infty$, in the adopted level of theory, the method remains exact, but can no more have a linear scaling.

The idea of DMM is to minimize the energy E by the help of the density matrix $\mathcal{P}$ and the GCP for the electrons. We have

$$N_e = \mathrm{tr}[\mathcal{P}] = \sum_i \mathcal{P}_{ii}, \tag{3.7}$$

where N_e is the number of electrons of the system and

$$E = \mathrm{tr}[\mathcal{P}H] = \sum_{i,j} \mathcal{P}_{ij} H_{ji} \tag{3.8}$$

for a Hamiltonian matrix H. The density matrix $\mathcal{P}$ has to fulfil the idempotency condition $\mathcal{P} = \mathcal{P}^2$. This ensures that it is the matrix representation of the operator which projects onto the subspace of occupied states. Another condition that has to be satisfied is to have either constant electron number N_e or constant chemical potential μ. Li et al. state that it is more convenient to keep the chemical potential μ fixed and minimize the GCP

$$\Omega = E - \mu N_e = \mathrm{tr}[\mathcal{P}(H - \mu I)]. \tag{3.9}$$

A complication is that one has to make sure that the eigenvalues of $\mathcal{P}$ lie in the interval $[0, 1]$. To overcome this problem and to satisfy the idempotency condition, as well, they replace $\mathcal{P}$ by a trial density matrix

$$\tilde{\mathcal{P}} = 3\mathcal{P}^2 - 2\mathcal{P}^3, \tag{3.10}$$

which is known as the McWeeny purification transformation [106].
The density matrix is therefore obtained by minimizing

$$\Omega = \mathrm{tr}[\tilde{\mathcal{P}}(H - \mu I)] \tag{3.11}$$

with respect to $\mathcal{P}$. This ensures that, if the eigenvalues of $\mathcal{P}$ are either close to 0 or 1, the eigenvalues of $\tilde{\mathcal{P}}$ are even closer to those values.
As DMM approximates through the above-described cut-off radius R_c, the sparsified density matrix is no longer idempotent. The McWeeny transformation, however, reduces this error. Regarding the question of parallelizability, one notes that DMM is not intrinsically parallel. Thus, it would become less efficient as there is a considerable amount of all-to-all communication that has to be realized [15].

3.2 Fermi Operator Expansion

A completely different approach to linear scaling is the idea of Fermi operator expansion (FOE), which has been developed by Goedecker, Colombo, and Teter [40, 42, 43]. It is based on the expression of the density matrix $\mathcal{P}$ as a matrix functional $f(\cdot)$ of the Hamiltonian matrix H

$$\mathcal{P} = f(H), \tag{3.12}$$

where

$$f(\varepsilon) = \frac{1}{1 + \exp\left(\frac{\varepsilon - \mu}{k_B T}\right)}. \tag{3.13}$$

Thereby, ε is an eigenvalue of H, μ the chemical potential, k_B Boltzmann's constant, and T the electronic temperature. The functional $f(\cdot)$ is called the Fermi distribution. The idea of FOE consists of a series expansion of the Fermi operator, that is to say the Fermi distribution of the system's Hamiltonian, $f(H)$. The easiest way is to approximate with a polynomial $p(H)$ with coefficients a_i

$$\mathcal{P} \approx p(H) = \sum_{i=0}^{n} a_i H^i, \tag{3.14}$$

but this entails numerical instability for a high polynomial degree n [40]. This can be circumvented by using Chebyshev polynomials $T_i(X)$ for matrices, which are defined as follows

$$\begin{aligned} T_0(X) &= I \\ T_1(X) &= X \\ T_{i+1}(X) &= 2\,X T_i(X) - T_{i-1}(X) \quad \text{for } i = 2, 3, \ldots \end{aligned} \tag{3.15}$$

Thereby, X is a matrix whose spectrum lies in the interval $[-1, 1]$. The Chebyshev polynomials are orthogonal so that it holds

$$\langle T_i, T_j \rangle = \begin{cases} 0, & i \neq j \\ \pi, & i = j = 0 \\ \pi/2, & i = j \neq 0. \end{cases} \tag{3.16}$$

The set $\{T_i\}_{i=0}^{n}$ forms a basis of the linear vector space of all polynomials with degree $\leq n$. So, we can write our polynomial $p(\cdot)$ from (3.14) as a linear

combination of this basis with known coefficients c_i [39]

$$p(H) = \frac{c_0}{2} I + \sum_{i=1}^{n} c_i T_i(H). \tag{3.17}$$

Goedecker [40] shows that such an expansion of a matrix in Chebyshev polynomials scales quadratically with the number of basis functions and therefore with the number of atoms in the system. He further explains how a localization region for each column of the Hamiltonian can be introduced. So, a truncated Hamiltonian is used and linear scaling is achieved.
For our tight-binding scheme [57], where local charge neutrality is enforced, the degree n of the polynomial of the Chebyshev expansion must be higher to reach the same high accuracy as in the non-self-consistent tight-binding case. It holds

$$n \propto \frac{(\varepsilon_{\max} - \varepsilon_{\min})}{\Delta\varepsilon}, \tag{3.18}$$

where $\varepsilon_{\max}$ and $\varepsilon_{\min}$ is the largest and smallest eigenvalue, respectively, and $\Delta\varepsilon$ is the spectrum width of the Hamiltonian.

It remains to be answered why one should use the FOE method for the calculation of the density matrix. To answer this question, we look again at the work of Goedecker [40]. One clear advantage is that the generation of density matrix requires only matrix-vector multiplications and is therefore computationally very efficient. Additionally, only FOE scales linearly with respect to the size of the localization region. In the important 3-dimensional case, it has both, the best asymptotic scaling behaviour and the smallest prefactor with respect to the accuracy. There is no initial guess for the density matrix needed and it is intrinsically parallel so that a good speed-up can be achieved.

Chapter 4

Self-Consistent Field Theory Based Molecular Dynamics with Linear System-Size Scaling

In this chapter, we present a field-theoretic method suitable for linear scaling molecular dynamics simulations using forces from self-consistent electronic structure calculations. It is based on an exact decomposition of the grand-canonical potential for independent fermions and does neither rely on the ability to localize the orbitals nor that the Hamiltonian operator is well-conditioned. Hence, this scheme enables highly accurate all-electron linear scaling calculations even for metallic systems. The inherent energy drift of Born-Oppenheimer molecular dynamics simulations, arising from an incomplete convergence of the self-consistent field cycle, is solved by means of a properly modified Langevin equation. The predictive power of this approach is illustrated using the example of liquid methane under extreme conditions.

This work has been presented up to minor changes in a publication by Dorothee Richters and Thomas D. Kühne [136]. The sections 4.1-4.5 are the sections I-IV, VI of [136], section V of this publication is integrated into section 5.3, where we present the results of the application of our method to liquid methane.

4.1 Introduction

Ab initio molecular dynamics (AIMD), where the forces are calculated on-the-fly by accurate electronic structure methods, has been very successful in explaining and predicting a large variety of physical phenomena and guiding experimental work [102]. However, the increased accuracy and predictive power of AIMD simulations comes at a significant computational cost, which has limited the attainable length and time scales in spite of recent progress [62, 85]. As a consequence, Hartree-Fock (HF), density functional theory (DFT) [66, 74], and even the semi-empirical tight-binding (TB) approach [140, 143] are to date the most commonly used electronic structure methods in conjunction with AIMD. However, for large systems the calculation of the electronic structure and hence total energies as well as nuclear forces of atoms and molecules is still computationally fairly expensive. This is due to the fact that solving the Schrödinger equation is a high-dimensional eigenvalue problem, whose solution requires diagonalizing the Hamiltonian of the corresponding system, which typically scales cubically with its size. Therefore, a method that scales linearly with the size of the system would be very desirable, thus making a new class of systems accessible to AIMD that were previously thought not feasible. For that reason, developing such methods is an important objective and would have a major impact in scientific areas such as nanotechnology or biophysics, just to name a few.

Several so called linear scaling methods have been proposed [9, 16, 35, 41, 90, 105, 122, 153] to circumvent the cubic scaling diagonalization that is the main bottleneck of DFT and TB. Underlying all of these methods is the concept of "nearsightedness" [73, 130], an intrinsic system dependent property, which states that at fixed chemical potential the electronic density depends just locally on the external potential, so that all matrices required to compute the Fermi operator will become sparse at last. Together with sparse matrix algebra techniques linear scaling in terms of memory requirement and computational cost can be eventually achieved. However, the crossover point after which linear scaling methods become advantageous is still rather large, in particular for metallic systems or if high accuracy is needed.

Therefore another method, based on the grand-canonical potential (GCP) for independent fermions, has been recently developed [1, 2]. Krajewski and Parrinello demonstrated that by decomposing the GCP it is possible to devise an approximate stochastic linear scaling scheme [78–80]. Since this approach does not rely on the ability to localize the electronic wavefunction, even metals can be treated. However, due to its stochastic nature extending such a method towards self-consistent TB, DFT or HF is far from straightforward.

This is where we start in this work. Following previous work of Ceriotti, Kühne, and Parrinello [27, 28], we compute here the finite-temperature density matrix, or Fermi matrix, in an efficient, accurate, and in particular deterministic fashion by a hybrid approach. Inspired by the Fermi operator expansion method pioneered by Goedecker and coworkers [40–42], the Fermi operator is described in terms of a Chebyshev polynomial expansion, but in addition is accompanied by fast summation as well as iterative matrix inversion techniques. The resulting algorithm has several important advantages. As before, the presented scheme does not rely on the ability to localize the orbitals, but requires only that the Hamiltonian matrix is sparse, a substantially weaker requirement. As a consequence not only metals, but even systems for which the Fermi matrix is not sparse yet can be treated with a linear scaling computational effort. Another advantage is that the algorithm is intrinsically parallel as the terms resulting from the decomposition of the GCP are independent of each other and can be separately calculated on different processors.
But, at variance to the original approach [78–80], the addition of Chebyshev polynomial expansion and fast summation techniques leads to a particularly efficient algorithm that obeys a sub-linear scaling with respect to the width of the Hamiltonian's spectrum, which is very attractive for all-electron calculations or when a high energy resolution is required. Since the present method allows for an essentially exact decomposition of the GCP, without invoking any high-temperature approximation, it facilitates highly accurate linear scaling *ab initio* simulations. However, the main advantage lies in the deterministic nature of the hybrid approach, which enables self-consistent electronic structure calculations. The fact that the present scheme is based on the GCP inherently entails finite electron temperature, which is not only in line with finite temper-

ature simulations such as AIMD, but furthermore also allows for computations of systems with excited electrons [141, 142]. We have thus put particular emphasis on adopting the hybrid approach within AIMD. Specifically, the modified Car-Parrinello-like propagation of the self-consistent Hamiltonian matrix [85] and how to accurately sample the Boltzmann distribution with noisy forces [79, 85] are discussed in detail. Beside describing the method itself, we will show that it is indeed possible to perform fully self-consistent AIMD simulations and demonstrate the present scheme on liquid methane at planetary pressure and temperature conditions in Chapter 5.

4.2 Basic Methodology

In this section, we summarize the basic methodology, first proposed by Krajewski and Parrinello [78–80]. We begin with the generic expression for the total energy E of an effective single-particle theory, such as HF, DFT or TB

$$E = 2\sum_{i=1}^{N} \varepsilon_i + V_{dc}. \tag{4.1}$$

The first term denotes the so-called band-structure energy, which is given by the sum of the lowest N doubly occupied eigenvalues ε_i of an arbitrary Hamiltonian H. In DFT, for instance, H is the KS matrix, while V_{dc} accounts for double counting terms as well as for the nuclear Coulomb interaction. In TB and other semi-empirical theories, H depends parametrically only on the nuclear positions and V_{dc} is a pairwise additive repulsion energy. While in either case it is well known how to calculate V_{dc} with linear scaling computational effort, the computation of all occupied orbitals by diagonalization requires $\mathcal{O}(N^3)$ operations. Due to the fact that the band-structure term can be equivalently expressed in terms of the density matrix $\mathcal{P}$, the total energy can be written as

$$E = 2\sum_{i=1}^{N} \varepsilon_i + V_{dc} = \mathrm{tr}[\mathcal{P}H] + V_{dc}. \tag{4.2}$$

As a consequence, the cubic scaling diagonalization of H can be bypassed by directly calculating $\mathcal{P}$ rather than all ε_i's.

To that extend, we follow Alavi and coworkers [1, 2] and consider the following

(Helmholtz) free energy functional

$$\mathcal{F} = \Omega + \mu N_e + V_{dc}, \tag{4.3}$$

where μ is the chemical potential, $N_e = 2N$ the number of electrons and Ω the GCP for non-interacting fermions

$$\begin{aligned} \Omega &= -\frac{2}{\beta} \ln \det \left(I + e^{\beta(\mu S - H)} \right) \\ &= -\frac{2}{\beta} \operatorname{tr} \ln \left(I + e^{\beta(\mu S - H)} \right). \end{aligned} \tag{4.4}$$

Here, S stands for the overlap matrix, which is equivalent to the identity matrix I if and only if the orbitals are expanded in mutually orthonormal basis functions. In the GCP, the electronic temperature is finite and given by $\beta^{-1} = k_B T_e$. However, in the low-temperature limit

$$\lim_{\beta \to \infty} \Omega = 2 \sum_{i=1}^{N} \varepsilon_i - \mu N_e, \tag{4.5}$$

the band-structure energy can be recovered and $\lim_{\beta\to\infty} \mathcal{F} = E$ holds. In order to make further progress, let us now factorize the operator of equation (4.4) into P terms. Given that P is even, which we shall assume in the following, Krajewski and Parrinello [79, 80] derived the following identity

$$\begin{aligned} I + e^{\beta(\mu S - H)} &= \prod_{l=1}^{P} \left(I - e^{\frac{i\pi}{P}(2l-1)} e^{\frac{\beta}{P}(\mu S - H)} \right) \\ &= \prod_{l=1}^{P} M_l = \prod_{l=1}^{P/2} M_l^* M_l, \end{aligned} \tag{4.6}$$

where the matrices M_l, with $l = 1, \ldots, P$ are defined by

$$M_l := I - e^{\frac{i\pi}{P}(2l-1)} e^{\frac{\beta}{P}(\mu S - H)}, \tag{4.7}$$

and * denotes complex conjugation. Similar to numerical path-integral calculations, it is possible to exploit the fact that if P is large enough, so that the effective temperature β/P is small, the exponential operator $e^{\frac{\beta}{P}(\mu S - H)}$ can be approximated by a Trotter decomposition or simply by a high-temperature expansion, i.e.

$$M_l = I - e^{\frac{i\pi}{P}(2l-1)} \left(I + \frac{\beta}{P}(\mu S - H) \right) + \mathcal{O}\left(\frac{1}{P^2} \right). \tag{4.8}$$

However, as we will see, here no such approximation is required, which is in contrast to the original approach [78–80]. In any case, the GCP can be rewritten as

$$\begin{aligned}\Omega &= -\frac{2}{\beta}\ln\det\prod_{l=1}^{P} M_l = -\frac{2}{\beta}\ln\prod_{l=1}^{P/2}\det\left(M_l^* M_l\right)\\ &= -\frac{2}{\beta}\sum_{l=1}^{P/2}\ln\det\left(M_l^* M_l\right)\\ &= \frac{4}{\beta}\sum_{l=1}^{P/2}\ln\left(\det\left(M_l^* M_l\right)\right)^{-\frac{1}{2}}. \end{aligned} \tag{4.9}$$

As is customary in lattice gauge field theory [110, p. 17], where the minus sign problem is avoided by sampling a positive definite distribution, the inverse square root of the determinant can be written as an integral over a complex field $\boldsymbol{\phi}_l$, which has the same dimension M as the full Hilbert space, i.e.

$$\det\left(M_l^* M_l\right)^{-1/2} = \frac{1}{(2\pi)^{\frac{M}{2}}}\int e^{-\frac{1}{2}\phi_l^* M_l^* M_l \phi_l}\,\mathrm{d}\phi_l. \tag{4.10}$$

Inserting equation (4.10) into equation (4.9) we end up with the following field-theoretic expression for the GCP:

$$\begin{aligned}\Omega &= \frac{4}{\beta}\sum_{l=1}^{P/2}\ln\left[\frac{1}{(2\pi)^{\frac{M}{2}}}\int e^{-\frac{1}{2}\phi_l^* M_l^* M_l \phi_l}\,\mathrm{d}\phi_l\right]\\ &= \frac{4}{\beta}\sum_{l=1}^{P/2}\ln\int e^{-\frac{1}{2}\phi_l^* M_l^* M_l \phi_l}\,\mathrm{d}\phi_l + c, \end{aligned} \tag{4.11}$$

where ϕ_l are appropriate vectors and c is an additive constant.

All physical relevant observables can be computed as functional derivatives of the GCP with respect to an appropriately chosen external parameter. For example, $N_e = -\partial\Omega/\partial\mu$ and $\lim_{\beta\to\infty}\Omega + \mu N_e = 2\sum_{i=1}^{N}\varepsilon_i$, so that

$$E = \lim_{\beta\to\infty}\mathcal{F} = 2\sum_{i=1}^{N}\varepsilon_i + V_{dc} = \frac{\partial(\beta\Omega)}{\partial\beta} - \mu\frac{\partial\Omega}{\partial\mu} + V_{dc}. \tag{4.12}$$

Since the functional derivative of the constant in equation (4.11) is identical

to zero, all physical interesting quantities can be computed analogue to

$$\frac{\partial\Omega}{\partial\lambda} = \frac{4}{\beta}\sum_{l=1}^{P/2}\frac{\int -\frac{1}{2}\phi_l^*\left(\frac{\partial(M_l^*M_l)}{\partial\lambda}\right)\phi_l e^{-\frac{1}{2}\phi_l^*M_l^*M_l\phi_l}\,\mathrm{d}\phi_l}{\int e^{-\frac{1}{2}\phi_l^*M_l^*M_l\phi_l}\,\mathrm{d}\phi_l} \tag{4.13}$$

$$= -\frac{2}{\beta}\sum_{l=1}^{P/2}\frac{\int\sum_{i,j=1}^{M}(\phi_l^*)_i\left(\frac{\partial(M_l^*M_l)}{\partial\lambda}\right)_{ij}(\phi_l)_j e^{-\frac{1}{2}\phi_l^*M_l^*M_l\phi_l}\,\mathrm{d}\phi_l}{\int e^{-\frac{1}{2}\phi_l^*M_l^*M_l\phi_l}\,\mathrm{d}\phi_l}$$

$$= -\frac{2}{\beta}\sum_{l=1}^{P/2}\sum_{i,j=1}^{M}\left(\frac{\partial(M_l^*M_l)}{\partial\lambda}\right)_{ij}\frac{\int(\phi_l^*)_i(\phi_l)_j e^{-\frac{1}{2}\phi_l^*M_l^*M_l\phi_l}\,\mathrm{d}\phi_l}{\int e^{-\frac{1}{2}\phi_l^*M_l^*M_l\phi_l}\,\mathrm{d}\phi_l}$$

$$= -\frac{2}{\beta}\sum_{l=1}^{P/2}\sum_{i,j=1}^{M}\left(\frac{\partial(M_l^*M_l)}{\partial\lambda}\right)_{ij}(M_l^*M_l)^{-1}_{ij} \tag{4.14}$$

$$= -\frac{2}{\beta}\sum_{l=1}^{P/2}\mathrm{tr}\left[(M_l^*M_l)^{-1}\frac{\partial(M_l^*M_l)}{\partial\lambda}\right] \tag{4.15}$$

$$= -\frac{2}{\beta}\sum_{l=1}^{P}\mathrm{tr}\left[M_l^{-1}\frac{\partial M_l}{\partial\lambda}\right]. \tag{4.16}$$

Thereby, equation (4.14) holds because of Montvay and Münster [110, p. 18], while equation (4.15) is due to the fact that $M_l^*M_l$ is symmetric positive definite.

Unlike equation (4.9), the determination of $\Omega = \partial(\beta\Omega)/\partial\beta$ does no longer require to calculate the inverse square root of a determinant, but only the inverse of M_l. But, since the inversion usually has to be performed P times, the computational scaling has presumably a rather large prefactor. Nevertheless, as we will see later this can be much ameliorated and all but very few matrix inversions can be avoided. On the other hand, M_l is not only very sparse, since it obeys the same sparsity pattern as H, but is furthermore also always better conditioned as the latter, so that all M_l^{-1} matrices are substantially sparser than the finite temperature density matrix and thus can be efficiently determined [27, 28]. Solving the N_e sets of linear equations $M_l\boldsymbol{\Phi}_j^l = \boldsymbol{\psi}_j$, where $\{\boldsymbol{\psi}_j\}$ is a complete set of basis functions, the inverse can be exactly computed as $M_l^{-1} = \sum_{j=1}^{N_e}\boldsymbol{\phi}_j^l\boldsymbol{\psi}_j^l$ within $\mathcal{O}(N^2)$ operations.

Comparing equation (4.2) with equation (4.5) it is easy to see that the GCP and similarly all physical significant observables can be written as the trace of a matrix product consisting of the Fermi matrix ρ, which in the low-temperature limit is equivalent to $\mathcal{P}$. Specifically, $\Omega = \partial(\beta\Omega)/\partial\beta = \mathrm{tr}[\rho H] - \mu N_e$, but because at the same time $N_e = \mathrm{tr}[\rho S]$ holds, the former can be simplified to

$$\Omega = \mathrm{tr}[\rho(H - \mu S)], \tag{4.17}$$

where $S = -\partial H/\partial\mu$ and $\rho = \partial\Omega/\partial H$. As a consequence, the GCP and all its functional derivatives can be reduced to evaluate ρ based on equation (4.16) with $\lambda = H_{ij}$. Using the standard basis $\{e_i\}$ and the identity [80]

$$\begin{aligned}\rho_{ij} = \frac{\partial\Omega}{\partial H_{ij}} &= \frac{2}{P}\sum_{l=1}^{P} e_j^* \left(I - M_l^{-1}\right) e_i = \frac{2}{P}\left(\sum_{l=1}^{P}\left(I - M_l^{-1}\right)\right)_{ji}, \\ &= \frac{4}{P}\sum_{l=1}^{P/2}\left(I - \left(M_l^* M_l\right)^{-1}\right)_{ji} = \frac{4}{P}\sum_{l=1}^{P/2}\left(I - \left(M_l^* M_l\right)^{-1}\right)_{ij},\end{aligned} \tag{4.18}$$

where the latter holds as the inverse of $\left(M_l^* M_l\right)^{-1}$ is symmetric, as well, we get

$$\begin{aligned}\rho = \frac{\partial\Omega}{\partial H} &= \frac{4}{P}\sum_{l=1}^{P/2}\left(I - \left(M_l^* M_l\right)^{-1}\right) \\ &= \frac{2}{P}\sum_{l=1}^{P}\left(I - M_l^{-1}\right).\end{aligned} \tag{4.19}$$

In other words, the origin of the method is the notion that the density matrix, the square of the wavefunction at low temperature and the Maxwell-Boltzmann distribution at high temperature, can be decomposed into a sum of M_l^{-1} matrices, each at higher effective temperature β/P and hence always sparser than ρ. Yet, contrary to the original approach [78–80], neither a Trotter decomposition nor a high-temperature expansion for equation (4.7) has been used, so far everything is exact for any P. Nevertheless, beside the aforementioned reduction from cubic to quadratic scaling no computational savings have been gained either. Quite the contrary, at first sight it might even appear that this scheme, which requires to invert P matrices, is less efficient than explicitly diagonalizing H. However, as already mentioned, in the next section we are

going to demonstrate that this can be circumvented for the most part by expressing all but very few matrix inversions through a Chebyshev polynomial expansion.

4.3 The Hybrid Approach

In this section, we describe the novel hybrid approach in order to make further progress and to achieve an even more favourable scaling. For that purpose, one can either approximate the propagator $e^{\frac{\beta}{P}(\mu S - H)}$ of equation (4.7), or exploit the fact that by increasing P in equation (4.19) the matrix exponential and hence M_l^{-1} can be ever simpler exactly calculated. Specifically, we employed the squaring and scaling technique to compute matrix exponentials, i.e. $e^A = (e^{A/m})^m$ [109], where we exploit the fact that $e^{A/m}$ is trivial to compute whenever m is large. In an analysis of the M_l matrices, we found that every M_l matrix is throughout better conditioned than H [27]. From this follows that for all l, M_l^{-1} always exhibits less non-zero entries and is therefore much easier to compute than the inverse of H, which would correspond to the complexity of calculating ρ directly.
In addition, the method can be even more improved by recognizing that H is real as well as symmetric and that the equality

$$M_l = M_{P-l+1}^* \tag{4.20}$$

holds. Using equation (4.20) and the fact that $\omega_l := e^{\frac{i\pi}{P}(2l-1)}$ denotes a point on the unit circle of the complex plane, we show that only the real parts of the M_l matrices are required to compute ρ

$$\begin{aligned}
M_l^* M_l &= \left(I - \omega_l e^{\frac{\beta}{P}(\mu S - H)}\right)^* \left(I - \omega_l e^{\frac{\beta}{P}(\mu S - H)}\right) \\
&= \left(I - \overline{\omega}_l \left(e^{\frac{\beta}{P}(\mu S - H)}\right)^*\right) \left(I - \omega_l e^{\frac{\beta}{P}(\mu S - H)}\right) \\
&= I - (\overline{\omega}_l + \omega_l) e^{\frac{\beta}{P}(\mu S - H)} + (\overline{\omega}_l \omega_l)\, e^{\frac{2\beta}{P}(\mu S - H)} \\
&= I - 2\,\mathrm{Re}\,\omega_l\, e^{\frac{\beta}{P}(\mu S - H)} + e^{\frac{2\beta}{P}(\mu S - H)} \\
&= \left(I + e^{\frac{2\beta}{P}(H - \mu S)} - 2\,\mathrm{Re}\,\omega_l\, e^{\frac{\beta}{P}(H - \mu S)}\right) e^{\frac{2\beta}{P}(\mu S - H)} \\
&=: N_l\, e^{\frac{2\beta}{P}(\mu S - H)} \in \mathbb{R}^{M \times M},
\end{aligned} \tag{4.21}$$

where M is the number of basis functions and therefore the dimension of the real matrix $M_l^* M_l$.
The latter proof entails substantial savings in terms of computational cost and memory requirement. From this it follows that equation (4.19) can be further simplified to

$$\rho = \frac{2}{P} \sum_{l=1}^{P/2} \left(I - \mathrm{Re}\, M_l^{-1} \right), \tag{4.22}$$

where the upper limit of index l is henceforth restricted to $P/2$. Moreover, it has been observed that just a handful of M_l matrices, where l is close to $P/2$, are ill-conditioned and only for them the inversion is computationally cumbersome. All other M_l matrices having a smaller index are rather well-conditioned, so that the matrix inversion can be very efficiently performed by a Chebyshev polynomial expansion [27]. This is to say that ρ can always be written as a sum of M_l^{-1} matrices, which are throughout pretty much sparser than ρ itself. The latter is in fact true, even if ρ is rather full, so that metallic systems can be very efficiently treated.

These complementary properties of the M_l matrices immediately suggest the following hybrid approach. Thereby, an optimal $\bar{l}$ is chosen such that $1 < \bar{l} < P/2$, where all M_l matrices with $l < \bar{l}$ are inverted by a Chebyshev polynomial expansion and only otherwise for $l \geq \bar{l}$ by an iterative Newton-Schulz matrix inversion. As long as M_l is not ill-conditioned, the former has the advantage of being essentially independent of P, so that increasing P will not increase the computational cost. Together with the fact that the number of ill-conditioned M_l matrices depends only on the particular system and β, but again not on P, the present hybrid approach allows to employ an arbitrary large P at basically no additional computational cost. In this way, the decomposition of the GCP in equation (4.9) can be made exact in any order essentially for free. From this it follows that the electronic temperature β^{-1} can be chosen to be rather low and is typically identical with the nuclear temperature.
Furthermore, it is possible to rewrite $\mathrm{Re}\, M_l^{-1}$ in the following way:

$$\mathrm{Re}\, M_l^{-1} = \frac{1}{2} \left(I + \left(e^{\frac{2\beta}{P}(H-\mu S)} - I \right) N_l^{-1} \right), \tag{4.23}$$

where N_l is the real valued matrix defined in equation (4.21). That is to say, the whole problem can be reduced to invert N_l. Pretty much as for the M_l matrix, if N_l is well-conditioned, its inverse can be expressed by a Chebyshev expansion. For this purpose let us rewrite N_l in terms of a shifted and scaled auxiliary matrix

$$X = \frac{e^{\frac{\beta}{P}(H-\mu S)} - z_0}{\zeta}, \tag{4.24}$$

whose spectrum lies between -1 and 1. The corresponding shifting and scaling parameters $z_0 = \left(e^{\varepsilon_{\max}/P} + e^{\varepsilon_{\min}/P}\right)/2$ and $\zeta = \left(e^{\varepsilon_{\max}/P} - e^{\varepsilon_{\min}/P}\right)/2$ are expressed in terms of the maximum and minimum eigenvalues of $\tilde{H} = H - \mu S$ in unit of $k_B T$, i.e. by $\varepsilon_{\max}$ and $\varepsilon_{\min}$ [28]. Since a rather crude estimate for $\varepsilon_{\max}$ and $\varepsilon_{\min}$ is sufficient, they can be efficiently approximated using Gershgorin's circle theorem [37] as

$$\varepsilon_{\max} \geq \max_i \left(\tilde{H}_{ii} + \sum_{i \neq j} |\tilde{H}_{ij}| \right), \tag{4.25}$$

$$\varepsilon_{\min} \leq \min_i \left(\tilde{H}_{ii} - \sum_{i \neq j} |\tilde{H}_{ij}| \right). \tag{4.26}$$

The difference $\Delta\varepsilon = \varepsilon_{\max} - \varepsilon_{\min}$ corresponds to the spectral width of $\tilde{H}$. The condition number $\kappa(N_l) \approx 1 + \Delta\varepsilon^2 \pi^{-2} (P - 2l)^{-2}$ is somewhat higher than $\kappa(M_l) \approx 1 + \Delta\varepsilon \pi^{-1} (P - 2l)^{-1}$, but is more rapidly declining with decreasing l. Therewith, for $l < \bar{l}$, we can approximate N_l^{-1} as a sum of Chebyshev polynomials of X by

$$N_l^{-1} \approx \sum_{i=0}^{m_C(l)} c_i(l) T_i(X), \tag{4.27}$$

where $T_i(X)$ are the Chebyshev polynomials as defined in equation (3.15) and $c_i(l)$ the corresponding coefficients. Note that $T_i(X)$ are independent of l so that they have to be calculated only once. The upper bound $m_C(l)$ and thus the number of terms in the summation to achieve a relative accuracy of 10^{-D} on N_l^{-1} is approximately

$$m_C(l) \approx \frac{1}{2} + \frac{\Delta\varepsilon D \ln 10}{\pi (P - 2l)}. \tag{4.28}$$

After having computed the inverse of all the well-conditioned N_l matrices, we have to deal with the very few ill-conditioned ones. As already indicated this is accomplished by the following Newton-Schulz iteration

$$A_{k+1} = 2A_k - A_k N_l A_k, \quad k = 0, 1, \ldots, \tag{4.29}$$

which converges quadratically to N_l^{-1} given that A_0 is within the respective area of convergence [139]. Even though for

$$A_0 = N_l^* \big(\|N_l\|_1 \|N_l\|_\infty\big)^{-1}, \tag{4.30}$$

equation (4.29) is already guaranteed to converge [123], but the computation of N_l^{-1} becomes even more efficient with the availability of a good initial guess for the matrix inverse. Fortunately, we can make use of $N_{\bar{l}+n-1}^{-1}$ as an initial guess for $N_{\bar{l}+n}^{-1}$, $n \in \{0, \ldots, P/2 - \bar{l}\}$ that is good enough to even converge rather ill-conditioned matrices usually within a few iterations. The number of matrix multiplications required to obtain a relative accuracy of 10^{-D} on N_l starting from $N_{\bar{l}-1}$ that has already been calculated by equation (4.27) is

$$m_N(l) = \frac{2}{\ln 2} \ln \frac{\ln(1 - \chi(l)) - D \ln 10}{\ln \chi(l)}, \tag{4.31}$$

$$\text{where } \chi(l) \approx \frac{4(P + 1 - 2l)}{(1 + (P + 1 - 2l))^2}. \tag{4.32}$$

Hence, the optimal value of $\bar{l}$ can be found by minimizing the estimated total number of matrix multiplications

$$m_{tot}(\bar{l}) = m_C(\bar{l}) + \sum_{l=\bar{l}}^{P/2} m_N(l) \tag{4.33}$$

under variation of $\bar{l}$. In general, $P/2 - \bar{l}$ is rather small and only weakly dependent on β, which implies that just a few N_l matrices needs to be explicitly inverted using equation (4.29), regardless of the electronic temperature.

However, the matrix-matrix multiplications of equation (4.29) causes that $\lim_{k\to\infty} A_{k+1}$ eventually becomes fairly occupied. For this reason in order to sustain linear scaling the intermediate matrices are truncated. Nevertheless, as already mentioned, the condition number of N_l is always lower than the one

of H and typically even rather well-conditioned, so that N_l^{-1} is by definition substantially sparser than ρ. From this it follows that the necessary truncation cut-off is relatively mild and the approximation therefore very small, so that highly accurate linear scaling electronic structure calculations are still possible.

4.4 Performing Self-Consistent Molecular Dynamics Simulations

In this section, we describe the implementation of the above presented hybrid approach within a self-consistent AIMD framework. Its chief advantage is not only that it allows for accurate linear scaling calculations, but is furthermore also deterministic. Hence, at variance to the original approach [78–80], where the corresponding matrices are inverted by an approximate stochastic method, it is now possible to perform calculations using Hamiltonian operators of fully self-consistent mean-field theories, such as HF, DFT, and self-consistent TB (SCTB) [32, 146].

We have tested the method in the context of electronic structure based MD using a SCTB model [57] and implemented it in the CMPTool program package [59, 107]. In the self-consistent field (SCF) optimization loop, self-consistency is realized by imposing local charge neutrality (LCN), to account for charge transfer processes, as well as bond breaking and formation. This means that the number of electrons of every atom α has to be equal to the number of its valence electrons q_α^0 within an adjustable tolerance, which we named $\Delta q_{\max}$. To that extend, during the SCF loop the diagonal elements of H are varied using a linear response function Θ until local charge neutrality is achieved. Specifically, in each MD step first H is built up, whereas in every SCF iteration we calculate the shift-vector Δ_H to the diagonal elements of H. The latter are the so called on-site energies $\epsilon_i = H_{ii}$, while the diagonal elements of ρS represents the occupancy of the corresponding orbital, hence $N_e = \mathrm{tr}[\rho S]$. Summing over all orbitals centred on any particular atom α, one obtains the associated on-site charge q_α. LCN is enforced by calculating $\Delta_H^k = \Theta(q_\alpha^k - q_\alpha^0)$ for every SCF iteration k and shifting the on-site energies using $\epsilon_i^{k+1} = \epsilon_i + \Delta_H^k$. So adapted, H^k is diagonalized using the above formalism until $\max_\alpha |q_\alpha - q_\alpha^0| \leq \Delta q_{\max}$. In that case, instead of being grand-canonical the simulation is performed at

constant N_e.

However, as already recognized by Kress et al. [82] using the present SCTB model [57], the SCF cycle is very slowly converging and the number of necessary iterations depends critically on $\Delta q_{\max}$. Nevertheless, this can remedied by adapting the method of Kühne et al. [85] in such a way that instead of a fully coupled electron-ion MD only the modified predictor-corrector integrator is used to propagate Δ_H in time. In the framework of DFT, this scheme has shown to be particularly effective for a large variety of different systems [22–25, 30, 33, 96, 97, 99, 154]. Inspired by the original scheme of Kolafa [76, 77] here

$$\Delta_H(t_n)^p = \sum_{m=1}^{K}(-1)^{m+1}m\frac{\binom{2K}{K-m}}{\binom{2K-2}{K-1}}\Delta_H(t_{n-m}) \tag{4.34}$$

is used as a modified predictor, where $\Delta_H(t_n)^p$ is an estimate for $\Delta_H(t_n)$ of the next MD time step t_n and is approximated using the weighted shifts of the K previous time steps. We claim that the weights

$$w_m := (-1)^{m+1}m\frac{\binom{2K}{K-m}}{\binom{2K-2}{K-1}} \tag{4.35}$$

always add up to 1. Thus, we show the following

$$\sum_{m=1}^{K} w_m = \sum_{m=1}^{K}(-1)^{m+1}m\frac{\binom{2K}{K-m}}{\binom{2K-2}{K-1}} = 1 \tag{4.36}$$

by making use of the Appendix of [76] and write

$$\sum_{m=1}^{K}(-1)^{m+1}m\binom{2K}{K-m} \tag{4.37}$$

$$= \sum_{m=1}^{K}(-1)^{m+1}m\left[\binom{2K-2}{K-m} + 2\binom{2K-2}{K-m-1} + \binom{2K-2}{K-m-2}\right]$$

$$= \sum_{m=1}^{K}(-1)^{m+1}\left[\frac{m(2K-2)!}{(K-m)!(K+m-2)!}\right.$$

$$\left. + \frac{2m(2K-2)!}{(K-m-1)!(K+m-1)!} + \frac{m(2K-2)!}{(K-m-2)!(K+m)!}\right]. \tag{4.38}$$

All but the first summand cancels out so that we get

$$\sum_{m=1}^{K}(-1)^{m+1}m\binom{2K}{K-m} = \binom{2(K-1)}{K-1}, \tag{4.39}$$

which after inserting it into equation (4.36) equals to

$$\sum_{m=1}^{K} w_m = \binom{2K-2}{K-1}^{-1} \cdot \binom{2(K-1)}{K-1} = 1. \tag{4.40}$$

As special case for $K = 1$, we have $w_1 = 1$, i.e. $\Delta_H(t_n)^p = \Delta_H(t_{n-1})$.

However, contrary to the second generation Car-Parrinello MD approach of Kühne et al. [85], where in each MD step only a single preconditioned electronic gradient calculation is required as the corrector, here the predicted $\Delta_H(t_n)^p$ is only used as an initial guess for the SCF cycle, which requires at least a single if not multiple diagonalizations. That is to say that instead of a genuine Car-Parrinello-like dynamics [21, 85], a less efficient accelerated Born-Oppenheimer MD (BOMD) [3, 6, 52, 119, 132, 149] is performed.
Nevertheless, in this way the convergence rate of the SCF cycle is much increased, while at the same time even allowing for a rather tight tolerance threshold. In fact, comparing with the employed convergence criterion of Kress et al. [82], here $\Delta q_{\max}$ can be chosen to be at least one to two orders of magnitude smaller without requiring numerous SCF iterations.

Due to the fact that the present scheme is equivalent to diagonalizing H, as for any SCF theory based BOMD simulation, the interatomic forces thus calculated are affected by a statistical noise Ξ_I^N, except for the unrealistic case that $\Delta q_{\max} = 0$. Hence, instead of the exact forces F_I, merely an approximation $F_I^{\mathrm{BOMD}} = F_I + \Xi_I^N$ is computed, where F_I^{BOMD} are the BOMD forces calculated by an arbitrary SCF based theory. Even though, Ξ_I^N can, to a very good approximation, be assumed as white [31, 62, 79], the line integral defining the net work is always positive and thus entails an energy drift during a microcanonical MD simulation. While the noise may be tiny and the forces highly accurate, as far as static calculations such as geometry optimization are concerned, the resulting energy drift is way more critical. An energy drift of as small as 1 μeV/atom/ps grows to an aberration of 10 K/ns and may cause

that liquid water, for instance, evaporates within a couple of nanoseconds simply because of the energy drift immanently present in any BOMD simulation [69, 84, 86, 87, 125]. Therefore, at least in principle, it is no longer guaranteed that by solving Newton's equation of motion the correct Boltzmann averages are obtained.

Fortunately, only based on the assumption that Ξ_I^N is unbiased, this can be rigorously corrected by devising a modified Langevin equation [79, 85]. Specifically, taking cue from the work of Krajewski and Parrinello [78, 79], we sample the canonical distribution using the following equation

$$M_I \ddot{R}_I = F_I + \Xi_I^N - \gamma_N M_I \dot{R}_I \tag{4.41}$$

$$= F_I^{\mathrm{BOMD}} - \gamma_N M_I \dot{R}_I, \tag{4.42}$$

where R_I are the positions of the nuclei, M_I their nuclear masses and γ_N a friction coefficient to compensate for the noise Ξ_I^N. The latter has to obey

$$\langle F_I(0) \Xi_I^N(t) \rangle \cong 0, \tag{4.43}$$

as well as the so called fluctuation-dissipation theorem

$$\langle \Xi_I^N(0) \Xi_I^N(t) \rangle = 2\gamma_N k_B T M_I \delta(t). \tag{4.44}$$

If we would know γ_N such that equation (4.44) is satisfied, a genuine Langevin equation is recovered, which guarantees for an accurate canonical sampling of the Boltzmann distribution. However, at first sight this may look like an impossible undertaking, since we neither know F_I, nor Ξ_I^N from which γ_N can be deduced. Nevertheless, it is possible, even without knowing Ξ_I^N except that it is approximately unbiased, to determine γ_N directly by simply varying it in such a way that the equipartition theorem $\left\langle \frac{1}{2} M_I \dot{R}_I^2 \right\rangle = \frac{3}{2} k_B T$ holds. Once γ_N is determined, it must be kept constant for the whole simulation. But then it is possible to exactly and very efficiently calculate static and even dynamic observables without knowing F_I, but just F_I^{BOMD}. Due to the fact that the same also holds for the noise introduced by truncating the intermediate matrices of equation (4.27), as well as using finite-precision arithmetic and a non-vanishing integration time step, the corresponding noise terms can be simply added to Ξ_I^N.

4.5 Conclusion

In conclusion, we would like to mention that the here presented method can be directly applied to fully self-consistent DFT calculations by writing V_{dc} of equation (4.1) as

$$\begin{aligned} V_{dc}[\rho(r)] = &- \frac{1}{2} \iint \frac{\rho(r)\rho(r')}{|r-r'|} \, \mathrm{d}r'\mathrm{d}r \\ &- \int \rho(r) \frac{\delta \Omega_{\mathrm{xc}}}{\delta \rho(r)} \, \mathrm{d}r + \Omega_{\mathrm{xc}} + E_{\mathrm{ii}}, \end{aligned} \tag{4.45}$$

where the first term on the right hand side is the double counting correction of the Hartree energy, while Ω_{xc} is the finite-temperature exchange and correlation grand-canonical functional and E_{ii} the nuclear Coulomb interaction. Except for the latter term, equation (4.45) accounts for the difference between the GCP for independent fermions Ω and the GCP for the interacting spin-$\frac{1}{2}$ Fermi gas [2]

$$\begin{aligned} \Omega_{int}[\rho(r)] = &- \frac{2}{\beta} \ln \det \left(I + e^{\beta(\mu S - H)} \right) \\ &- \frac{1}{2} \iint \frac{\rho(r)\rho(r')}{|r-r'|} \, \mathrm{d}r'\mathrm{d}r - \int \rho(r) \frac{\delta \Omega_{\mathrm{xc}}}{\delta \rho(r)} \, \mathrm{d}r + \Omega_{\mathrm{xc}}. \end{aligned} \tag{4.46}$$

As before, in the low-temperature limit $\Omega_{int}[\rho(r)] + \mu N_e$ equals to the band-structure energy, whereas Ω_{xc} corresponds to the familiar exchange and correlation energy, so that in this limit $\mathcal{F} = \Omega + \mu N_e + V_{dc} = \Omega_{int}[\rho(r)] + \mu N_e + E_{\mathrm{ii}}$ is equivalent to the Harris-Foulkes energy functional [34, 48]. Such as the latter, $\mathcal{F}$ is explicitly defined for any $\rho(r)$ and obeys exactly the same stationary point as the finite-temperature functional of Mermin [108].

The formal analogy of the decomposition to the Trotter factorization immediately suggests the possibility to apply some of the here presented ideas with benefit to numerical path-integral calculations [26]. The same applies for a related area where these methods are extensively used, namely the lattice gauge theory to quantum chromodynamics [71], whose action is rather similar to the one of equation (4.10).

Chapter 5

Liquid Methane at Extreme Temperature and Pressure: Implications for Models of Uranus and Neptune

In this chapter, we present a study on liquid methane (CH_4) at extreme conditions, meaning high pressure and temperature. Methane occurs in the middle ice layer of the giant gas planets Uranus and Neptune. In this layer, at a depth of one-third of the planetary radius, pressure and temperature range from 20 GPa and 2000 K to 600 GPa and 8000 K, which we simulate by means of large-scale electronic structure based molecular dynamics. In doing so, we employ the method described in Chapter 4 to illustrate its predictive power. We address the controversy of whether or not the interior of Uranus and Neptune consists of diamond. We show that there is no evidence for the formation of diamond, but rather carbon chains and sp^2-bonded polymeric carbon. We predict that at high temperature hydrogen may exist in its mono-atomic and metallic state.

This work has been presented up to minor changes in a publication by Dorothee Richters and Thomas D. Kühne [135]. We merge this with section V of [136], where results on methane have been presented, as well.

5.1 Introduction

Being the most abundant organic molecule in the universe, liquid CH_4 at high temperature and pressure is of great relevance for planetary science. The here considered pressure and temperature conditions follow the isentrope in the middle ice layers of Neptune and Uranus at a depth of one-third the planetary radius below the atmosphere. The gravity fields and mean densities of the outer gas giants Neptune and Uranus allude to a three-layer model: a relatively small central rocky core composed of iron, oxygen, magnesium and silicon, followed by an ice mantle and a predominantly hydrogen atmosphere. The middle ice layer consists of CH_4, NH_3 as well as H_2O and, in spite of its name, is not solid but gaseous in the outer atmospheres and a hot liquid in the interior [129]. At variance to the planetary models of Saturn and Jupiter, the observed values for mass and radius indicate that hydrogen cannot be an integral part of either Neptune and Uranus. Since it is moreover not primordial, the detected abundance of hydrogen in the atmospheres of both planets implies that it may initially originate from deep within the planets and brought to the outmost layer by convection, where it does not substantially contribute to the total mass [60]. It has to be mentioned that there is an uncertainty on the relative masses of the ice layer with respect to the rocky core [128]. However, we do not rely on any of these planetary models but only on the occurrence of a sizeable amount of ammonia, which most planetary models have in common [60, 128, 129].

In any case, information on the interior structure of Neptune and Uranus are scarce and experimentally only indirectly accessible by means of Voyager II flyby measurements [51, 61, 112], shock-wave compression [113, 114], as well as laser-heated diamond anvil cell experiments [11]. Even though CH_4 is the most stable hydrocarbon at ambient conditions, based on these shock-wave experiments as well as theoretical ground state calculations [133], it has been suggested that CH_4 may dissociate around $P = 20$ GPa and $T = 2000$ K into H_2 and diamond [137]. While there is little doubt that in the cores of Uranus and Neptune CH_4 dissociates into diamond, this would be anyhow rather consequential as it implies that in the interiors of these giant planets there is no

CH_4 at all, but a huge diamond mine instead.

On the other hand, *ab initio* molecular dynamics (AIMD) simulations predicted that the formation of diamond is preempted by the appearance of hydrocarbons [5]. Notwithstanding that their finding had been subsequently confirmed by laser-heated diamond anvil cell experiments, which, at pressure $P = 19$ GPa and temperature $T = 2000$-3000 K, indicate the presence of both polymeric carbon as well as diamond [11]. This view was further strengthened by subsequent AIMD simulations, even though none of them found any evidence for diamond formation [81, 82, 145]. Nevertheless, AIMD simulations are particularly appropriate to directly probe CH_4 under the extreme pressure and temperature conditions predominating in the middle ice layer, in particular as here in either case covalent bonds are broken and formed. Moreover, all of the AIMD simulations show that the intricate interplay between temperature and pressure is essential to grasp CH_4 at planetary conditions, where covalent C-H bonds are broken by heat, while compression favours condensation of the dissociated carbon atoms. It is therefore suggestive that in AIMD simulations at even higher pressure, but still in the middle ice layer, carbon may nonetheless spontaneously transform into diamond. However, what causes the large discrepancy in the pressure between theory and experiment when diamond is formed is unknown.

5.2 The Setting

In this and the following section, we revisit the behaviour of liquid CH_4 by means of the field-theoretic approach presented in Chapter 4. However, contrary to previous AIMD simulations, where the considered systems sizes have been rather small (16-128 CH_4 molecules) [5, 82, 145], here we use as many as 1000 CH_4 molecules in a periodic cubic simulation box of length $L = 25.55\,\text{Å}$ as our unit cell. All of our calculations have been performed in the canonical ensemble at $T = 2000$ - 8000 K and volume $V = 10.04$ cm^3/mol, which corresponds to the second-shock at $P = 92$ GPa and $T = 4000$ K of a two-stage light-gun shock compression experiment [114]. The agreement with the time-averaged pressure $\langle P \rangle = 72$ GPa of our AIMD simulation at $T = 4000$ K, as

calculated using the Nielsen-Martin stress-theorem [118], is more than satisfactory.
For all of our large scale MD simulations we have employed the self-consistent tight-binding model for hydrocarbons [57] as implemented in the CMPTool code [59]. The atoms are propagated by integrating the equation of motion of the modified Langevin equation (4.42) using a discretized time step of $\Delta t = 0.5$ fs [85, 86]. The LCN threshold of the SCF loop is chosen $\Delta q_{\max} = 0.05$.

Well-equilibrated and long trajectories are essential to ensure an accurate sampling. To that extend we have at first carefully equilibrated each of our simulations at $T = 2000, 4000, 6000$, and 8000 K, before accumulating statistics for overall 50 ps. Even though dissociation processes typically happen on rather short timescales, it is important to note that the temperature for dissociation and dehydrogenation as determined by direct MD simulations only represents an upper bound. A major advantage of our novel grand-canonical simulation technique is that, at variance to conventional ground-state AIMD simulations, excited electrons can be employed. Due to the fact that they are known to dramatically weaken covalent bonds [141, 142], and therefore may facilitate the dissociation of CH_4, we have hence chosen the electronic temperature to be identical with the nuclear temperature. Nuclear quantum effects, such as zero-point energies, are less important for the high temperature regime examined here and are therefore neglected. However, entropy effects have been shown to be very relevant, so that the dissociation of CH_4 is supposedly much more sensitive to temperature than it is to pressure [145].

By following the approach described in Chapter 4, the GCP for the electrons (4.19) had been decomposed using $P = 10000$. The minimization of equation (4.33) with respect to $\bar{l}$ yields $\bar{l} = P/2 - 2$, which implies that all except for two N_l matrices can be efficiently computed by a Chebyshev polynomial expansion with an estimated $m_C(\bar{l}) \leq 61$. Nevertheless, since equation (4.33) is merely an approximation, in practice the overall efficiency can be further increased by reducing $\bar{l}$. Here we have employed $\bar{l} = P/2 - 4$, which results in $m_C(\bar{l}) \approx 30$.

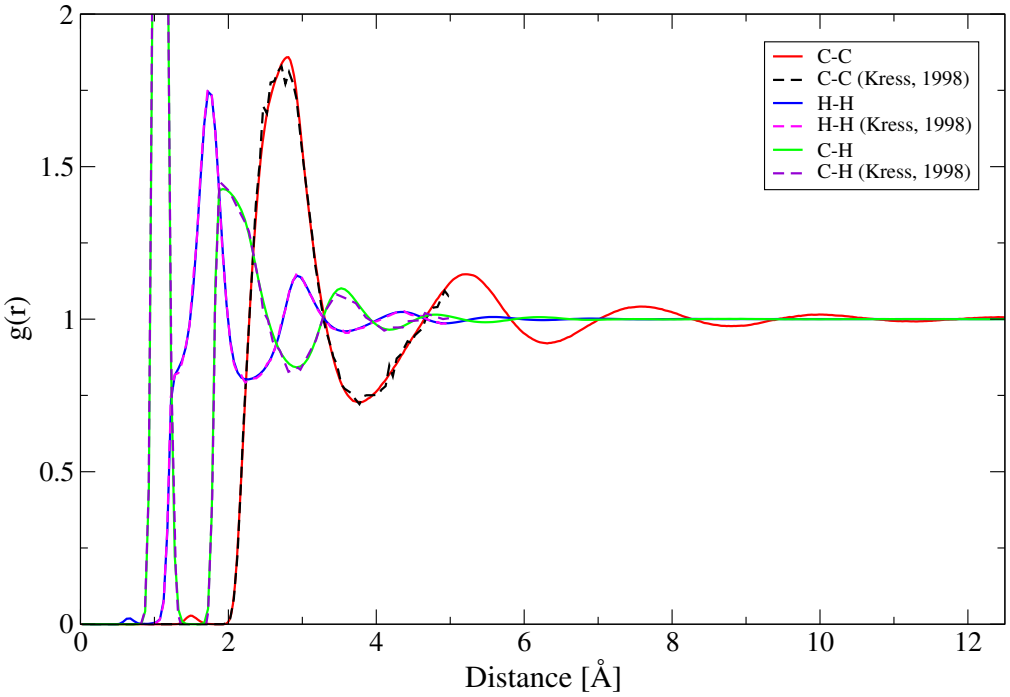

Figure 5.1: Partial pair-correlation functions g(r) of liquid CH_4 at 2000 K.

In order to sustain linear scaling in terms of computational cost and at the same time and memory requirement, all sparse matrices are stored in the common Compressed Row Storage (CRS) format. Due to the fact that the algorithm heavily relies on the multiplication of sparse matrices, we have put particular emphasis on an efficient parallel implementation. In that the data are distributed to the individual processor cores by employing a space-filling Hilbert curve to keep the load balanced [18]. While for solid state systems a very good scalability has been observed, for disordered liquids studied here the situation is substantially less favourable. A more efficient scheme, which dynamically rearranges the matrices between the various processor cores, or even distributes them fully at random is a desirable aim and future work.

5.3 Results and Discussion

In this section, we investigate the dissociation of methane and its implications for giant gas planets such as Uranus and Neptune at different temperatures.

Furthermore, we show that the algorithms scales linearly with the system size. To assess the accuracy of our method, we study a sample comprising of 1000 CH_4 molecules (2×2×2 the size of our unit cell) at $T = 2000$ K and compare the partial pair-correlation functions, as obtained by the present scheme, with the results of Kress et al. [82] using exactly the same model [57]. As can be seen in Figure 5.1, the agreement is excellent.

To demonstrate that linear system size scaling is indeed attained, in Figure 5.2 the average runtime for a complete SCTB MD step at $T = 2000$ K is shown for various system sizes using a single core of a 2.40 GHz Intel Westmere processor. Specifically, we have considered eight different systems, with 40, 320, 625, 2560, 5000, 16875, 40000, and 78125 atoms, respectively. As can be seen in Figure 5.2, the scaling is essentially linear with system size. Comparing the runtime with a divide and conquer diagonalization algorithm unveils that the crossing point, after which the linear scaling algorithm becomes computationally more favourable, is at $N_C \approx 425$ atoms.
To assess the almost perfectly linear scaling, we use linear regression. This means that we assume to have linear scaling and then we test this hypothesis a posteriori. The best possible linear function to the given values is fitted through the least squares approach. We evaluate the coefficient of determination

$$R^2 = \frac{\sum_{i=1}^{n}(f(x_i) - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{5.1}$$

where $f : \mathbb{R} \to \mathbb{R}$ is our regression line, (x_i, y_i) are the pair of variates of the number of atoms and the corresponding computational time, and $\bar{y}$ is the average of all values y_i. The coefficient of determination is a statistical measure of how well the regression line approximates the given data. If $R^2 = 1$, the regression line fits the data perfectly. We did regression analysis with both all data points and the data points from 2560 atoms on. In the first case, we get $R^2 = 0.995$ and in the second case, we get $R^2 = 0.998$. So, our scaling is almost perfectly linear. The regression line for the full set can be found in Figure 5.2.

However, beside the formal scaling with system size, the corresponding prefactor is also rather important and depends on the spectral width of the Hamil-

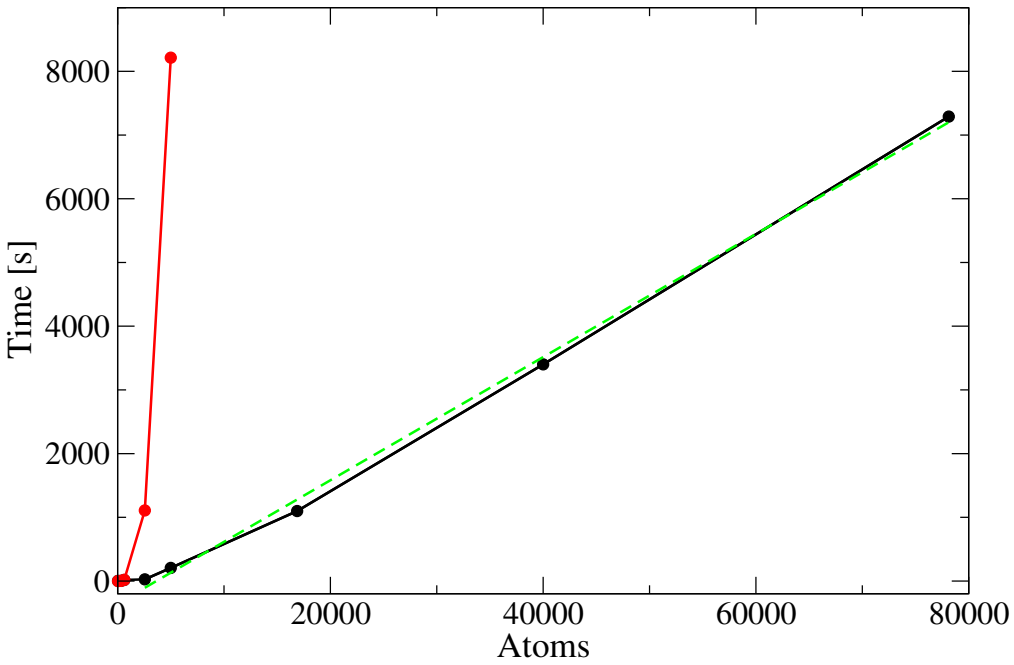

Figure 5.2: The average walltime for a single SCTB MD step versus the number of atoms on a single core of a 2.40 GHz Intel Westmere processor. The walltime using a divide and conquer diagonalization algorithm is shown in red, while the present linear scaling scheme is denoted in black. The dashed green line is the regression line and illustrates perfect linear system size scaling.

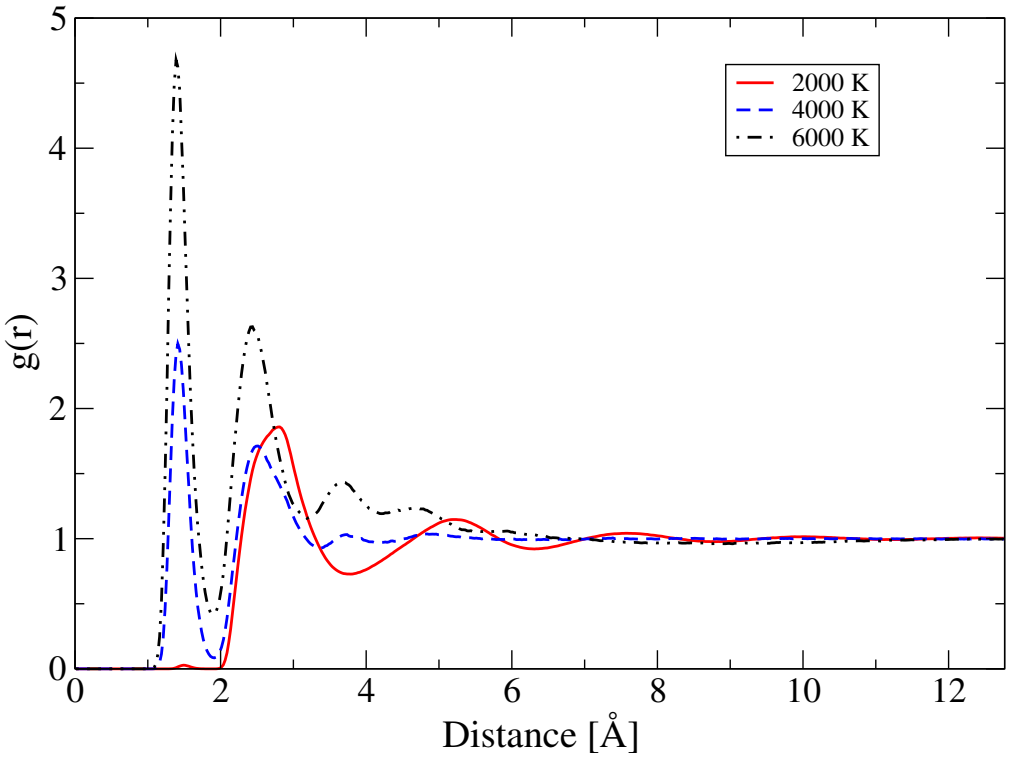

Figure 5.3: Comparison of the C-C PCF at 2000 K (solid red line), 4000 K (dashed blue line), and 6000 K (dot-dashed black line).

tonian $\Delta\varepsilon$. In the case of Chebyshev polynomial based Fermi operator expansion methods, the computational cost to achieve an accuracy of 10^{-D} has been found to scale like $D\beta\Delta\varepsilon$ [8] as described in Chapter 4. Apparently, this entails a fairly large prefactor if either high accuracy is required, or the electronic temperature is low, or when $\Delta\varepsilon$ is large. The latter is typically the case for an all-electron calculation, or if a plane wave basis set is employed. Nevertheless, the usage of fast polynomial summation methods leads to the more favourable scaling $\sqrt{\beta\Delta\varepsilon}$ [91, 92, 148]. For the present hybrid approach this results in an even better sub-linear scaling of $\sqrt[3]{\beta\Delta\varepsilon}$ [28], which makes it particularly attractive for highly accurate all-electron *ab initio* calculations, or when a high energy resolution is required. Together with the methods proposed by Lin et al. [94, 95], this is the best scaling with respect to β and $\Delta\varepsilon$ reported so far. Based on a multipole representation of the Fermi operator, the latter scales as $\ln(\beta\Delta\varepsilon)\ln(\ln(\beta\Delta\varepsilon))$, which depending on the actual value of $\beta\Delta\varepsilon$ is either slightly lower or larger than the present cubic root scaling.

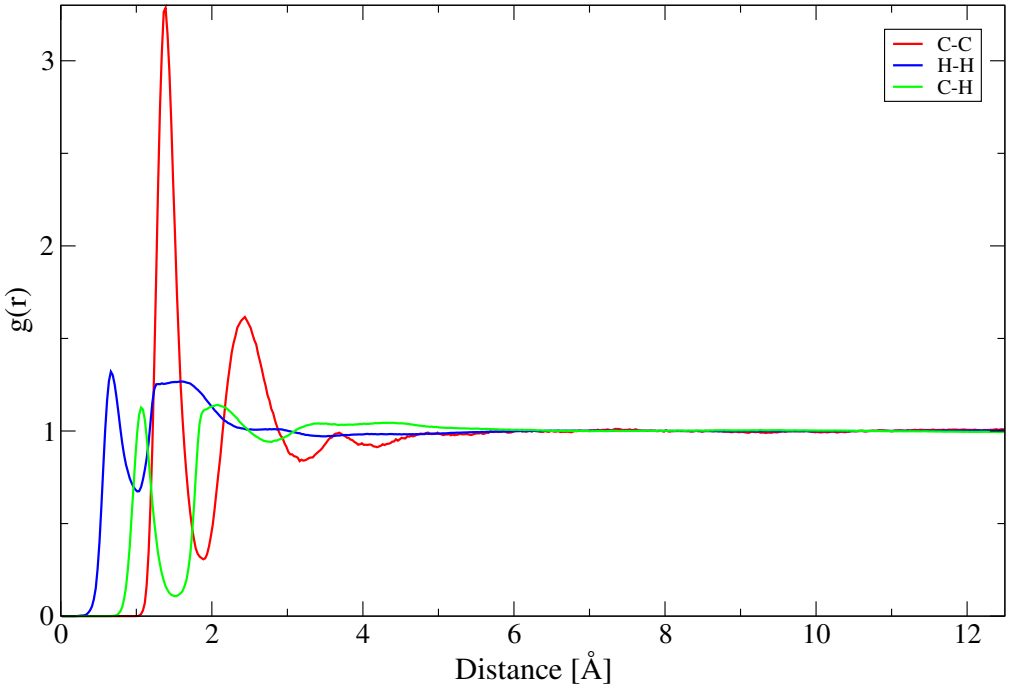

Figure 5.4: Partial pair-correlation functions g(r) of liquid CH_4 at 8000 K.

The fact that even metallic systems can be treated with linear system size scaling is demonstrated on exactly the same system, though at $T = T_e = 8000$ K. We find that at this temperature the CH_4 molecules are partially dissociated, as indicated by the reduced intramolecular C-H peak in Figure 5.4. Similar, from the first C-C and H-H peaks, the occurrence of covalent C-C bonds and H_2 molecules can be deduced. Moreover, a noticeable fraction of monoatomic hydrogen can be identified, which immediately suggests that hydrogen is on the verge of a liquid-liquid phase transition into an atomic fluid phase that is in agreement with recent AIMD calculations [111, 147]. Eventually, the electronic band-gap is vanishing, which is most likely due to the emergence of monoatomic hydrogen.

Now, we study the behaviour of liquid methane at 4000 and 6000 K and discuss its implications. The partial pair correlation functions (PCF) of our simulations are shown in Figures 5.3-5.6. As can be seen in Figure 5.3, as well as Figure 5.6, at $T = 2000$ K essentially no covalent C-C and H-H bonds are present. The only remaining significant peaks at 2.81 Å and 1.74 Å represent the average C-C and H-H distances between two adjacent CH_4 molecules,

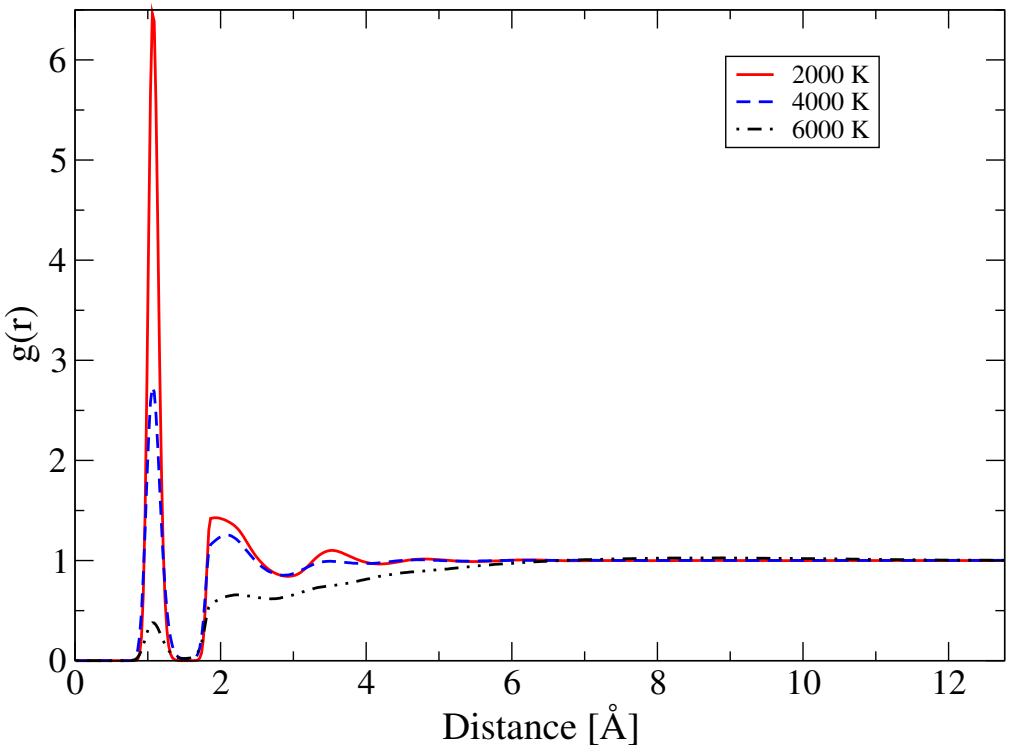

Figure 5.5: Comparison of the C-H PCF at 2000 K (solid red line), 4000 K (dashed blue line), and 6000 K (dot-dashed black line).

respectively. The insignificant peaks at 1.47 Å in Figure 5.3 and 0.75 Å in Figure 5.6 do not point to an onset of dissociation, as proposed by experiment [11, 137], but are rather due to fleetingly broken C-H bonds caused by finite temperature. Consequently, the sharp intramolecular peak in Figure 5.5 at around 1.075 Å can be ascribed to covalent C-H bonds. The corresponding coordination numbers, as obtained by integrating the associated PCFs up to their first minima, are shown in Table 5.1. In the case of C-H the partial coordination number is 3.984, which indicates that the liquid at $T = 2000$ K, except for single fleetingly broken C-H bonds, is nearly exclusively made up of undissociated CH_4 molecules. This view is consistent with other AIMD studies [5, 82, 145], but at variance to theoretical ground-state calculations [133], as well as experimental measurements [11, 113, 137], no signs for dissociation have been found.

For the most relevant case at $T = 4000$ K and $P \approx 100$ GPa, the situation is much different and evidences for dissociation can indeed be observed. As can

be seen in Figure 5.3, covalent C-C bonds are appearing as well as covalently bonded H_2 dimers, as shown in Figure 5.6. As a consequence, the height of the intramolecular C-H peak in Figure 5.5 is much reduced, though still existing. From Table 5.1 it can be deduced that nearly half of the covalent C-H bonds are broken, which indicates that methane does dissociate only partially to form hydrocarbon chains with mainly two and three carbon atoms, as well as H_2. More precisely, the CH_4 molecules dissociate and recombine to form C_2H_6 and to a smaller extend C_3H_8, which is in agreement with previous AIMD studies [5, 82, 145]. However, we find no sustained sign for the presence of C_2H_2, which has been detected in the atmosphere of Neptune [29]. But, we do find seeds of somewhat longer sp^2-bonded chains and ring-like carbon structures, but definitely no sign of sp^3 carbon bonds, i.e. no diamond-like carbon. This is consistent with the computed vibrational density of states of Spanu et al. [145], who reported a noticeable feature at 1600 cm^{-1} that can be attributed to threefold coordinated carbon atoms in graphite-like configurations. That is to say that on the one hand our calculations are in agreement with experiment by implying that in the middle ice layer CH_4 molecules itself are not present, but merely its dissociated constituents, which confirms that the interior chemistry of Uranus and Neptune is more complex than previously assumed. On the other hand, our results differ in the sense that at $T = 4000$ K we do not find any evidence for diamond-like carbon, as reported by the very same experiments [11, 113, 137].

Even deeper within the planet at even higher temperature of $T = 6000$ K, the remaining CH_4 have fully dissociated as indicated by the vanishing intramolecular C-H peak in Figure 5.5. On the other hand, the first peak in Figure 5.3, which is due to C-C bonds, as well as the covalent H-H peak in Figure 5.6 are even more pronounced as is the case for $T = 4000$ K. As shown in Table 5.1, the partial C-H coordination number is rather small, which entails that contrary to $T = 4000$ K hydrocarbon chains are no longer present, but have completely dehydrogenated into polymeric carbon and hydrogen. As before, no evidence for sp^3-bonded diamond could be found, which indicates that even higher pressures are required to condense carbon into diamond.

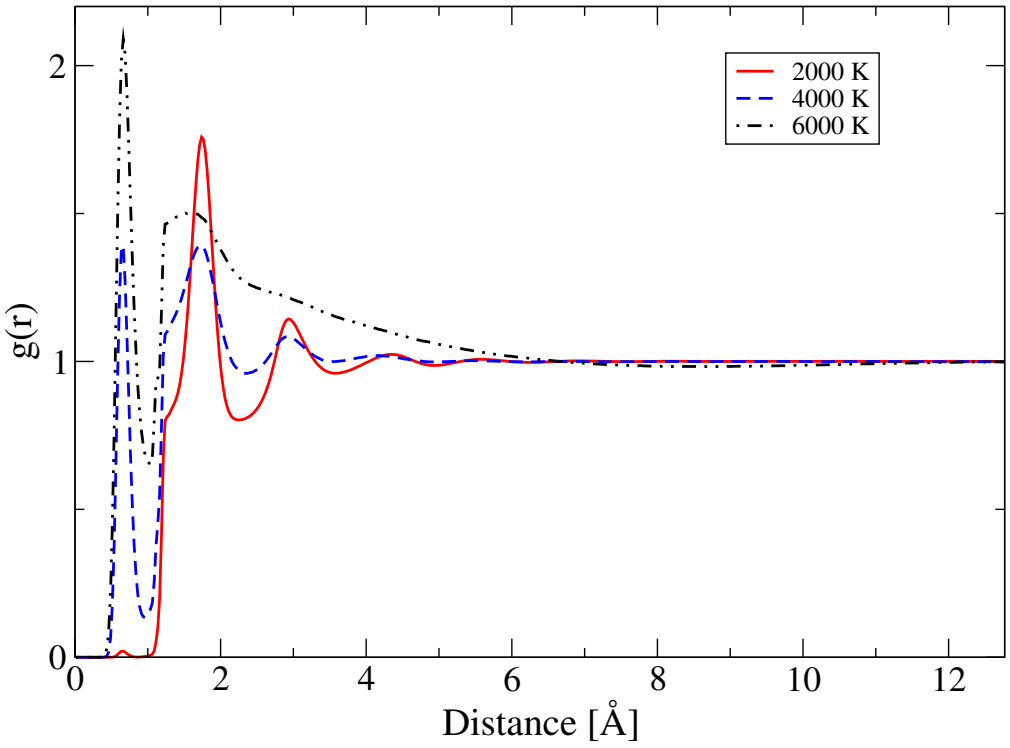

Figure 5.6: Comparison of the H-H PCF at 2000 K (solid red line), 4000 K (dashed blue line), and 6000 K (dot-dashed black line).

Table 5.1: Partial coordination numbers, as obtained by integrating the associated PCFs up to their first minima, for all investigated temperatures.

Temperature	C-C	C-H	H-H
2000 K	0.080	3.984	0.004
4000 K	1.244	2.298	0.408
6000 K	2.556	0.417	1.115

However, in contrast to previous AIMD simulations [5, 82, 145] we find for the first time that at $T = 6000$ K hydrogen is no longer solely molecular, but a noticeable fraction of monoatomic hydrogen can be identified. This immediately suggests that at $T = 6000$ K the present hydrogen molecules are on the verge of a liquid-liquid phase transition into an atomic fluid phase, which is in agreement with recent calculations [111, 147]. In our calculations we find that the band gap is rapidly decreasing with temperature and vanishing at $T = 6000$ K. Together with the fact that dissociation has shown to be accompanied with the metallization [7, 138], our prediction of liquid atomic hydrogen may lead to an explanation for the large magnetic fields of planets such as Uranus and Neptune through a dynamo-like mechanism by electrical currents in the liquid metallic regions of their interiors [116, 117]. Since we find that liquid CH_4, where it is stable at $T = 2000$ K, is a wide band-gap insulator, it can only fully dehydrogenated contribute to the magnetic field in the form of metallic hydrogen, which indeed has been established experimentally at rather similar conditions by shock-compression experiments [115, 152]. Moreover, the motion of charged particles trapped in such magnetic fields causes the generation of radio waves. In fact, planetary radio experiments aboard the Voyager II flyby mission detected a wide variety of radio emissions for both planets [150, 151].

If large enough, amorphous or crystalline carbon clusters precipitate and sink towards the planetary center as sediment via gravitational settling. The corresponding release of energy has been estimated to be a substantial fraction of the internal heat production, which would explain for instance why Neptune radiates more than twice the energy it receives from the sun [11]. It is also likely the cause for the externally observed high luminosity and could even contribute to the convective motions of its fluid interior of Neptune. The reason why Uranus does not have such internal heat flow mechanism is still unknown [61]. Nevertheless, the similarity of the internal structures of these two planets suggests that the suppressed convection of Uranus may be a consequence of its closer proximity to the sun. In contrast, saturated hydrocarbons such as C_2H_6 and H_2, being the products of the above ascertained decomposition of CH_4 at $T = 4000$ K, do not precipitate and instead rather rise to join the atmo-

sphere. As a consequence, this process could be responsible for the anomalous abundance of H_2 in the atmospheres of both planets, and in the case of Neptune may also account for the observed wealth of atmospheric C_2H_6, where it might be brought up from the deep interior by the afore-elucidated convection process. Therefore, the present results imply that chemical processes such as phase transformations at extreme temperatures and pressures must be considered in order to provide a more realistic model of the interiors of giant gas planets.

5.4 Conclusion

Even though our calculations provide a consistent picture of the deep chemistry of Neptune and Uranus, the remaining question is why no diamond formation could be observed, whereas experimentally it is reported to occur from $P = 20$ GPa and $T = 2000$ K on. Due to the fact that liquid methane is optically transparent and can not simply be heated by a laser beam, it is therefore common practice to include a noble metal absorber within laser-heated diamond anvil cell experiments. Spanu et al. reported that without a metallic absorber no formation of complex hydrocarbons and H_2 at $T = 2000$ K could be determined, which not only agrees with the findings of the present work but also indicates that liquid CH_4 resides in a metastable state. On the contrary, at the presence of a nobel metal, liquid CH_4 readily dissociates [145].

We conclude by noting that another possibility to explain the discrepancy between theory and experiment may be the existence of a homogeneous nucleation mechanism, similar to the one recently proposed by Khaliullin et al. for the direct graphite-to-diamond transition [38, 67, 68].

Chapter 6

An Algorithm to Calculate the Inverse Principal p-th Root of Symmetric Positive Definite Matrices

In this chapter, we address the general mathematical problem of computing the p-th root of a given matrix in a fast way by the help of an iteration function. As we have seen in Chapter 4, where we explained the Newton-Schulz iteration to compute the inverse of a given matrix, this problem directly affects our computations.

We present a new iteration function that enables calculating the inverse p-th root of a given matrix for an arbitrary p. We evaluate the order of convergence contingent upon a parameter q. By choosing q adaptively, better results than with before known formulas of this type can be achieved as less iterations and matrix-matrix multiplications are required.

The performance is evaluated by a MATLAB code using symmetric positive definite random matrices with various densities, condition numbers and spectral radii.

6.1 Introduction

The first attempts to calculate the inverse of a matrix by the help of an iterative scheme were amongst others made by Schulz [139] in the early thirties of the last century. This resulted in the well-known Newton-Schulz iteration scheme that is widely used to approximate the inverse of a given matrix. One of the advantages of this method is that for a particular start matrix as initial guess convergence is guaranteed [123]. The convergence is of order two, which is already quite satisfying, but there were a lot of attempts to speed up this iteration scheme and to extend it to a formula to calculate not only the inverse but a general inverse p-th root of a given matrix [12, 54, 63, 64, 144]. This is an important task because, besides the pure mathematical interest, in many applications in physics or chemistry one needs efficient methods to calculate the (inverse) square root or the inverse of a matrix. An example is our linear scaling scheme presented in Chapter 4, or Löwdin's method of symmetric orthogonalization [98]. The latter transforms the eigenvalue problem for overlapping orbitals into an equivalent problem with orthogonal orbitals, whereby the inverse square root of the overlap matrix has to be calculated. This is for instance necessary in the extended Hückel method [55], and also used in tight-binding [44, 124].

Common problems in the attempts cited above are the stability of the iteration formula and its convergence. For $p \neq 1$, most of the iteration schemes have quadratic order of convergence, a rare exception is for instance Halley's method [45, 47, 64], which is of order three. Altman [4] however generalized the Newton-Schulz iteration to an iterative method of inverting a linear bounded operator in a Hilbert space. He constructed the so called hyperpower method of any order of convergence and proved that the method of degree three is the optimum one, as it gives the best accuracy for the same number of multiplications.
Here, we describe an iteration function for the calculation of the inverse p-th root of a given matrix A. In this scheme, we have two variables, the natural p and another natural $q \geq 2$ that represents the order of expansion. We show that two special cases of this formula are Newton's method for matrices

[12, 45, 46, 58, 63, 64, 144] and Altman's hyperpower method [4].

6.2 Previous Work

The study of the calculation of the inverse p-th root, where p is a natural, has been treated prevalently by various authors. The characterization of the problem is quite simple, in general, for a given matrix A, one wants to find a matrix B that fulfils $B^{-p} = A$. If A is non-singular, one can always find such a matrix B, but B is not unique. The problem of computing the p-th root of a A is strongly connected with the spectral analysis of A. If, for example, A is a real or complex matrix of order n with no eigenvalues on the closed negative real axis, B can be uniquely defined [12]. As we deal only with symmetric positive definite matrices, we can restrict ourselves to this unique solution, which is called the principal p-th root and guaranteed by the following theorem.

Theorem 1 (Higham [53], 2008)**.** *Let $A \in \mathbb{C}^{n\times n}$ have no eigenvalues on $\mathbb{R}^-$. There is a unique p-th root B of A all of whose eigenvalues lie in the segment $\{z : -\pi/|p| < \arg(z) < \pi/|p|\}$, and it is a primary matrix function of A. We refer to B as the principal p-th root of A and write $B = A^{1/p}$. If A is real then $A^{1/p}$ is real.*

Remark 1. *Here, $p < 0$ is also included, so that* Theorem 1 *holds also for the calculation of inverse p-th roots.*

The calculation of such a root is usually done by the help of an iteration function, as computation by brute force is computationally very demanding or even infeasible for large matrices. Iteration functions can also be very helpful if the scaling of the computation for sparse matrices should be reduced because the intermediately occurring matrices can be truncated. One should always keep in mind that the inverse p-th roots of sparse matrices are in general not anymore sparse but usually full matrices.

One of the most discussed iteration schemes for computing the p-th root of a matrix is based on Newton's method for finding roots of functions. One can approximate the root $\hat{x}$ of $f : \mathbb{R} \to \mathbb{R}$, meaning that we have $f(\hat{x}) = 0$, by the

iteration

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}. \tag{6.1}$$

Here, $x_k \to \hat{x}$ for $k \to \infty$ if x_0 is an appropriate initial guess. If one chooses $f(x) = x^p - a$ for an arbitrary a, then we get

$$x_{k+1} = \frac{1}{p}\left((p-1)x_k + a x_k^{1-p}\right). \tag{6.2}$$

One can also deal with matrices and study the resulting rational matrix function $F(X) = X^p - A$ [58], where $F : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$ and F' is the Fréchet derivative of F, see [144]. This has been the subject of a not so unimportant number of papers [10, 12, 45–47, 53, 54, 58, 63, 64, 88, 123, 127, 131, 144]. It is clear that this iteration converges to the p-th root of the matrix A if X_0 is chosen close enough to the true root. Mathematically, this means that we need to fulfil the condition $\|I - AX_0^p\| < 1$. The convergence is quadratic as soon as the iterates are close enough to limit of the iteration [58]. Smith [144] also shows that Newton's method for matrices has some issues concerning numerical stability for not so well-conditioned matrices, but this is not the topic of this work.

In their paper [12], Bini, Higham, and Meini proved that the matrix iteration

$$B_{k+1} = \frac{1}{p}\left[(p+1)B_k - B_k^{p+1}A\right], \quad B_0 \in \mathbb{R}^{n\times n} \tag{6.3}$$

converges quadratically to the inverse p-th root $A^{-1/p}$ if $\|I - B_0^p A\| < 1$, $B_0 A = A B_0$ and $\rho(A) < p+1$ hold. Here, $\rho(A)$ is the spectral radius, which is defined as the largest absolute eigenvalue of A.

In his work dated by 1959, Altman described the hyperpower method [4]. Let V be a Banach space and $A : V \to V$ a linear, bounded, and non-singular operator and B_0 an approximate reciprocal of A satisfying $\|I - AB_0\| < 1$. For the iteration

$$B_{k+1} = B_k(I + R_k + R_k^2 + \ldots + R_k^{q-1}), \quad B_0 \in V, \tag{6.4}$$

the sequence $(B_k)_{k\in\mathbb{N}_0}$ converges towards the inverse of A. Here, $R_k = I - B_k^p A$ is the k-th residual.
Altman proved that the natural q corresponds to the order of convergence

of (6.4) so that in principle a method of any order can be constructed. He described the optimum method as those who gives the best accuracy for the same number of multiplications, and demonstrated that the optimum method is obtained for $q = 3$.

To close this section, we recall some basic definitions, which are crucial for the next section. In the following, the iteration function $\varphi : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$ is assumed to be sufficiently often continuously differentiable.

Definition 1. *Let $\varphi : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$ be an iteration function. The process*

$$B_{k+1} = \varphi(B_k), \quad k = 0, 1, \ldots \tag{6.5}$$

is called convergent *to $Z \in \mathbb{C}^{n\times n}$ if it exists a constant $0 < c < 1$ so that for all start matrices $B_0 \in \mathbb{C}^{n\times n}$ with $\|I - B_0 Z\| \leq c$, we have $\|B_k - Z\| \to 0$ if $k \to \infty$.*

Definition 2. *A* fixed point *Z of the iteration function* (6.5) *is such that $\varphi(Z) = Z$ and is said to be* attractive *if $\|\varphi'(Z)\| < 1$.*

Definition 3. *Let $\varphi : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$ be an iteration function with fixed point $Z \in \mathbb{C}^{n\times n}$. The process* (6.5) *is called* convergent of order *$q \in \mathbb{N}$ if it exists a constant $0 < c < 1$ so that for all start matrices $B_0 \in \mathbb{C}^{n\times n}$ with $\|I - B_0 Z\| \leq c$ we have*

$$\|B_{k+1} - Z\| \leq C\|B_k - Z\|^q \text{ for } k = 0, 1, \ldots, \tag{6.6}$$

where $C > 0$ is a constant with $C < 1$ if $q = 1$.

We also want to remind a well-known theorem concerning the order of convergence.

Theorem 2. *Let $f(x)$ be a function with fixed point $x^\star$. If $f(x)$ is q-times continuously differentiable in a neighbourhood of $x^\star$ with $q \geq 2$ then $f(x)$ has order of convergence q if and only if*

$$f'(x^\star) = \ldots = f^{(q-1)}(x^\star) = 0 \text{ and } f^{(q)}(x^\star) \neq 0. \tag{6.7}$$

Proof. The proof can be found in the book of Gautschi [36, pp. 235–236]. □

6.3 Generalization of the Problem

In this section, we present a new iteration scheme to compute the p-th root of a matrix which contains Newton's method for matrices and the hyperpower method as special cases. Thus, we study a general expression to compute the p-th root of a matrix which comes along with variable q as the order of expansion. In Altman's case, q is the order of convergence, but as we will see, this does not hold generally for our formula. Nevertheless, choosing q larger than two leads often to an increase in performance, meaning that less multiplications, iterations, and computational time are required. We discuss in the next section how q can be chosen adaptively. The central message of this chapter is the following

Theorem 3. *Let $A \in \mathbb{C}^{n\times n}$ be a matrix, and p, $q \in \mathbb{N}$ with $q \geq 2$. We define the function*

$$\begin{aligned} &\varphi : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n} \\ &\qquad X \mapsto \frac{1}{p}\left[(p-1)I - \sum_{i=1}^{q} \binom{q}{i} (-1)^i (X^p A)^{i-1}\right] X, \end{aligned} \tag{6.8}$$

where X has to be fulfil $AX = XA$, and the iteration

$$B_{k+1} = \varphi(B_k), \quad B_0 \in \mathbb{C}^{n\times n}. \tag{6.9}$$

If $\|I - B_0^p A\| < 1$ and $B_0 A = A B_0$, then it holds

$$\lim_{k\to\infty} B_k = A^{-1/p}. \tag{6.10}$$

If $p > 1$, the order of convergence of (6.8) is quadratic. If $p = 1$, the order of convergence of (6.8) is equal to q.

Remark 2. *One can use the same formula to calculate the p-th root of A where p is negative or even choose $p \in \mathbb{Q} \setminus \{0\}$. But this is not a competitive method because for negative p, one has to compute inverse matrices in every iteration step and for non-integers the binomial theorem and the calculation of powers of matrices are not that simple anymore.*

Proof. Now, we prove Theorem 3. To make the representation of (6.8) more convenient, we rearrange the following term

$$\sum_{i=1}^{q} \binom{q}{i} (-1)^i (X^p A)^{i-1} \tag{6.11}$$

$$= \left(\sum_{i=1}^{q} \binom{q}{i} (-1)^i (X^p A)^i \right) (X^p A)^{-1}$$

$$= \left(\sum_{i=0}^{q} \binom{q}{i} (-1)^i (X^p A)^i - I \right) (X^p A)^{-1}$$

$$= ((I - X^p A)^q - I)) (X^p A)^{-1}, \tag{6.12}$$

so that we finally get

$$\varphi(X) = \frac{1}{p} \left[(p-1)X - ((I - X^p A)^q - I) X^{1-p} A^{-1} \right]. \tag{6.13}$$

Due to Definition 2, one can easily see that $A^{-1/p}$ is an attractive fixed point of (6.20) by calculating $\varphi(A^{-1/p})$ and $\varphi'(A^{-1/p})$

$$\begin{aligned} \varphi(A^{-1/p}) &= \frac{1}{p} \left[(p-1)A^{-1/p} - \left((I - (A^{-1/p})^p A)^q - I \right) \right) (A^{-1/p})^{1-p} A^{-1} \right] \\ &= \frac{1}{p} \left[(p-1)A^{-1/p} + A^{-1/p} \right] = A^{-1/p}, \end{aligned} \tag{6.14}$$

$$\begin{aligned} \varphi'(X) &= q(I - X^p A)^{q-1} \\ &\quad + \frac{p-1}{p} \left(((I - X^p A)^q - I) (X^p A)^{-1} + I \right), \end{aligned} \tag{6.15}$$

$$\varphi'(A^{-1/p}) = 0. \tag{6.16}$$

To apply Theorem 2, we calculate for our iteration

$$\varphi(X) = \frac{1}{p} \left[(p-1)X - ((I - X^p A)^q - I) (X^p A)^{-1} X \right]$$

not only the first (cf. equation (6.15)) but also the second derivative

$$\begin{aligned}\varphi^{(2)}(X) = &- pq(q-1)X^{p-1}A(I-X^pA)^{q-2} \\ &+ q(1-p)X^{-1}(I-X^pA)^{q-1} \\ &+ (p-1)X^{-p-1}A^{-1}(I-X^pA)^q \\ &- (p-1)X^{-p-1}A^{-1}.\end{aligned} \tag{6.17}$$

We have already seen that $\varphi'(A^{-1/p}) = 0$ for every p, but for the second derivative holds

$$\varphi^{(2)}(A^{-1/p}) = 0 \iff p = 1. \tag{6.18}$$

One can also show that $\varphi^{(j)}(A^{-1/p}) = 0$ if and only if $p = 1$ for $j = 3, \ldots, q-1$ because we have

$$\varphi^{(j)}(X) = \sum_{i=0}^{j} J_{i,j}(I-X^pA)^{q-i} + (p-1)J_jX^{1-p-j}A^{-1}, \tag{6.19}$$

where $J_{i,j}$ and J_j are rational non-zero numbers. So, we have convergence of an arbitrary order.

This implies that according to the Theorem 2, the convergence of the presented formula (6.8) is exactly quadratic for any q if $p \neq 1$. For $p = 1$, we have that the order of convergence is identical with the chosen q. This is what had to be demonstrated. □

As a trivial consequence, we conclude that a larger q does in general not lead to a higher order of convergence. However, we make several tests with the help of MATLAB [103], to get the number of iterations, matrix-matrix multiplications, and the computational time needed to obtain the p-th root of a given matrix within a predefined accuracy contingent upon q. We find that a higher order of expansion in the sum (6.11) leads in almost all cases to a better performance. Details are presented in the next section.

From now on, we deal with equation (6.13) for the iteration of matrices $B_k \in \mathbb{C}^{n\times n}$ and define $B_{k+1} = \varphi(B_k)$. We assume that the start matrix B_0 satisfies $B_0A = AB_0$ and $\|I - B_0^pA\| < 1$. One can show that in that case, it holds $B_kA = AB_k$ for every $k \in \mathbb{N}$, see [144]. Thus, we have

$$\begin{aligned} B_{k+1} &= \frac{1}{p}\left[(p-1)B_k - \left((I - B_k^p A)^q - I\right) B_k^{1-p} A^{-1}\right], \\ B_0 &\in \mathbb{C}^{n\times n}. \end{aligned} \tag{6.20}$$

For $p = 1$ and $q = 2$, we get the already mentioned Newton-Schulz iteration that converges quadratically to the inverse of A

$$\begin{aligned} B_{k+1} &= -\left((I - B_k A)^2 - I\right) A^{-1} \\ &= (-(B_k A)^2 + 2B_k A)A^{-1} \\ &= 2B_k - B_k^2 A. \end{aligned} \tag{6.21}$$

For $q = 2$ and any p, we get the iteration

$$\begin{aligned} B_{k+1} &= \frac{1}{p}\left[(p-1)B_k - \left((I - B_k^p A)^2 - I\right) B_k^{1-p} A^{-1}\right] \\ &= \frac{1}{p}\left[(p-1)B_k - \left((B_k^p A)^2 - 2B_k^p A\right) B_k^{1-p} A^{-1}\right] \\ &= \frac{1}{p}\left[(p-1)B_k - \left(B_k^{p+1} A - 2B_k\right)\right] \\ &= \frac{1}{p}\left[(p+1)B_k - B_k^{p+1} A\right]. \end{aligned} \tag{6.22}$$

This is exactly the matrix iteration (6.3) that has been discussed in the work of Bini, Higham, and Meini [12].

We now proceed by dealing with the iteration formula in the case $p = 1$ for higher orders ($q > 2$) and show that it converges faster. For that purpose, we take equation (6.20) and calculate for $p = 1$

$$\begin{aligned} B_{k+1} &= \frac{1}{p}\left[(p-1)B_k - \left((I - B_k^p A)^q - I\right) B_k^{1-p} A^{-1}\right] \\ &\overset{p=1}{=} \left[I - (I - B_k A)^q\right] A^{-1}. \end{aligned} \tag{6.23}$$

We now prove that this is convergent of order q in the sense of Definition 3.

$$\begin{aligned}
\|B_{k+1} - A^{-1}\| &= \|(I - (I - B_k A)^q)A^{-1} - A^{-1}\| \\
&= \|(I - B_k A)^q A^{-1}\| \\
&= \|(I - B_k A)^q A^{-1} A^{-q} A^q\| \\
&\leq \|A\|^{q-1} \cdot \|(I - B_k A)^q (A^{-1})^q\| \\
&\leq \|A\|^{q-1} \cdot \|A^{-1} - B_k\|^q. \qquad (6.24)
\end{aligned}$$

Now, we show why iteration (6.20) coincides for $p = 1$ with Altman's work on the hyperpower method [4]. For that purpose, we rewrite the $(k+1)$-st iterate B_{k+1} in terms of the powers of the k-th residual $R_k = I - B_k^p A$.

$$\begin{aligned}
B_{k+1} &= \frac{1}{p}\left[(p-1)B_k - ((I - B_k^p A)^q - I)\,(B_k^p A)^{-1} B_k\right] \\
&= \frac{1}{p}\left[(p-1)B_k - (R_k^q - I)(I - R_k)^{-1} B_k\right] \\
&= \frac{1}{p}\left[(p-1)B_k + (R_k^{q-1} + R_k^{q-2} + \ldots + R_k + I)B_k\right] \\
&= \frac{1}{p} B_k \left[pI + \left(\sum_{j=1}^{q-1} R_k^j\right)\right]. \qquad (6.25)
\end{aligned}$$

Altman however proved convergence of any order of the iteration scheme (6.4)

$$B_{k+1} = B_k(I + R_k + R_k^2 + \ldots + R_k^{q-1}), \quad B_0 \in V$$

for the calculation of the inverse of a given linear, bounded, and non-singular operator $A \in V$. If we take for the Banach space $V = \mathbb{C}^{n\times n}$, this is exactly equation (6.25) with $p = 1$.

6.4 Numerical Results

Even if the mathematical analysis of our iteration function results in the awareness that larger q does not lead to a better order of convergence, we make numerical tests by varying p and q. Concerning the matrix A whose inverse p-th root should be determined, we take real symmetric positive definite random matrices with different densities and condition numbers. We do this by the

help of MATLAB [103] and elaborate the computational time in seconds (time), the number of iterations (#it) and matrix-matrix multiplications (#mult) until the calculated inverse p-th root is close enough to the true inverse p-th root.

6.4.1 The Scalar Case

First, we run the program in the scalar case. As the computation of roots of matrices is strongly connected with its eigenvalues, it is logical to study the formula for scalar quantities λ first. Thus, we have $n = 1$ in equation (6.20). We take values λ varying from 10^{-9} to 1.9 and choose $b_0 = 1$ as start value. For that choice of b_0, we have guaranteed convergence for $\lambda \in (0, 2)$. We calculate the inverse p-th root of λ, thus $\lambda^{-1/p}$. We set the threshold ε for the exit of the program to 10^{-8} and the maximum number of iterations to 35. Usually, a smaller tolerance ε should be sufficient but to see differences in the computational time, we choose this small threshold. Here, one should note that in the scalar case the computational time is not very meaningful due to its small differences for varying q. In the scalar case, we do not consider, as in the matrix case, the norm of the residual $r_k = 1 - b_k^p \cdot \lambda$ to compute the error, but the difference between the k-th iterate and the correct inverse p-th root, thus $\tilde{r}_k = b_k - \left(\frac{1}{\lambda}\right)^{1/p}$. This is due to the fact that in the scalar case, one can easily get $\left(\frac{1}{\lambda}\right)^{1/p}$ by a straightforward calculation. Note that we have not necessarily $|\tilde{r}_k| < 1$, but only $|r_k| < 1$. For better distinguishing the scalar from the matrix case, we write in the scalar case, as above, b_k, λ, $\tilde{r}_k$, and r_k instead of B_k, A, and R_k.

The range of q is chosen rather wide to see the influence of this value on the number of iterations and the computational time. In agreement with our observations in the matrix case, which we describe in the next subsection, there is a coherence between the choice of q, the number of multiplications, the number of iterations, and the computational time. In the following, we elaborate general rules for the optimal choice of q.
For a certain number of iterations needed, the number of multiplications and divisions is always lowest for the lowest q that entails this specific number of iterations. But it can happen that the number of iterations is not steadily

Table 6.1: Results for $p = 2$, $\lambda = 1.5$. Optimal values in bold.

time	#it	#mult	q
1.7e−05	5	37	2
1.9e−05	4	34	3
1.7e−05	**3**	**29**	**4**
1.8e−05	4	42	5
1.6e−05	**3**	35	6
1.9e−05	4	50	7
1.7e−05	4	54	8

Table 6.2: Results for $p = 2$, $\lambda = 10^{-9}$. Optimal values in bold.

time	#it	#mult	q
2.7e−05	27	191	2
3.4e−05	17	138	3
2.9e−05	14	128	4
2.6e−05	12	**122**	5
2.2e−05	11	123	6
2.1e−05	**10**	**122**	**7**
2.6e−05	**10**	132	8

decreasing, as one can see in the case for computing the inverse square root of $\lambda = 1.5$ (Table 6.1). Nevertheless, we conclude that the best choice is $q = 4$, as we have the lowest number of iterations and multiplications and almost the lowest computational time.

In other cases, the evaluation is simpler. For example for computing the inverse square root of $\lambda = 10^{-9}$, we have, when varying q from 2 to 8 clearly the best result for $q = 7$, as in that case, the number of iterations, the number of multiplications as well as the computational time is lowest (Table 6.2).

In most of the cases, it is not that easy to decide which q is the best for a certain tuple (p, λ). It can also happen that lowest number of iterations is attained for really large q, meaning $q > 20$. To present a general rule of thumb, we pick q from 3 to 8 as this usually gives a good and fast approximation of the inverse p-th root of a given $\lambda \in (0, 2)$. Hereby, we observe that for values close to 1, the optimal choice is in most, but not all cases, $q = 3$ and for values close to the borders of the interval, mostly $q = 6$ is the best choice. This implies that the further the value of λ is away from 1, the more important it is to chose a larger q. This can also be explained by the fact that in the scalar case, we start with $b_0 = 1$ as a first initial guess for the inverse p-th root.

Nonetheless, as one can see in Figure 6.1 for the exemplary case $p = 1$, $\lambda = 10^{-6}$, a larger q causes the iteration to enter after less iterations the quadratic convergence in each of the cases (cf. Appendix A for detailed results). This is due to the fact that in every iteration it is calculated

$$\begin{aligned} b_{k+1} &= \frac{1}{p}\left[(p-1) - \left((1-b_k^p\lambda)^q - 1\right)/(b_k^p\lambda)\right] b_k \\ &= b_k + \frac{1}{p}\left(\sum_{i=1}^{q-1} r_k^i\right) b_k. \end{aligned} \tag{6.26}$$

It can be seen that a larger q leads to more summands in (6.26) and therefore to larger steps and variation of b_{k+1}. This holds also for negative r_k, as we have $r_k \in (-1, 1)$ and therefore $|r_k^{i+1}| < |r_k^i|$. But a larger value for q obviously increases the performance just up to a certain limit. This is due to the fact that a larger value of q implies that $b_k^p\lambda$ is raised by larger exponents. Therefore the number of multiplications increases, what is, especially in the matrix case, the time consuming part. This is why we take into account the number of multiplications, the computational time, and the number of iterations.

6.4.2 The Matrix Case

For evaluating the performance of our formula for matrices, we make different set-ups by varying the variables p, q, the density d and the condition number

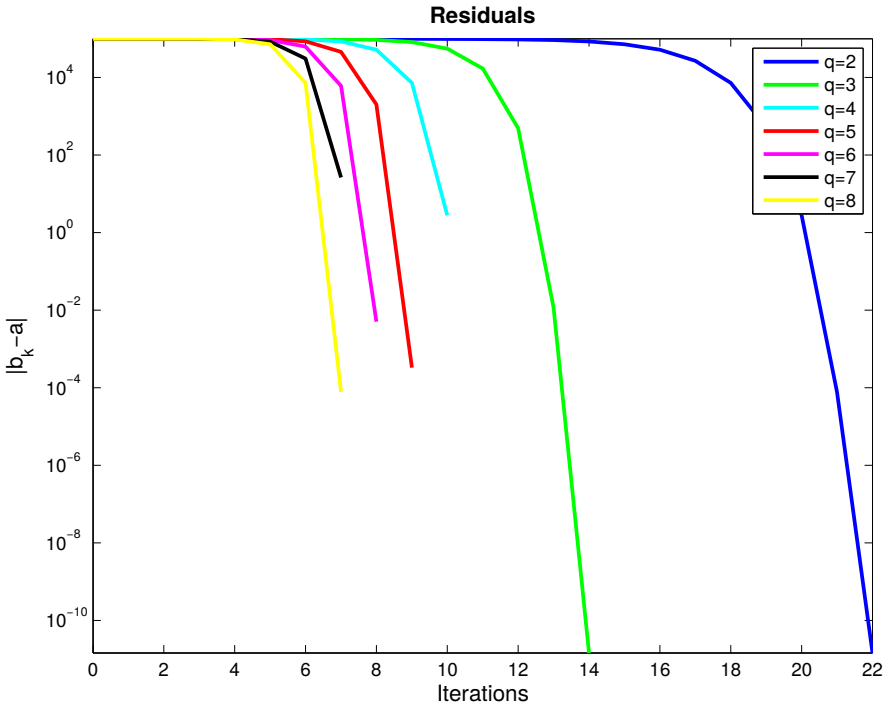

Figure 6.1: Residuals for $p = 1$ and $\lambda = 10^{-6}$.
Optimal choice is here $q = 5$, as $q = 7$ and $q = 8$ require less iterations but more multiplications.

c. The density of a matrix is defined as the number of its non-zero elements divided by its total number of elements. The condition number of a normal matrix, for which $AA^* = A^*A$ holds, is defined as the quotient of its largest absolute eigenvalue and its smallest absolute eigenvalue. The bigger the condition number is, the more ill-conditioned A is. Well-conditioned matrices have condition numbers close to 1. For each set-up, we take ten symmetric positive definite matrices $A \in \mathbb{R}^{1000\times 1000}$ with random entries, generated by the MATLAB [103] function `sprandsym`. This yields matrices with a spectral radius $\rho(A) < 1$. We store the number of iterations that the iteration needs to converge, the number of matrix-matrix multiplications as well as the computational time for each random matrix. Then, we average this values over the number of random matrices. The threshold is set to $\varepsilon = 10^{-4}$ and the maximum number of iterations t to 100. The maximum number of iterations has never been reached, so that setting $t = 30$ would have been sufficient. We choose q varying from 2 to 6, as for ill-conditioned matrices larger q sometimes causes divergence in the cases $p = 4$ and $p = 5$ due to numerical errors. In all other cases, where no divergence occurred, $q > 6$ was never the best choice as the number of iterations decreases not further, but the number of multiplications increases. Thus, this choice is not a restriction.

By using the sum representation like presented in (6.8)

$$B_{k+1} = \frac{1}{p}\left[(p-1)I - \sum_{i=1}^{q}\binom{q}{i}(-1)^i (B_k^p A)^{i-1}\right] B_k, \tag{6.27}$$

and by temporarily saving $B_k^p A$, we minimize the number of multiplications. It is evident that the number of matrix-matrix multiplications for the same number of iterations is lowest for smallest q. But a larger q can also mean less iterations, so usually the best q is the smallest q for the lowest possible number of iterations for a certain set-up.
In general, one can determine the number of matrix-matrix multiplications m as a function of p, q and the number of iterations j. We have

$$m = m(p, q, j) = p + ((q-1) + p)\, j. \tag{6.28}$$

We want to find out which parameters are decisive for the optimal choice of q, so we fix p and figure out which q is the best for

- sparse ($d \in \{0.001, 0.003\}$) and well-conditioned matrices ($c = 1.25$),
- sparse and ill-conditioned matrices ($c \in \{10, 50, 500\}$),
- not so sparse ($d = 0.01$) and well-conditioned matrices,
- not so sparse and ill-conditioned matrices,
- full ($d \in \{0.1, 0.8\}$) and well-conditioned matrices,
- full and ill-conditioned matrices.

It is not recommended to choose matrices with much larger condition number ($c \gg 500$), as this may lead to divergence due to very small eigenvalues.

To fulfil the conditions of Theorem 3, we claim that the start matrix B_0 commutes with the given matrix A. If B_0 is chosen as a plus-signed multiple of the identity matrix or the matrix A itself, $B_0 = \alpha I$ or $B_0 = \alpha A$ for $\alpha > 0$, then it is obvious that it holds $AB_0 = B_0A$ and AB_0 is symmetric positive definite.

In the first part of the calculations, we deal only with matrices A that have a spectral radius smaller than 1. If we take $B_0 = I$, then we have $\|I - B_0A\|_2 < 1$ due to the following

Lemma 1. *Let $C \in \mathbb{C}^{n\times n}$ be a Hermitian positive definite matrix with $\|C\|_2 \leq 1$. Then, it holds $\|I - C\|_2 < 1$.*

Proof. It is clear that $\|I\|_2 = 1$. Let U be the unitary matrix such that $U^{-1}CU = \operatorname{diag}(\mu_1, \ldots, \mu_n)$, where $\mu_i \in (0, 1]$ are the eigenvalues of C. Then we have

$$\begin{aligned} \|I - C\|_2 &= \|U^{-1}(I - C)U\|_2 \\ &= \|I - \operatorname{diag}(\mu_1, \ldots, \mu_n)\|_2 \\ &= \|\operatorname{diag}(1 - \mu_1, \ldots, 1 - \mu_n)\|_2 \\ &= 1 - \mu_{\min} < 1, \end{aligned} \tag{6.29}$$

where $\mu_{\min}$ is the smallest eigenvalue of C. □

We examine here a couple of exemplary cases. As they interest us most, we first choose sparse, ill-conditioned matrices A with $d = 0.003$ and $\text{cond}(A) = 500$ and obtain the following results. For $p = 1$, the number of matrix-matrix multiplications it lowest for $q = 3$, but concerning the number of iterations and the computational time, $q = 6$ is best. As one can see in Table 6.3, one needs 24% more matrix-matrix multiplications, but only 62.5% of the number of iterations. Every iteration is time consuming as intermediate results and the residuals have to be calculated and stored. Thus, one would conclude that for 1000×1000 matrices, the optimal method entails $q = 6$ but the situation is most probably different for larger matrices. For $p = 2$, the situation is similar, the least number of multiplications is reached in the case $q = 3$, but the computational time and the number of iterations are optimal for $q = 5$ (cf. Table 6.4). One should here note that the computational time is, contrary to the other two criteria, not a fully reliable quantity as the attended time can be influenced by the workload of the computer. As the differences concerning the number of multiplications are not so significant, $q = 5$ is best for those and also for reasonable larger matrices. In the other cases, optimal q is easy to determine. As one can see in Table 6.5, for $p = 4$, the number of iterations, multiplications as well as the computational time is lowest for $q = 4$. The situation is similar for $p = 3$ and $q = 5$, where $q = 3$ and $q = 4$ respectively are optimal.

We now consider another example to show that also in the matrix case, quadratic convergence is reached after less iterations for $q > 2$. We take again ill-conditioned matrices ($c = 500$), as hereby the differences in the iteration schemes occur more clearly. We take $p = 3$ and $d = 0.003$. As can be seen in Figure 6.2, the lowest number of matrix-matrix multiplications is clearly achieved for the case $q = 3$. Even if the cases $q = 5$ and $q = 6$ require one iteration less, they come along with considerably more multiplications. A larger q causes the iteration scheme to reach quadratic convergence after less iterations in the matrix case, as well. This can be seen in Figure 6.3 and in the Appendix, where the residuals are presented in detail. The optimal choice is here $q = 3$, as it is also the method that is computationally the least demanding.

Table 6.3: Results for $c = 500$, $p = 1$, $d = 0.003$.
Optimal values in bold.

time	#it	#mult	q
72.65	13	27	2
51.673	8	**25**	3
51.52	7	29	4
48.807	6	31	5
43.393	**5**	31	6

Table 6.4: Results for $c = 500$, $p = 2$, $d = 0.003$.
Optimal values in bold.

time	#it	#mult	q
71.23	11	35	2
52.213	7	**30**	3
48.593	6	32	4
44.23	**5**	32	**5**
47.353	**5**	37	6

However, after having studied a few exemplary cases, we want for a general rule to decide which method should be used for different types of matrices. Comparing all studied cases, we notice that the density does only slightly influence the choice of the optimal q. The important parameter for determining the optimal q is the condition number of the matrices. This can be explained by the fact that the condition number is strongly connected with the spectrum of the matrix. We have already seen in the scalar case how the efficiency of different values of q varies with the (eigen-)value.

For well-conditioned matrices, we have a clear result. For $p = 1$, the optimal value is $q = 6$ and in all other cases, $q = 3$ is best. For ill-conditioned matrices,

Table 6.5: Results for $c = 500$, $p = 4$, $d = 0.003$.
Optimal values in bold.

time	#it	#mult	q
74.35	10	54	2
50.013	6	40	3
44.4	**5**	**39**	**4**
48.263	**5**	44	5
52.72	**5**	49	6

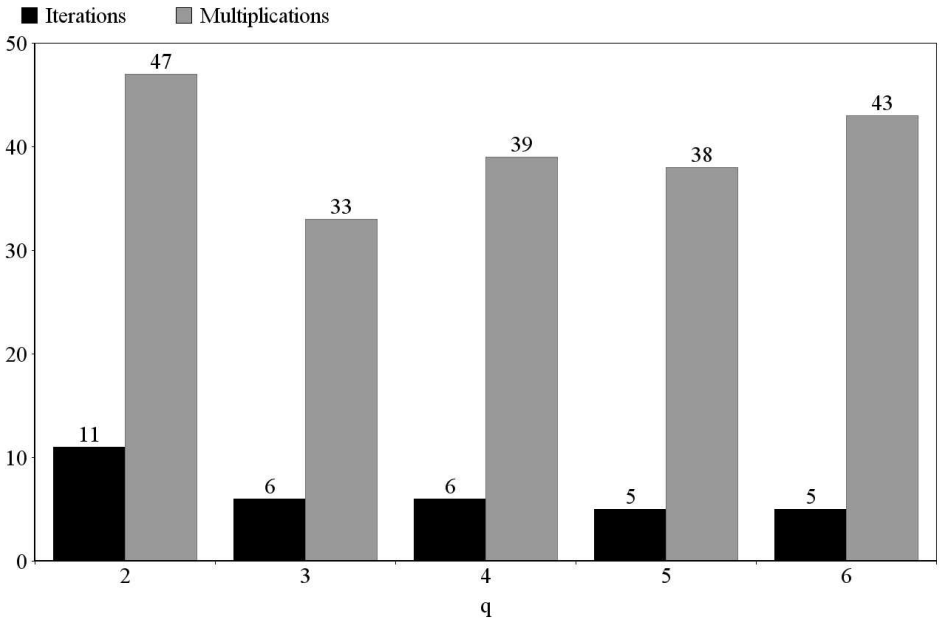

Figure 6.2: Number of iterations and multiplications for $c = 500$, $p = 3$, and $d = 0.003$.

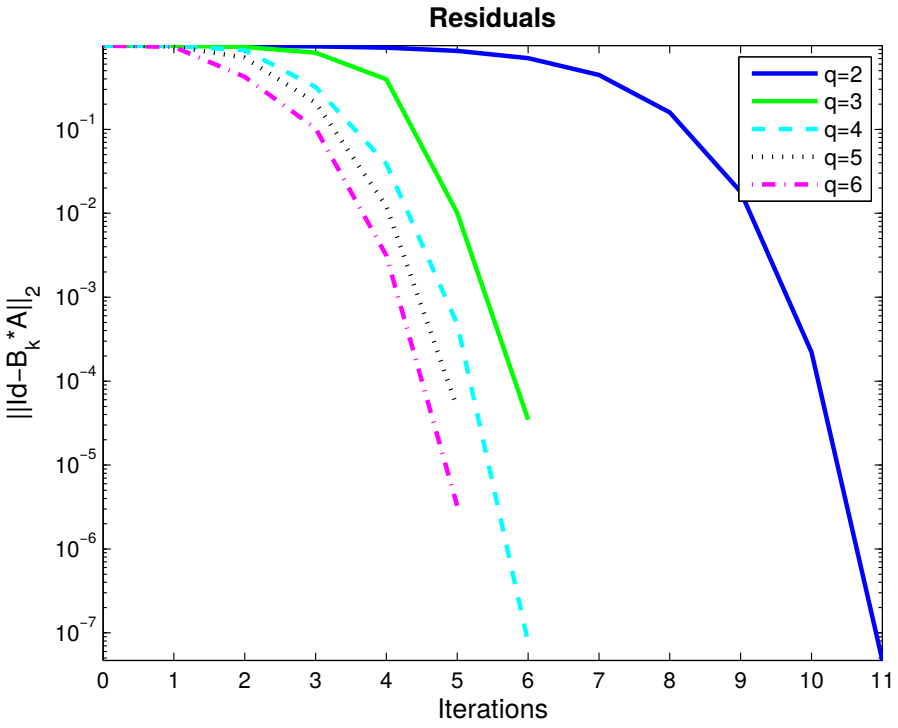

Figure 6.3: Residuals for $c = 500$, $p = 3$, and $d = 0.003$.

the results are ambiguous in a few cases, like already shown in Tables 6.3 and 6.4. For $p \in \{1, 2, 3\}$, we often have that a smaller q ($q = 3$) requires the lowest number of multiplications but a larger q ($q = 5$ or $q = 6$) requires less iterations and is therefore the better choice for the studied size of matrices. The complete results for all studied cases are presented in Table 6.6, where we point out both q in these ambiguous cases.

6.4.3 General Matrices and Applications

For the type of applications in chemistry and physics that are the subject of this thesis, the most interesting are sparse and ill-conditioned matrices with an arbitrary spectral radius. For obtaining a code that scales linearly with the number of rows/columns of the occurring matrices, it is crucial to have sparse matrices. For well-conditioned matrices, various methods for calculating their root already exist, as for example the expansion in Chebyshev polynomials for inverting matrices as described in Chapter 3 and Chapter 4. The problem of ill-conditioned matrices consists of a large quotient of its largest and smallest

Table 6.6: Results for $\rho(A) < 1$. Best q is outlined for different p and c.

c \ p	1	2	3	4	5
1.25	6	3	3	3	3
10	5	3	3/5	3	3
50	3/5	3	3/5	4	4
500	3/6	5	3	4	4

eigenvalue and the resulting issue of error propagation. We have already seen in the scalar case that the number of iteration steps needed until quadratic convergence is achieved as well as the total number of iteration steps varies significantly with the (eigen-)value.

As aforementioned, matrices coming from applications have generally a spectral radius that is larger than 1. This is in contrast to the fact that we dealt so far only with matrices A with $\rho(A) < 1$. In [12], Bini et al. proved the following

Proposition 1. *Suppose that all the eigenvalues of A are real and positive. The iteration* (6.3)

$$B_{k+1} = \frac{1}{p}\left[(p+1)B_k - B_k^{p+1}A\right],$$

with $B_0 = I$ converges to $A^{-1/p}$ if $\rho(A) < p+1$. If $\rho(A) = p+1$ the iteration does not converge to the inverse of any p-th root of A.

As already discussed, if we take $q = 2$ in (6.20), the iteration coincides with Bini's iteration (6.3). Numerical investigation provides that the analogue of Proposition 1 is also true for (6.20) in the case $q \neq 2$, apart from a few random matrices for $p = 1$ and varying q where the iteration converges. But this is so erratically occurring that we are not able to present a general rule.

However, we want to deal with matrices having a spectral radius larger than 1, as well. To present a solution to this problem, we assert first that for deal-

ing with general matrices A, one can either scale the matrix such that is has a spectral radius $\rho(A) < 1$ or choose the matrix B_0 such that $\|I - B_0^p A\| < 1$ is satisfied. We pursue the latter approach.

In Chapter 4, we dealt with the problem of inverting large, sparse matrices. For well-conditioned matrices, this problem was solved by Chebyshev polynomial expansion and for the occurring ill-conditioned matrices the Newton-Schulz iteration, so our iteration for $p = 1$ and $q = 2$, was used. As already explained, the Newton-Schulz iteration is in most of the cases not the optimal choice, so there is room for improvement. The N_q matrices that have to be inverted in Chapter 4 are all real and symmetric positive definite and have therefore only positive real eigenvalues. Thus, we can in principle apply iteration (6.20) but as the spectral radius of these matrices is larger than $p+1 = 2$, we need to be careful with the initial guess. For the Newton-Schulz iteration, we can take

$$B_0 = (\|A\|_1 \|A\|_\infty)^{-1} A^T \tag{6.30}$$

as start matrix, as Pan and Reif proved that then convergence is guaranteed [123]. We solve the problem for an arbitrary q by the following

Proposition 2. *Let A be a symmetric positive definite matrix with $\rho(A) \geq 1$ and B_0 like in equation* (6.30)*. Then, $\|I - B_0^p A\|_2 < 1$ is guaranteed.*

Proof. As A is symmetric positive definite, we have $\|A\|_2 = \lambda_{\max}$. We also remind the relationship $\|A\|_2 \leq \sqrt{\|A\|_1 \|A\|_\infty}$. Furthermore, we make use of Lemma 1 and have therefore just to show that

$$\|B_0^p A\|_2 = \|\left((\|A\|_1 \|A\|_\infty)^{-1} A^T\right)^p A\|_2 \leq 1. \tag{6.31}$$

As A is symmetric, we have $A^T = A$ and therefore commutativity of B_0 and A. Additionally, the relation

$$(\|A\|_1 \|A\|_\infty)^{-1} \leq \frac{1}{\lambda_{\max}^2} \tag{6.32}$$

holds. So we have

$$\begin{aligned} \|B_0^p A\|_2 &\leq \frac{1}{(\lambda_{\max}^2)^p} \|A^{p+1}\|_2 \leq \frac{1}{\lambda_{\max}^{2p}} \|A\|_2^{p+1} \\ &= \frac{1}{\lambda_{\max}^{2p}} \lambda_{\max}^{p+1} = \frac{1}{\lambda_{\max}^{p-1}} \leq 1, \end{aligned} \tag{6.33}$$

as we deal with matrices with $\rho(A) \geq 1$. □

Remark 3. *By replacing A^T with A^* in equation* (6.30)*,* Proposition 2 *is also true for Hermitian positive definite matrices.*

Consequently, the initial guess (6.30) fits also for the general case, and not only for the Newton-Schulz iteration. We have therefore convergence in (6.20) for any $p \geq 1$ and $q \geq 2$, regardless of the spectral radius.
Let us hereby note that the calculations are done in the spectral norm, as the guess by Pan and Reif works only for this norm. One issue is that in MATLAB, the spectral norm is not available for sparse matrices. Practical calculations show that the initial guess and use of the infinity norm leads also to convergence with the same number of iterations, and requires only about half of the computational time. But then, we can not guarantee that $\|I - B_0^p A\|_\infty < 1$, as requested in Theorem 3, and therefore the residuals do not form a monotonically decreasing sequence. Therefore, it might happen that the iteration diverges but this never occurred in practice.

We perform calculations using matrices A having the same densities and condition numbers as in the case $\rho(A) < 1$, except for $d = 0.001$. We scale the matrices such that we have $\rho(A) \in \{10, 50\}$ and use the initial guess by Pan and Reif (6.30). We observe also in the general case that the best q is independent of the density of the matrix and depends only on the condition number. Nevertheless, it is important to consider matrices of different densities to find the general rule. As an example, we study ill-conditioned matrices (c=500) with spectral radius $\rho(A) = 10$ for the case $p = 3$. It can be seen that the number of matrix-matrix multiplications is always lowest for the case $q = 5$. Also the computational time is lowest for the case $q = 5$. The number of iterations decreases further with q so that the lowest number of iterations is always achieved for the case $q = 6$. But this is not a better choice as we have to build more matrix-matrix products which is the more an issue, the larger the matrices are. The complete results can be found in table 6.8, where the optimal choices are highlighted in bold.

Table 6.7: Results for the case $p = 3$, $\rho(A) = 10$, $c = 500$.
Optimal values in bold.

d	time	#it	#mult	q
	406.29	55.333	224.33	2
	189.68	22.667	116.33	3
0.003	167.94	18	111	4
	152.86	15	**108**	**5**
	153.05	**14**	115	6
	417.93	55.667	225.67	2
	198.37	23.333	119.67	3
0.01	170.63	18	111	4
	155.33	15	**108**	**5**
	158.54	**14**	115	6
	429.66	57.333	232.33	2
	219.35	25	128	3
0.1	189.38	19	117	4
	179.09	16	**115**	**5**
	187.66	**15**	123	6
	529.29	61.667	249.67	2
	277.02	27	138	3
0.8	247.98	20.667	127	4
	235.74	17.667	**126.67**	**5**
	237.65	**16**	131	6

All the other cases can be studied in the same way. The decision is not always as clear as in the above considered case. An example is the calculation of the inverse 4-th root of matrices with spectral radius $\rho(A) = 50$ and condition number $c = 10$. We have hereby that for sparse matrices the choice $q = 5$ is optimal, but for more dense matrices $q = 6$ is the optimal value. However, $q = 6$ is also an acceptable choice for sparse matrices, so that we can choose $q = 6$ as the best option if we want a general rule. The complete results can

be found in Table 6.8, where the optimal choices are highlighted in bold.

Table 6.8: Results for the case $p = 4$, $\rho(A) = 50$, $c = 10$.
Optimal values in bold.

d	time	#it	#mult	q
	240.52	32.333	165.67	2
	154.84	18.333	114	3
0.003	132.94	14.333	104.33	4
	119.96	12	**100**	**5**
	120.04	**11**	103	6
	251.14	33.667	172.33	2
	161.69	19.333	120	3
0.01	140.19	15	109	4
	134.29	13	**108**	**5**
	131.08	**12**	112	6
	286.65	37	189	2
	187.38	20.667	128	3
0.1	167.03	16	116	4
	164.89	14	116	5
	161.47	**12.333**	**115**	**6**
	391.31	42.667	217.33	2
	258.66	24	148	3
0.8	224.17	18	**130**	4
	225.29	16	132	5
	216.07	**14**	**130**	**6**

By doing the complete analysis for all chosen densities, condition numbers, and values of p, we are able to formulate general results for both cases, $\rho(A) = 10$ and $\rho(A) = 50$. These are presented in Tables 6.9 and 6.10, respectively. We focused primarily on the number of matrix-matrix multiplication as this

Table 6.9: Results for $\rho(A) = 10$. Best q is outlined for different p and c.

c \ p	1	2	3	4	5
1.25	4	3	3	4	4
10	3	3	5	4	6
50	3	4	5	6	6
500	3	4	5	6	6

Table 6.10: Results for $\rho(A) = 50$. Best q is outlined for different p and c.

c \ p	1	2	3	4	5
1.25	3	3	3	4	4
10	3	3	5	6	6
50	3	4	5	6	6
500	3	3	5	6	6

is the most significant criterion. In the cases where the decision was not clear, like in the above presented example, we looked in detail at the time and the number of iterations. Fortunately, this happens only a few cases. Also for matrices with a larger spectral radius, we have no configuration where $q = 2$ is the best choice. On the contrary, as one can see in Tables 6.7 and 6.8, the evaluation of the inverse p-th root with $q = 2$ requires generally up to two times the computational time and number of matrix-matrix-multiplications, and up to three times the number of iterations, compared to the optimal choice of q. One notices that the tables look rather similar but are not identical. Generally, it can be said that the larger p is, the more profitable it is to choose a larger q. A broader classification for matrices with general spectral radii is future work.

6.4.4 Non-Commutative Matrices

The method presented here works well if the given matrix A and the start matrix B_0 commute, which can be easily achieved by using equation (6.30) as an initial guess. However, we note that the proof of Theorem 3 does not work for non-commutative matrices, as we can not make use of the binomial theorem. In the case of the calculation of inverse matrices, we have shown that our formula coincides with Altman's work. Altman however does not assume commutativity in his iteration (6.4). Thus, we have an iteration scheme that also works for the inversion of matrices that do not commute with the start matrix. This might be important in our linear scaling algorithm presented in Chapter 4, where we use the Newton-Schulz iteration with a start matrix that does not commute with the matrix that has to be inverted, but is very close to the true inverse. It remains for future work to investigate to what extent our method is improved when using our iteration scheme with a larger q.

6.5 Conclusion

We presented a new general iteration scheme to calculate the inverse p-th root of symmetric positive definite matrices. It includes as special cases the methods of Altman [4] and Bini [12]. The variable q that in Altman's work equals the order of convergence for the iterative inversion of matrices, represents here the order of expansion. We figured out that $q > 2$ does not lead to a higher order of convergence in the case $p \neq 1$. However, the iteration converges while less iterations and matrix-matrix multiplications are needed, as quadratic convergence is reached faster. The computational time and the number of matrix-matrix multiplications is up to two times lower, and the number of iterations is up to three times lower.

To decide which order of expansion is optimal for matrices with different densities and condition numbers, one has always to take into account the condition number of the matrix whose inverse p-th root is to be calculated. Also the parameter p is decisive, whereas the density of the given matrix only affects the results slightly. As we have seen in Tables 6.6, 6.9, and 6.10, $q = 2$ is not optimal for any of the studied configurations. Thus, we have described a considerable amelioration to the previously known methods.

Chapter 7

Conclusion and Outlook

In this thesis, we presented a linear scaling approach to Fermi operator expansion. The approach is hybrid for the inversion of sparse matrices, as we distinguished between well- and ill-conditioned matrices. For the well-conditioned matrices, the inversion can be done easily using Chebyshev expansion. For the ill-conditioned matrices, the Newton-Schulz iteration was employed. This method is not necessarily the optimal choice, as we have seen in Chapter 6, where we generalized the problem of finding inverse p-th roots of matrices. We suggest that there is still room for improvement in the calculations.
We presented a new iteration scheme to calculate the inverse p-th root of symmetric positive definite matrices. Our method is more efficient and more general than before known formulas, which emerge as special cases. The iteration converges with less iterations and matrix-matrix multiplications, as quadratic convergence is reached faster.

The efficiency and accuracy of our linear scaling approach has been illustrated by the application to liquid methane at planetary conditions. We used a cubic simulation cell with periodic boundary conditions containing of 1000 methane molecules and simulated the extreme conditions ($2000 - 8000$ K and $20 - 600$ GPa) of the middle ice layer of the giant gas planets Uranus and Neptune. At 4000 K and more, we found no evidence of diamond but large carbon clusters and ring-like carbon structures. We also detected molecular and atomic hydrogen as well as small hydrocarbons like ethane and propane. We so were able to explain the Voyager II fly-by measurements but are left with

the question as to why no diamond has been found. One possible explanation is that diamond formation requires even higher pressures.
Following the research line in this work, a future perspective is to apply our method to other interesting and relevant problems in chemistry and physics. One of the possible tasks is the large scale simulation of liquid water using our linear scaling method, which is a current work in progress.

Appendix A

Residuals

We present the residuals for two exemplary cases. We show that we enter the quadratic convergence after less iterations if we choose $q > 2$, as presented in Figures 6.1 and 6.3. Where we have quadratic convergence, is highlighted in bold.

Table A.1: Residuals for the scalar case $p = 1$ and $\lambda = 10^{-6}$

q	$\lvert b_k - (1/\lambda)^{1/p}\rvert$	$\lvert 1 - b_k^p \lambda\rvert$	#it
	9.999800e+04	9.999800e−01	
	9.999600e+04	9.999600e−01	
	9.999200e+04	9.999200e−01	
	9.998400e+04	9.998400e−01	
	9.996800e+04	9.996800e−01	
	9.993602e+04	9.993602e−01	
	9.987208e+04	9.987208e−01	
	9.974433e+04	9.974433e−01	
	9.948931e+04	9.948931e−01	
	9.898122e+04	9.898122e−01	
2	9.797282e+04	9.797282e−01	
	9.598673e+04	9.598673e−01	
	9.213453e+04	9.213453e−01	
	8.488771e+04	8.488771e−01	
	7.205924e+04	7.205924e−01	

Table A.1 – Continued from previous page

q	$\lvert b_k - (1/\lambda)^{1/p}\rvert$	$\lvert 1 - b_k^p\lambda\rvert$	#it
	5.192534e+04	5.192534e−01	
	2.696241e+04	2.696241e−01	
	7.269715e+03	7.269715e−02	
	5.284876e+02	5.284876e−03	20
	2.792991e+00	2.792991e−**05**	
	7.800796e−05	7.800797e−**10**	
	1.455192e−11	0.000000e+00	
	9.999700e+04	9.999700e−01	
	9.999100e+04	9.999100e−01	
	9.997300e+04	9.997300e−01	
	9.991903e+04	9.991903e−01	
	9.975729e+04	9.975729e−01	
	9.927365e+04	9.927365e−01	
3	9.783673e+04	9.783673e−01	
	9.364957e+04	9.364957e−01	
	8.213294e+04	8.213294e−01	
	5.540541e+04	5.540541e−01	
	1.700813e+04	1.700813e−01	12
	4.920050e+02	4.920050e−**03**	
	1.190991e−02	1.190991e−**07**	
	1.455192e−11	0.000000e+00	
	9.999600e+04	9.999600e−01	
	9.998400e+04	9.998400e−01	
	9.993602e+04	9.993602e−01	
	9.974433e+04	9.974433e−01	
	9.898122e+04	9.898122e−01	
4	9.598673e+04	9.598673e−01	
	8.488771e+04	8.488771e−01	
	5.192534e+04	5.192534e−01	9
	7.269715e+03	7.269715e−**02**	
	2.792991e+00	2.792991e−**05**	

Table A.1 – Continued from previous page

q	$\lvert b_k - (1/\lambda)^{1/p}\rvert$	$\lvert 1 - b_k^p\lambda\rvert$	#it
	0.000000e+00	1.110223e−**16**	
	9.999500e+04	9.999500e−01	
	9.997500e+04	9.997500e−01	
	9.987508e+04	9.987508e−01	
	9.937695e+04	9.937695e−01	
5	9.692331e+04	9.692331e−01	
	8.553447e+04	8.553447e−01	
	4.578316e+04	4.578316e−01	8
	2.011540e+03	2.011540e−**02**	
	3.293392e−04	3.293392e−**09**	
	0.000000e+00	1.110223e−**16**	
	9.999400e+04	9.999400e−01	
	9.996401e+04	9.996401e−01	
	9.978423e+04	9.978423e−01	
	9.871236e+04	9.871236e−01	
6	9.251861e+04	9.251861e−01	
	6.271545e+04	6.271545e−01	7
	6.084814e+03	6.084814e−**02**	
	5.075559e−03	5.075559e−**08**	
	0.000000e+00	1.110223e−**16**	
	9.999300e+04	9.999300e−01	
	9.995101e+04	9.995101e−01	
	9.965759e+04	9.965759e−01	
7	9.762758e+04	9.762758e−01	
	8.452940e+04	8.452940e−01	
	3.083574e+04	3.083574e−01	7
	2.650820e+01	2.650820e−**04**	
	0.000000e+00	1.110223e−**16**	
	9.999200e+04	9.999200e−01	
	9.993602e+04	9.993602e−01	
	9.948931e+04	9.948931e−01	

Table A.1 – Continued from previous page

q	$\lvert b_k - (1/\lambda)^{1/p}\rvert$	$\lvert 1 - b_k^p\lambda\rvert$	#it
8	9.598673e+04	9.598673e−01	
	7.205924e+04	7.205924e−01	6
	7.269715e+03	7.269715e−**02**	
	7.800797e−05	7.800798e−**10**	

Table A.2: Residuals for the matrix case $p = 3$, $d = 0.003$, and $c = 500$

q	Residuals	#it
	9.980000e−01	
	9.952664e−01	
	9.888193e−01	
	9.737193e−01	
	9.389248e−01	
2	8.617598e−01	
	7.051328e−01	
	4.445128e−01	
	1.591981e−01	8
	1.811528e−**02**	
	2.205411e−**04**	
	4.700491e−**08**	
	9.980000e−01	
	9.907740e−01	
	9.579903e−01	
3	8.196543e−01	
	3.947808e−01	4
	1.011714e−**02**	
	3.476818e−**05**	
	9.980000e−01	
	9.840957e−01	
	8.786780e−01	

Table A.2 – Continued from previous page

q	Residuals	#it
4	3.199332e−01	3
	3.883428e−**02**	
	4.796801e−**04**	
	7.665675e−**08**	
5	9.980000e−01	3
	9.748093e−01	
	7.125061e−01	
	2.083850e−01	
	1.209819e−**02**	
	4.814179e−**05**	
6	9.980000e−01	3
	9.624980e−01	
	4.222389e−01	
	1.024713e−01	
	3.140923e−**03**	
	3.277028e−**06**	

Bibliography

[1] A. Alavi and D. Frenkel. Grand-canonical simulations of solvated ideal fermions. Evidence for phase separation. *J. Chem. Phys.*, 97(12):9249–9257, 1992.

[2] A. Alavi, J. Kohanoff, M. Parrinello, and D. Frenkel. Ab initio molecular dynamics with excited electrons. *Phys. Rev. Lett.*, 73(19):2599–2602, 1994.

[3] D. Alfè. Ab initio molecular dynamics, a simple algorithm for charge extrapolation. *Comp. Phys. Comm.*, 118(1):31–33, 1999.

[4] M. Altman. An optimum cubically convergent iterative method of inverting a linear bounded operator in Hilbert space. *Pacific J. Math.*, 10(4):1107–1113, 1960.

[5] F. Ancilotto, G. L. Chiarotti, S. Scandolo, and E. Tosatti. Dissociation of methane into hydrocarbons at extreme (planetary) pressure and temperature. *Science*, 275(5304):1288–1290, 1997.

[6] T. A. Arias, M. C. Payne, and J. D. Joannopoulos. Ab initio molecular dynamics: Analytically continued energy functionals and insights into iterative solutions. *Phys. Rev. Lett.*, 69(7):1077–1080, 1992.

[7] S. Azadi and T. D. Kühne. Absence of metallization in solid molecular hydrogen. *JETP Lett.*, 95(9):449–453, 2012.

[8] R. Baer and M. Head-Gordon. Chebyshev expansion methods for electronic structure calculations on large molecular systems. *J. Chem. Phys.*, 23(23):10003, 1997.

[9] S. Baroni and P. Giannozzi. Towards very large-scale electronic-structure calculations. *Europhys. Lett.*, 17(6):547, 1992.

[10] A. Ben-Israel. A note on an iterative method for generalized inversion of matrices. *Math. Comput.*, 20:439–440, 1966.

[11] L. R. Benedetti, J. H. Nguyen, W. A. Caldwell, H. Liu, M. Kruger, and R. Jeanloz. Dissociation of CH_4 at high pressures and temperatures: Diamond formation in giant planet interiors? *Science*, 286(5437):100–102, 1999.

[12] D. A. Bini, N. J. Higham, and B. Meini. Algorithms for the matrix pth root. *Numerical Algorithms*, 39(4):349–378, 2005.

[13] F. Bloch. Über die Quantenmechanik der Elektronen in Kristallgittern. *Zeitschrift f. Physik*, 52(7-8):555–600, 1929.

[14] M. Born and J. R. Oppenheimer. Zur Quantentheorie der Molekeln. *Annalen der Physik*, 84(4):457–484, 1927.

[15] D. R. Bowler, M. Aoki, C. M. Goringe, A. P. Horsfield, and D. G. Pettifor. A comparison of linear scaling tight binding methods. *Mat. Sci. Eng.*, 5:199, 1997.

[16] D. R. Bowler and T. Miyazaki. $\mathcal{O}(N)$ methods in electronic structure calculations. *Rep. Prog. Phys.*, 75(3):036503, 2012.

[17] D. R. Bowler, T. Miyazaki, and M. J. Gillan. Recent progress in linear scaling ab initio electronic structure techniques. *J. Phys. Cond. Mat.*, 14(11):2781, 2002.

[18] V. Brázdová and D. R. Bowler. Automatic data distribution and load balancing with space-filling curves: Implementation in CONQUEST. *J. Phys. Cond. Mat.*, 20(27):275223, 2008.

[19] A. Canning, G. Galli, F. Mauri, A. D. Vita, and R. Car. O(N) tight-binding molecular dynamics on massively parallel computers: an orbital decomposition approach. *Comp. Phys. Comm.*, 94:89–102, 1996.

[20] K. Capelle. A bird's-eye view of density-functional theory. *Braz. J. Phys.*, 36(4):1318–1343, 2006.

[21] R. Car and M. Parrinello. Unified approach for molecular dynamics and density-functional theory. *Phys. Rev. Lett.*, 55(22):2471–2474, 1985.

[22] S. Caravati, M. Bernasconi, T. D. Kühne, M. Krack, and M. Parrinello. Coexistence of tetrahedral- and octahedral-like sites in amorphous phase change materials. *Appl. Phys. Lett.*, 91(17):171906–171906–3, 2007.

[23] S. Caravati, M. Bernasconi, T. D. Kühne, M. Krack, and M. Parrinello. First-principles study of crystalline and amorphous $Ge_2Sb_2Te_5$ and the effects of stoichiometric defects. *J. Phys. Cond. Mat.*, 21(25):255501, 2009.

[24] S. Caravati, M. Bernasconi, T. D. Kühne, M. Krack, and M. Parrinello. Unravelling the mechanism of pressure induced amorphization of phase change materials. *Phys. Rev. Lett.*, 102(20):205502, 2009.

[25] S. Caravati, D. Colleoni, R. Mazzarello, T. D. Kühne, M. Krack, M. Bernasconi, and M. Parrinello. First-principles study of nitrogen doping in cubic and amorphous $Ge_2Sb_2Te_5$. *J. Phys. Cond. Mat.*, 23(26):265801, 2011.

[26] D. M. Ceperley. Path integrals in the theory of condensed helium. *Rev. Mod. Phys.*, 67(2):279–355, 1995.

[27] M. Ceriotti, T. D. Kühne, and M. Parinello. An efficient and accurate decomposition of the Fermi operator. *J. Chem. Phys.*, 129(2):024707, 2008.

[28] M. Ceriotti, T. D. Kühne, and M. Parinello. A hybrid approach to Fermi operator expansion. *AIP Conf. Proc.*, 1148(1):658–661, 2009.

[29] B. Conrath, F. M. Flasar, R. Hanel, V. Kunde, W. Maguire, J. Pearl, J. Pirraglia, R. Samuelson, P. Gierasch, A. Weir, B. Bezard, D. Gautier, D. Cruikshank, L. Horn, R. Springer, and W. Shaffer. Infrared observations of the neptunian system. *Science*, 246(4936):1454–1459, 1989.

[30] C. S. Cucinotta, G. Miceli, P. Raiteri, M. Krack, T. D. Kühne, M. Bernasconi, and M. Parrinello. Superionic conduction in substoichiometric LiAl alloy: An ab initio study. *Phys. Rev. Lett.*, 103(12):125901, 2009.

[31] J. Dai and J. Yuan. Large-scale efficient Langevin dynamics, and why it works. *Europhys. Lett.*, 88(2):20001, 2009.

[32] M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai, and G. Seifert. Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Phys. Rev. B*, 58(11):7260–7268, 1998.

[33] M. Farnesi Camellone, T. D. Kühne, and D. Passerone. Density functional theory study of self-trapped holes in disordered SiO_2. *Phys. Rev. B*, 80(3):033203, 2009.

[34] W. M. C. Foulkes and R. Haydock. Tight-binding models and density-functional theory. *Phys. Rev. B*, 39(17):12520–12536, 1989.

[35] G. Galli and M. Parrinello. Large scale electronic structure calculations. *Phys. Rev. Lett.*, 69(24):3547–3450, 1992.

[36] W. Gautschi. *Numerical Analysis: An Introduction*. Birkhäuser, 1997.

[37] S. Gershgorin. Über die Abgrenzung der Eigenwerte einer Matrix. *Izv. Akad. Nauk USSR Otd. Fiz.-Mat. Nauk*, 6:749–754, 1931.

[38] L. M. Ghiringhelli, C. Valeriani, J. H. Los, E. J. Meijer, A. Fasolino, and D. Frenkel. State-of-the-art models for the phase diagram of carbon and diamond nucleation. *Mol. Phys.*, 106(16-18):2011–2038, 2008.

[39] A. Gil, J. Segura, and N. M. Temme. *Numerical methods for special functions*. Siam, 2007.

[40] S. Goedecker. Linear scaling electronic structure methods. *Phys. Rev. Lett.*, 73(1):122–125, 1994.

[41] S. Goedecker. Linear scaling electronic structure methods. *Rev. Mod. Phys.*, 71(4):1085–1123, 1999.

[42] S. Goedecker and L. Colombo. Efficient linear scaling algorithm for tight-binding molecular dynamics. *Phys. Rev. Lett.*, 73(1):122–125, 1994.

[43] S. Goedecker and M. Teter. Tight-binding electronic-structure calculations and tight-binding molecular dynamics with localized orbitals. *Phys. Rev. B*, 51(15):9455–9464, 1995.

[44] C. M. Goringe, D. R. Bowler, and E. Hernàndez. Tight-binding modelling of materials. *Rep. Prog. Phys.*, 60(12):1447, 1997.

[45] C.-H. Guo. On Newton's method and Halley's method for the principal pth root of a matrix. *Linear Algebra Appl.*, 432(8):1905–1922, 2010.

[46] C.-H. Guo and N. J. Higham. A Schur-Newton method for the matrix p'th roots and its inverse. *Siam J. Matrix Anal. Appl.*, 28(3):788–804, 2006.

[47] E. Halley. Methodus nova, accurata facilis inveniendi radices aequationum quarumcumque generaliter, sine praevia reductione. *Trans. Roy. Soc. London*, 18(207-214):136–148, 1694.

[48] J. Harris. Simplified method for calculating the energy of weakly interacting fragments. *Phys. Rev. B*, 31(4):1770–1779, 1985.

[49] P. Haynes and M. Payne. Failure of density-matrix minimization methods for linear-scaling density-functional theory using the Kohn penalty-functional. *Solid State Communications*, 108(10):737–741, 1998.

[50] P. D. Haynes and M. C. Payne. Corrected penalty-functional method for linear-scaling calculations within density-functional theory. *Phys. Rev. B*, 59(19):12173–12176, 1999.

[51] R. Helled, J. D. Anderson, M. Podolak, and G. Schubert. Interior models of Uranus and Neptune. *Astrophys. J.*, 726(1):15–21, 2011.

[52] J. M. Herbert and M. Head-Gordon. Accelerated, energy-conserving Born-Oppenheimer molecular dynamics via Fock matrix extrapolation. *Phys. Chem. Chem. Phys.*, 7(18):3269–3275, 2005.

[53] N. J. Higham. *Functions of Matrices : Theory and Computation*. SIAM, 2008.

[54] N. J. Higham and A. H. Al-Mohy. Computing matrix functions. *Acta Numerica*, 19:159–208, 2010.

[55] R. Hoffmann. An extended Hückel theory. I. Hydrocarbons. *J. Chem. Phys.*, 39(6):1397–1412, 1963.

[56] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136(3B):B864–B871, 1964.

[57] A. Horsfield, P. D. Godwin, D. G. Pettifor, and A. P. Sutton. Computational materials synthesis. I. A tight-binding scheme for hydrocarbons. *Phys. Rev. B*, 54(22):15773–15775, 1996.

[58] W. D. Hoskins and D. J. Walton. A faster, more stable method for computing the pth roots of positive definite matrices. *Linear Algebra Appl.*, 26:139–163, 1979.

[59] https://cmsportal.caspur.it/index.php/CMPTool.

[60] W. B. Hubbard. Interiors of the giant planets. *Science*, 214(4517):145–149, 1981.

[61] W. B. Hubbard, W. J. Nellis, A. Mitchell, N. C. Holmes, S. S. Limaye, and P. C. McCandless. Interior structure of Neptune: Comparison with Uranus. *Science*, 253(5020):648–651, 1991.

[62] J. Hutter. Car-Parrinello molecular dynamics. *WIREs Comput. Mol. Sci.*, 2(4):604–612, 2012.

[63] B. Iannazzo. On the Newton method for the matrix pth root. *Siam J. Matrix Anal. Appl.*, 28(2):503–523, 2006.

[64] B. Iannazzo. A family of rational iterations and its application to the computation of the matrix pth root. *Siam J. Matrix Anal. Appl.*, 30(4):1445–1462, 2008.

[65] S. Itoh, P. Ordejón, and R. M. Martin. Order-N tight-binding molecular dynamics on parallel computers. *Comp. Phys. Comm.*, 88(2-3):173–185, 1995.

[66] R. O. Jones and O. Gunnarsson. The density functional formalism, its applications and prospects. *Rev. Mod. Phys.*, 61(3):689–746, 1989.

[67] R. Z. Khaliullin, H. Eshet, T. D. Kühne, J. Behler, and M. Parrinello. Graphite-diamond phase coexistence study employing a neural-network mapping of the ab initio potential energy surface. *Phys. Rev. B*, 81(10):100103(R), 2010.

[68] R. Z. Khaliullin, H. Eshet, T. D. Kühne, J. Behler, and M. Parrinello. Nucleation mechanism for the direct graphite-to-diamond phase transition. *Nature Mat.*, 10(9):693–697, 2011.

[69] R. Z. Khaliullin and T. D. Kühne. Microscopic properties of liquid water from combined ab initio molecular dynamics and energy decomposition studies. *Phys. Chem. Chem. Phys*, 15(38):15746–15766, 2013.

[70] J. Kim, F. Mauri, and G. Galli. Total-energy global optimizations using nonorthogonal localized orbitals. *Phys. Rev. B*, 52(3):1640–1648, 1995.

[71] J. B. Kogut. The lattice gauge theory approach to quantum chromodynamics. *Rev. Mod. Phys.*, 55(3):775–836, 1983.

[72] J. Kohanoff. *Electronic structure calculations for solids and molecules.* Cambridge University Press, 2006.

[73] W. Kohn. Density functional and density matrix method scaling linearly with the number of atoms. *Phys. Rev. Lett.*, 76(17):3168–3171, 1996.

[74] W. Kohn. Nobel lecture: Electronic structure of matter-wave functions and density functionals. *Rev. Mod. Phys.*, 71(5):1253–1266, 1999.

[75] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140(4A):A1133–A1138, 1965.

[76] J. Kolafa. Time-reversible always stable predictor corrector method for molecular dynamics of polarizable molecules. *J. Comp. Chem.*, 25(3):335, 2004.

[77] J. Kolafa. Gear formalism of the always stable predictor-corrector method for molecular dynamics of polarizable molecules. *J. Chem. Phys.*, 122(16):164105, 2005.

[78] F. R. Krajewski and M. Parrinello. Stochastic linear scaling for metals and nonmetals. *Phys. Rev. B*, 71(23):233105, 2005.

[79] F. R. Krajewski and M. Parrinello. Linear scaling electronic structure calculations and accurate statistical mechanics sampling with noisy forces. *Phys. Rev. B*, 73(4):041105, 2006.

[80] F. R. Krajewski and M. Parrinello. Linear scaling for quasi-one-dimensional systems. *Phys. Rev. B*, 74(12):125107, 2006.

[81] J. D. Kress, S. R. Bickham, L. A. Collins, B. L. Holian, and S. Goedecker. Tight-binding molecular dynamics of shock waves in methane. *Phys. Rev. Lett.*, 83(19):3896–3899, 1999.

[82] J. D. Kress, S. Goedecker, A. Hoisie, H. Wasserman, O. Lubeck, L. A. Collins, and B. L. Holian. Parallel O(N) tight-binding molecular dynamics of polyetylene and compressed methane. *J. Comp.-Aided Mat. Design*, 5(2-3):295–316, 1998.

[83] T. D. Kühne. Second generation Car-Parrinello molecular dynamics. *WIREs Comput. Mol. Sci.*, 2013.

[84] T. D. Kühne and R. Z. Khaliullin. Electronic signature of the instantaneous asymmetry in the first coordination shell of liquid water. *Nature Comm.*, page 1450, 2013.

[85] T. D. Kühne, M. Krack, F. R. Mohamed, and M. Parrinello. Efficient and accurate Car-Parrinello-like approach to Born-Oppenheimer molecular dynamics. *Phys. Rev. Lett.*, 98(6):066401, 2007.

[86] T. D. Kühne, M. Krack, and M. Parrinello. Static and dynamical properties of liquid water from first principles by a novel Car-Parrinello-like approach. *J. Chem. Theory Comput.*, 5:235–241, 2009.

[87] T. D. Kühne, T. A. Pascal, E. Kaxiras, and Y. Jung. New insights into the structure of the vapor/water interface from large-scale first-principles simulations. *J. Phys. Chem. Lett.*, 2:105–113, 2011.

[88] S. Lakić. On the computation of the matrix k-th root. *Z. Angew. Math. Mech.*, 78(3):167–172, 1998.

[89] M. Levy. Universal variational functionals of electron densities, first-order density matrices, and natural spin-orbitals and solution of the v-representability problem. *Proc. Nat. Acad. Sci.*, 76(12):6062–6065, 1979.

[90] X.-P. Li, R. W. Nunes, and D. Vanderbilt. Density-matrix electronic-structure method with linear system-size scaling. *Phys. Rev. B*, 47(16):10891–10894, 1993.

[91] W. Liang, R. Baer, C. Saravanan, Y. Shao, A. T. Bell, and M. Head-Gordon. Fast methods for resumming matrix polynomials and Chebyshev matrix polynomials. *J. Comp. Phys.*, 194(2):575–587, 2004.

[92] W. Liang, C. Saravanan, Y. Shao, R. Baer, A. T. Bell, and M. Head-Gordon. Improved Fermi operator expansion methods for fast electronic structure calculations. *J. Chem. Phys.*, 119(8):4117–4125, 2003.

[93] E. H. Lieb. Density functionals for Coulomb systems. *Int. J. Quantum Chem.*, 24(3):243–277, 1983.

[94] L. Lin, J. Lu, R. Car, and W. E. Multipole representation of the Fermi operator with application to the electronic structure analysis of metallic systems. *Phys. Rev. B*, 79(11):115133, 2009.

[95] L. Lin, J. Lu, L. Ying, and W. E. Pole-based approximation of the Fermi-Dirac function. *Chin. Ann. Math. B*, 30(6):729–742, 2009.

[96] J. H. Los, T. D. Kühne, S. Gabardi, and M. Bernasconi. First principles simulation of amorphous InSb. *Phys. Rev. B*, 87(18):184201, 2013.

[97] J. H. Los, T. D. Kühne, S. Gabardi, and M. Bernasconi. First-principles study of the amorphous In_3SbTe_2 phase change compound. *Phys. Rev. B*, 88(17):174203, 2013.

[98] P.-O. Löwdin. On the non-orthogonality problem connected with the use of atomic wave functions in the theory of molecules and crystals. *J. Chem. Phys.*, 18(3):365–375, 1950.

[99] G. A. Ludueña, T. D. Kühne, and D. Sebastiani. Mixed Grotthuss and vehicle transport mechanism in proton conducting polymers from ab initio molecular dynamics simulations. *Chemistry of Materials*, 23(6):1424–1429, 2011.

[100] M. P. Marder. *Condensed Matter Physics*. WILEY-VCH, 2010.

[101] R. M. Martin. *Electronic Structure: Basic Theory and Practical Methods (Vol 1)*. Cambridge University Press, 2004.

[102] D. Marx and J. Hutter. *Ab Initio Molecular Dynamics*. Cambridge University Press, 2009.

[103] MATLAB and Statistics Toolbox Release 2013a. *Version 8.1*. The MathWorks Inc., Natick, Massachusetts, 2013.

[104] F. Mauri and G. Galli. Electronic-structure calculations and molecular-dynamics simulations with linear system-size scaling. *Phys. Rev. B*, 50:4316–4326, 1994.

[105] F. Mauri, G. Galli, and R. Car. Orbital formulation for electronic-structure calculations with linear system-size scaling. *Phys. Rev. B*, 47(15):9973–9976, 1993.

[106] R. McWeeny. Some recent advances in density matrix theory. *Rev. Mod. Phys.*, 32(2):335–369, 1960.

[107] S. Meloni, M. Rosati, A. Federico, L. Ferraro, A. Mattoni, and L. Colombo. Computational materials science application programming interface (CMSapi): A tool for developing applications for atomistic simulations. *Comp. Phys. Comm.*, 169(1-3):462–466, 2005.

[108] N. D. Mermin. Thermal properties of the inhomogeneous electron gas. *Phys. Rev.*, 137(5A):A1441–A1443, 1965.

[109] C. Moler and C. V. Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Rev.*, 45(1):3, 2003.

[110] I. Montvay and G. Münster. *Quantum Fields on a Lattice*. Cambridge Monographs on Mathematical Physics, 1994.

[111] M. A. Morales, C. Pierleoni, E. Schwegler, and D. M. Ceperley. Evidence for a first-order liquid-liquid transition in high-pressure hydrogen from ab initio simulations. *Proc. Nat. Acad. Sci.*, 107(29):12799–12803, 2010.

[112] W. J. Nellis, D. C. Hamilton, N. C. Holmes, H. B. Radousky, F. H. Ree, A. C. Mitchell, and M. Nicol. The nature of the interior of Uranus based on studies of planetary ices at high dynamic pressure. *Science*, 240(4853):779–781, 1988.

[113] W. J. Nellis, D. C. Hamilton, and A. C. Mitchell. Electrical conductivities of methane, benzene, and polybutene shock compressed to 60 GPa (600 kbar). *J. Chem. Phys.*, 115(2):1015–1019, 2001.

[114] W. J. Nellis, F. H. Ree, M. van Thiel, and A. C. Mitchell. Shock compression of liquid carbon monoxide and methane to 90 GPa (900 kbar). *J. Chem. Phys.*, 75(6):3055–3063, 1981.

[115] W. J. Nellis, S. T. Weir, and A. C. Mitchell. Minimum metallic conductivity of fluid hydrogen at 140 GPa (1.4 mbar). *Phys. Rev. B*, 59(5):3434–3449, 1999.

[116] N. F. Ness, M. H. Acuña, K. W. Behannon, L. F. Burlaga, J. E. P. Connerney, R. P. Lepping, and F. M. Neubauer. Magnetic fields at Uranus. *Science*, 233(4759):85, 1986.

[117] N. F. Ness, M. H. Acuña, L. F. Burlaga, J. E. P. Connerney, R. P. Lepping, and F. M. Neubauer. Magnetic fields at Neptune. *Science*, 246(4936):1473, 1989.

[118] O. H. Nielsen and R. M. Martin. First-principles calculation of stress. *Phys. Rev. Lett.*, 50(9):697–700, 1983.

[119] A. M. N. Niklasson. Extended Born-Oppenheimer molecular dynamics. *Phys. Rev. Lett.*, 100(12):123004, 2008.

[120] P. Ordejón, D. A. Drabold, M. P. Grumbach, and R. M. Martin. Unconstrained minimization approach for electronic computations that scales linearly with system size. *Phys. Rev. B*, 48(19):14646–14649, 1993.

[121] P. Ordejón, D. A. Drabold, R. M. Martin, and M. P. Grumbach. Linear system-size scaling methods for electronic-structure calculations. *Phys. Rev. B*, 51(3):1456–1476, 1995.

[122] A. H. R. Palser and D. E. Manolopoulos. Canonical purification of the density matrix in electronic-structure theory. *Phys. Rev. B*, 58(19):12704–12711, 1998.

[123] V. Pan and J. Reif. Efficient parallel solution of linear systems. *Proceedings of the 17th Annual ACM Symposium on the Theory of Computing, New York*, pages 143–152, 1985.

[124] D. A. Papaconstantopoulos and M. J. Mehl. The Slater-Koster tight-binding method: a computationally efficient and accurate approach. *J. Phys. Cond. Mat.*, 15(10):R413–R440, 2003.

[125] T. A. Pascal, D. Schärf, Y. Jung, and T. D. Kühne. On the absolute thermodynamics of water from computer simulations: A comparison of first-principles molecular dynamics, reactive and empirical force fields. *J. Chem. Phys.*, 137(24):244507, 2012.

[126] J. P. Perdew and K. Schmidt. Jacob's ladder of density functional approximations for the exchange-correlation energy. *AIP Conf. Proc.*, 577(1):1–20, 2001.

[127] W. V. Petryshyn. On the inversion of matrices and linear operators. *Proc. AMS*, 16(5):893–901, 1965.

[128] M. Podolak, J. I. Podolak, and M. Marley. Further investigations of random models of Uranus and Neptune. *Planet. Space Sci.*, 48(2-3):143–151, 2000.

[129] M. Podolak, A. Weizman, and M. Marley. Comparative models of Uranus and Neptune. *Planet. Space Sci.*, 43(12):1517–1522, 1995.

[130] E. Prodan and W. Kohn. Nearsightedness of electronic matter. *PNAS*, 102(33):11635–11638, 2005.

[131] P. J. Psarrakos. On the mth roots of a complex matrix. *Electronic Journal of Linear Algebra*, 9:32–41, 2002.

[132] P. Pulay and G. Fogarasi. Fock matrix dynamics. *Chem. Phys. Lett.*, 386(4-6):272–278, 2004.

[133] F. H. Ree. Systematics of high-pressure and high-temperature behavior of hydrocarbons. *J. Chem. Phys.*, 70(2):974–983, 1979.

[134] S. Reich, C. Thomsen, and J. Maultzsch. *Carbon Nanotubes: Basic Concepts and Physical Properties*. WILEY-VCH, 2004.

[135] D. Richters and T. D. Kühne. Liquid methane at extreme temperature and pressure: Implications for models of Uranus and Neptune. *JETP Lett.*, 97(4):184–187, 2013.

[136] D. Richters and T. D. Kühne. Self-consistent field theory based molecular dynamics with linear system-size scaling. *J. Chem. Phys.*, 140(13):134109, 2014.

[137] M. Ross. The ice layer in Uranus and Neptune – diamonds in the sky? *Nature*, 292(5822):435–436, 1981.

[138] S. Scandolo. Liquid-liquid phase transition in compressed hydrogen from first-principles simulations. *Proc. Nat. Acad. Sci.*, 100(6):3051–3053, 2003.

[139] G. Schulz. Iterative Berechung der reziproken Matrix. *Z. Angew. Math. Mech.*, 13(1):57–59, 1933.

[140] G. Seifert and J.-O. Joswig. Density-functional tight binding – an approximate density-functional theory method. *WIREs Comput. Mol. Sci.*, 2(3):456–465, 2012.

[141] P. L. Silvestrelli, A. Alavi, M. Parrinello, and D. Frenkel. Ab initio molecular dynamics simulation of laser melting of silicon. *Phys. Rev. Lett.*, 77(15):3149–3152, 1996.

[142] P. L. Silvestrelli, A. Alavi, M. Parrinello, and D. Frenkel. Structural, dynamical, electronic, and bonding properties of laser-heated silicon: An ab initio molecular-dynamics study. *Phys. Rev. B*, 56:3806–3812, 1997.

[143] J. C. Slater and G. F. Koster. Simplified LCAO method for the periodic potential problem. *Phys. Rev.*, 94(6):1498–1524, 1954.

[144] M. Smith. A Schur algorithm for computing marix pth roots. *Siam J. Matrix Anal. Appl.*, 24(4):971–989, 2003.

[145] L. Spanu, D. Donadio, D. Hohl, E. Schwegler, and G. Galli. Stability of hydrocarbons at deep earth pressures and temperatures. *PNAS*, 108(17):6843–6846, 2011.

[146] A. P. Sutton, M. W. Finnis, D. G. Pettifor, and Y. Ohta. The tight-binding bond model. *J. Chem. Phys.*, 21(1):35–66, 1988.

[147] I. Tamblyn and S. A. Bonev. Structure and phase boundaries of compressed liquid hydrogen. *Phys. Rev. Lett.*, 104(6):065702, 2010.

[148] C. Van Loan. A note on the evaluation of matrix polynomials. *Automatic Control, IEEE Transactions on*, 24(2):320–321, 1979.

[149] J. VandeVondele, M. Krack, F. Mohamed, M. Parrinello, T. Chassaing, and J. Hutter. Quickstep: Fast and accurate density functional calculations using a mixed gaussian and plane waves approach. *Comp. Phys. Comm.*, 167(2):103–128, 2005.

[150] J. W. Warwick, D. R. Evans, G. R. Peltzer, R. G. Peltzer, J. H. Romig, C. B. Sawyer, A. C. Riddle, A. E. Schweitzer, M. D. Desch, M. L. Kaiser, W. M. Farrell, T. D. Carr, I. de Pater, D. H. Staelin, S. Gulkis, R. L. Poynter, A. Boischot, F. Genova, Y. Leblanc, A. Lecacheux, B. M. Pedersen, and P. Zarka. Voyager planetary radio astronomy at Neptune. *Science*, 246(4936):1498–1501, 1989.

[151] J. W. Warwick, D. R. Evans, J. H. Romig, C. B. Sawyer, M. D. Desch, M. L. Kaiser, J. K. Alexander, T. D. Carr, D. H. Staelin, S. Gulki, R. L. Poynter, M. Aubier, A. Boischot, Y. Leblanc, A. Lecacheux, B. M. Pedersen, and P. Zarka. Voyager 2 radio observations of Uranus. *Science*, 233(4759):102–106, 1986.

[152] S. T. Weir, A. C. Mitchell, and W. J. Nellis. Metallization of fluid molecular hydrogen at 140 GPa (1.4 mbar). *Phys. Rev. Lett.*, 76(11):1860–1863, 1996.

[153] W. Yang. Direct calculation of electron density in density-functional theory. *Phys. Rev. Lett.*, 66(11):1438–1441, 1991.

[154] C. Zhang, R. Z. Khaliullin, D. Bovi, L. Guidoni, and T. D. Kühne. Vibrational signature of water molecules in asymmetric hydrogen bonding environments. *J. Phys. Chem. Lett.*, 4(19):3245–3250, 2013.

Declaration

I hereby declare that I wrote the dissertation submitted without any unauthorized external assistance and used only sources acknowledged in the work. All textual passages which are appropriated verbatim or paraphrased from published and unpublished texts as well as all information obtained from oral sources are duly indicated and listed in accordance with bibliographical rules. In carrying out this research, I complied with the rules of standard scientific practice as formulated in the statutes of Johannes Gutenberg-University Mainz to insure standard scientific practice.

Mainz, June 26, 2014

Dorothee Richters