Doctoral Thesis

Dissipative quantum transport simulations in two-dimensional semiconductor devices from first principles

Author(s): Szabó, Áron

Publication Date: 2016

Permanent Link: https://doi.org/10.3929/ethz-a-010659234

Rights / License: In Copyright - Non-Commercial Use Permitted
Dissipative quantum transport simulations in two-dimensional semiconductor devices from first principles

A thesis submitted to attain the degree of DOCTOR of SCIENCES of ETH ZURICH (Dr. sc. ETH Zurich)

presented by
ÁRON SZABÓ
MSc. in Engineering Physics, Budapest University of Technology and Economics
born on April 16th, 1987
citizen of Hungary

accepted on the recommendation of
Prof. Dr. Mathieu Luisier, examiner
Prof. Dr. Nicola Marzari, co-examiner

2016
Acknowledgments

First of all, I would like to express my sincere gratitude to my advisor Prof. Dr. Mathieu Luisier for the opportunity to join his group at the Integrated Systems Laboratory (IIS) and for the continuous support of my Ph.D study and related research. I am thankful for his trust, patience, guidance, and the freedom he gave me in my work. I am especially grateful to him for the thorough revision of this thesis.

I would also like to thank Prof. Dr. Nicola Marzari for accepting to be the co-examiner of this thesis, and Prof. Dr. Steven J. Koester and Prof. Dr. Lukas Novotny for their collaborations in our researches.

My gratitude also goes to Christine Haller, Hansjörg Gisler, and Christoph Wicki for maintaining excellent working conditions at the IIS, allowing me to focus on research only.

I am thankful to all the members of the Nano-TCAD group for the friendly environment they created. Special thanks go to Mauro Calderara and Sascha Brück with whom I shared our office for the longest term. They were always eager to help me, engage in enduring discussions, and they also knew how to stay quiet and let me concentrate on my work. I’m also grateful to Anne Ziegler for all the effort she put into trying to get us out of the office sometimes.

I would like to thank my excellent teachers at my alma mater and especially my MSc. advisor Prof. Dr. Ádám Gali for providing me with all the background I needed to start my work here.

Finally, I would like to express my deepest gratitude to my parents, Sándor and Krisztina, and to my siblings, Judit, Dávid, and Laura, for all the love and support they gave me in every circumstances through all my life and for the examples they have set.
Abstract

In this thesis a simulation framework for efficient and accurate atomic-level treatments of electron transport in the presence of electron-phonon interactions in nanoscale devices is developed. It is based on the non-equilibrium Green’s function (NEGF) formalism with every building block of the considered systems determined from ab initio density-functional theory (DFT) calculations. The strength of this approach is demonstrated through the investigation of transistors made of single- and few-layer semiconductors as well as van der Waals heterostructures.

DFT simulations represent a robust technique in the modeling of nanostructures, but their practical use is restricted to scales below realistic device sizes due to the heavy computational burden associated with them. This limitation is circumvented here by transforming the delocalized crystal electron states into maximally localized Wannier functions (MLWFs) that serve as the basis for the employed NEGF quantum-transport solver. The electron-phonon coupling strengths are also extracted from DFT calculations where the atoms are displaced from their original positions by small numerical values.

The analysis of MoS\textsubscript{2} field-effect transistors (FETs) reveals that the inclusion of electron-phonon scattering in the computational model is essential for transition metal dichalcogenide (TMD) devices, the ballistic approach leading to unphysical negative differential resistance phenomena. It is also found that due to the strong screening effects a double-gate design is required to benefit from the improved carrier mobility in triple-layer MoS\textsubscript{2} compared to a single- or double-layer device. The performance of other TMDs as well as black phosphorus FETs are studied and compared to each other in order to support
the on-going experimental efforts in the pursuit for ultimate high-performance logic devices.

The proposed framework has been extended to simulate heterostructures in addition to homogeneous FET channels. A MoTe$_2$-SnS$_2$ van der Waals heterojunction tunneling transistor is investigated as possible efficient subthermionic low-power switch. The effect of metal contacts on two-dimensional semiconductors is also examined. The thesis provides detailed explanations with step-by-step tutorials on the application of the MLWF technique in transport problems. Such approaches have recently started to gain increasing attention from the device modeling community.
Zusammenfassung


Die Analyse von MoS\textsubscript{2} Feldeffekt Transistoren (Field-Effect Transistor - FET) offenbart, dass der Einbezug von Elektron-Phonon Streuung im Rechenmodell für Übergangsmetall-Dichalkogenide (Transition Metal Dichalcogenide - TMD) essentiell ist, da der ballistische Ansatz zu unphysikalischen negativen Differentialwiderständen führt. Durch starke Abschirmeffekte ist auch ein Doppel-Gate Design nötig, um von der verbesserten Ladungsträger-Mobilität in Triple-Schicht
MoS$_2$ zu profitieren, die es im Gegensatz zu Einzel- oder Doppelschichten besitzt. Die Leistung anderer TMDs, sowie von schwarzen Phosphor-FETs, wird untersucht und miteinander verglichen, um die laufenden Experimente bei der Herstellung von High-Performance Logik-Schaltungen zu unterstützen.

# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledgments</td>
<td>v</td>
</tr>
<tr>
<td>Abstract</td>
<td>vii</td>
</tr>
<tr>
<td>Zusammenfassung</td>
<td>ix</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>2 Theory of nanoscale device simulations</td>
<td>7</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>7</td>
</tr>
<tr>
<td>2.2 Numerical quantum transport simulation basics</td>
<td>11</td>
</tr>
<tr>
<td>2.2.1 The Schrödinger equation as a matrix equation</td>
<td>11</td>
</tr>
<tr>
<td>2.2.2 Open vs. closed boundary conditions</td>
<td>13</td>
</tr>
<tr>
<td>2.2.3 Observables in an open system</td>
<td>16</td>
</tr>
<tr>
<td>2.3 Transport in the NEGF formalism</td>
<td>19</td>
</tr>
<tr>
<td>2.3.1 Green’s functions in the ballistic limit</td>
<td>20</td>
</tr>
<tr>
<td>2.3.2 The NEGF equations in two dimensions</td>
<td>22</td>
</tr>
<tr>
<td>2.3.3 Electron-phonon coupling</td>
<td>29</td>
</tr>
<tr>
<td>2.3.4 Mobility calculations</td>
<td>31</td>
</tr>
<tr>
<td>2.4 Density-functional theory</td>
<td>32</td>
</tr>
<tr>
<td>2.4.1 The Kohn-Sham equations</td>
<td>32</td>
</tr>
<tr>
<td>2.4.2 The exchange-correlation functional</td>
<td>35</td>
</tr>
<tr>
<td>2.5 Maximally localized Wannier functions</td>
<td>38</td>
</tr>
<tr>
<td>2.5.1 The Wannier transformation</td>
<td>39</td>
</tr>
<tr>
<td>2.5.2 Maximal localization</td>
<td>41</td>
</tr>
<tr>
<td>2.5.3 Entangled bands</td>
<td>45</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>------------------------------------------------------------</td>
</tr>
<tr>
<td>2.5.4</td>
<td>Derivatives of the MLWF Hamiltonian</td>
</tr>
<tr>
<td>2.6</td>
<td>Summary of the workflow</td>
</tr>
<tr>
<td>3</td>
<td>Simulation of MoS\textsubscript{2} FETs</td>
</tr>
<tr>
<td>3.1</td>
<td>Introduction</td>
</tr>
<tr>
<td>3.2</td>
<td>Computational details</td>
</tr>
<tr>
<td>3.3</td>
<td>Phonon-limited electron mobilities</td>
</tr>
<tr>
<td>3.4</td>
<td>Ballistic device simulations</td>
</tr>
<tr>
<td>3.5</td>
<td>Dissipative device simulations</td>
</tr>
<tr>
<td>3.6</td>
<td>Conclusions</td>
</tr>
<tr>
<td>4</td>
<td>Black phosphorus: an emerging transistor alternative</td>
</tr>
<tr>
<td>4.1</td>
<td>Introduction</td>
</tr>
<tr>
<td>4.2</td>
<td>Simulation setup</td>
</tr>
<tr>
<td>4.3</td>
<td>Investigation of electron- and hole mobilities</td>
</tr>
<tr>
<td>4.4</td>
<td>Analysis of phosphorene FETs</td>
</tr>
<tr>
<td>4.5</td>
<td>Conclusions</td>
</tr>
<tr>
<td>5</td>
<td>Van der Waals heterojunction tunneling FETs</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction</td>
</tr>
<tr>
<td>5.2</td>
<td>Material selection</td>
</tr>
<tr>
<td>5.3</td>
<td>Simulation approach</td>
</tr>
<tr>
<td>5.4</td>
<td>Device simulations</td>
</tr>
<tr>
<td>5.5</td>
<td>Conclusions</td>
</tr>
<tr>
<td>6</td>
<td>Metal contacts in 2D semiconductors</td>
</tr>
<tr>
<td>6.1</td>
<td>Introduction</td>
</tr>
<tr>
<td>6.2</td>
<td>Simulation setup</td>
</tr>
<tr>
<td>6.3</td>
<td>Investigation of the transfer length</td>
</tr>
<tr>
<td>6.4</td>
<td>Contact resistance calculation</td>
</tr>
<tr>
<td>6.5</td>
<td>Conclusions</td>
</tr>
<tr>
<td>7</td>
<td>Conclusion and outlook</td>
</tr>
<tr>
<td>A</td>
<td>Assembling the device Hamiltonian</td>
</tr>
<tr>
<td>B</td>
<td>Modeling of interfaces</td>
</tr>
<tr>
<td>B.1</td>
<td>Van der Waals heterojunctions</td>
</tr>
<tr>
<td>B.2</td>
<td>Metal-semiconductor interfaces</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

The technological advancement in the last 50 years has changed our everyday life to an extent comparable to the industrial revolution. In terms of impact, the 20th century counterpart of the steam engine is the transistor. Although it took some time from the first demonstration of a point-contact transistor by Bardeen, Shockley, and Brattain at Bell Labs in 1947 to the mass production of modern computer processors, it can be said without any doubts that transistors have shaped today’s world. To reach this point the semiconductor industry has aggressively pushed back its limits for decades, so that the ever growing computational power of the ever shrinking microchips could radically transform our civilization, from economy, science, or healthcare, to social life.

The main force behind the exponential growth of computer performance is the miniaturization of dimensions. What was first observed in 1965 by Gordon Moore [1], the fact that the number of components per chip area doubles every two years, became a law that every manufacturers tried to follow in a highly competitive race. Although significant improvements have been made in the optimization of the chip designs, it is ultimately the scaling of the active components – the transistors – from the micrometer to the nanometer scale that gave rise to modern digital technology. The ability to pack hundreds of millions of logic switches with enhanced capabilities onto a decreasing area has led to modern computers with far-reaching functionalities and
paved the way for integrating electronics into cars, phones, household
devices, and every gadgets, turning it into an organic part of our life.

The semiconductor industry itself has also greatly benefited from
its own product. Computers play an important role not only in the
automated fabrication processes, but also in the development of novel
microchip technologies. The use of technology computer aided design
(TCAD) in supporting long and expensive experimental processes has
became standard in the semiconductor companies. Designing a new
transistor strongly relies on this cost- and time-efficient method. To
continue this success story in the future the abilities and accuracy of
next-generation computer simulation tools will have to improve to the
point where systems composed of hundreds of thousands of atoms can
be handled on the atomistic, quantum-mechanical level.

The development of simulation tools with increasing numerical
efficiency and continuously improving physical models is becoming a
goal of utmost importance now that the conventional silicon metal-
oxide-semiconductor field-effect transistor (MOSFET) technology is
approaching its limits. An enormous design space has to be explored
in order to find the best active components that will hold the core of
future’s electronics.

For this purpose we have assembled a simulation framework based
on first principles (ab initio) density-functional theory (DFT) [2] and
the non-equilibrium Green’s function (NEGF) technique [3] [4]
connected through the use of maximally localized Wannier functions
(MLWFs) [5], an increasingly popular combination in the field of
computational device simulations. This thesis aims at providing a
tutorial about this method by reviewing its theoretical background
and highlighting its key features in Chapter 2. Furthermore, each
step will be explained in details via examples in Appendices A and
B. The codes implemented along with the presented examples are
publicly available on the website of the Nano-TCAD group of ETH
Zürich at https://www.nano-tcad.ethz.ch.

To keep up with Moore’s law and overcome the parasitic effects
of ultra-scaled devices, Intel has already replaced planar MOSFETs
with three-dimensional FinFETs at the 22 nm technology node [6]. Al-
though FinFETs possess the potential to satisfy the industry require-
ments for the next 2-3 technology nodes, in the long run novel device
architectures or materials will be required to minimize current leakages, self-heating, power consumption, and maintain a good electrostatic control of the active region of the transistors. The most intensively studied novel designs include silicon-on-insulator (SOI) ultra-thin-body (UTB)-[7], and gate-all-around (GAA) nanowire (NW) field-effect-transistors (FETs) [8][9][10]. As for the channel material, III-V compounds such as InAs or InGaAs are of great interest [11][12].

Another material family made of atomically thin structures has just gained attention as potential replacements for silicon in few years. The discovery of graphene by Novoselov and his co-workers in 2004[13] generated an unprecedented hype in material sciences. However, due to the lack of an electronic band gap, graphene is not well suited for digital logic applications. Six years later, the first experimental realization of a single-layer MoS$_2$ FET [14] redirected the spotlight to these van der Waals crystals, especially to few-layer metal-dichalcogenides (MDs). They are given by the chemical formula $MX_2$, where $M$ is a metal (Mo, W, Zr, Hf, Sn . . . ) and $X$ a chalcogen (S, Se, Te).

Single- or few-layer MDs and their heterostructures [15][16] have demonstrated their strength in various applications such as light emitting diodes (LEDs) [17], lasers [18], photodetectors [19][20][21], photovoltaics [22], biosensors [23][24], thermal- [25] and mechanical [26] energy harvesters, supercapacitors [27], memory cells [28], or as catalyst for water splitting [29]. It is, however, not clear – despite extensive studies – whether they will outperform silicon in the area of high-performance (HP) logic switches. This is why Chapter 3 of this thesis is dedicated to finding the answer to this question by means of computer simulations based on advanced atomic-level physical models.

Other than MDs, black phosphorus – which has been first isolated in single-layer thickness in 2014 [30] – is another similar candidate as HP transistor [31]. Black phosphorus (BP) and MDs share a layered structure with strong covalent or ionic bonding within an atomic layer, but only weak van der Waals forces holding different sheets together. Single- or few-layer samples of these materials can hence be easily obtained by mechanical exfoliation. As an alternative, some of them can be controllably grown by chemical/physical vapor deposition [32][33]. The confinement of electrons in an atomically thin
plane and a missing inversion symmetry gives rise to fascinating new properties \cite{34} \cite{35} \cite{36} \cite{37}. The strong light-matter interactions in these two-dimensional (2D) crystals \cite{38} \cite{39} make them appealing for optoelectronic applications \cite{40}, while the high carrier mobilities \cite{30} \cite{31} \cite{41} \cite{42}, the sizeable band gaps of the semiconducting allotropes, and the excellent electrostatic control that the 2D geometry offers render them quite desirable as FET materials. Unlike most MDs, phosphorene also exhibits extraordinarily high anisotropy \cite{43}. Its implications on the performance of single-layer BP FETs will be discussed in Chapter 4.

MOSFETs suffer from a relatively high power dissipation because their supply voltage has not scaled as fast as their dimensions. To attenuate this limitation a completely different device concept has been emerged at the beginning of the years 2000, the tunneling FETs (TFETs). TFETs rely on the quantum-mechanical tunneling of electrons from one band into another \cite{44}. The modulation of the tunneling path of cold electrons instead of the hot ones as in thermionic currents alleviates the theoretical limit imposed to the current gain in MOSFETs, a maximum of one decade of drain current increase per 60 mV change in the gate voltage at room temperature \cite{45} \cite{46}. Overcoming this performance barrier makes it possible to drastically reduce current leakages \cite{47} \cite{48} \cite{49}. Still, delivering an ON-state current high enough for practical applications without increasing the leakage in the OFF-state remains a huge challenge in TFETs.

Van der Waals heterostructures (vdWHs) made of 2D semiconductors might help resolve this issue, because they decrease the tunneling length to its minimum, offering the possibility to reach higher currents than in III-V devices or monolayer homojunctions. Furthermore, the absence of dangling bonds at the interface of vertical van der Waals heterojunctions reduces defects-induced performance deterioration \cite{16} and enables the combinations of materials with different lattice constants. Hence vdWHs greatly widen the available design space in discovering the material combination that will lead to the best device performance. In Chapter 5 a promising van der Waals heterostructure is proposed and its TFET operation analyzed.

While 2D semiconductors hold great promises to face the challenges of present-day electronics, their integration into electric circuits is accompanied by several technical difficulties. They suffer from a
large contact resistance at the interface of bulk metal interconnects that severely limits their performance [50]. Due to the atomic scale of the semiconductor layer, most techniques developed for reducing the contact resistance at silicon-metal interfaces are not applicable here [51]. In order to gain a better understanding of the underlying physical mechanisms and to support the ongoing experimental efforts [42] [52] [53], various computer simulations have been recently performed. However, only the equilibrium properties of the interfaces have been studied [54] [55], mainly focusing on the Schottky barrier height, which is only one piece of the puzzle. Realistic transport simulations in metals connected to 2D systems have not been undertaken yet, but the framework established in this work is suitable to deal with this task, as illustrated in Chapter [6].
Chapter 2

Theory of nanoscale device simulations

2.1 Introduction

When approaching the realm of the atomic scale, classical physics approximations hopelessly fail. The behavior of microscopic systems is governed by the laws of quantum mechanics instead. For decades the classical drift-diffusion model of electron flow has been the most eagerly applied theory in the study of transistors and other solid-state electronic devices. As the channel lengths shrink towards 10 nm and ultimately below it is no longer a viable option. In order to understand the physics of such nano-scale devices and be able to make useful predictions about their behavior, we have to face the problem at its root and solve the Schrödinger equation of the investigated system, which is not a straightforward task in most cases. Neglecting the time-dependence and focusing on the stationary problem of $N$ particles, we are still confronted to a $3N$ dimensional eigenvalue equation where the many-particle wavefunction $\Psi(r_1, \ldots r_N)$ is the unknown quantity governed by the following equation
\[
\sum_i \left( -\frac{\hbar^2}{2m_i} \nabla_i^2 + V(\mathbf{r}_1, \ldots, \mathbf{r}_N) \right) \Psi(\mathbf{r}_1, \ldots, \mathbf{r}_N) = E \Psi(\mathbf{r}_1, \ldots, \mathbf{r}_N).
\]

(2.1)

Any attempt to solve this equation numerically by brute force even for a moderated amount of particles is doomed to failure due to the astronomically large data size that such a calculation involves. For example, if we sample the wavefunction of 10 electrons on a grid with 1000 points along each spatial directions, we are left with the task of diagonalizing a matrix with the linear size of $1000^3 \cdot 10^9 = 10^{90}$. For comparison, the estimated number of atoms in the observable universe is approximately $10^{80}$.

The only way to render this problem feasible is by drastically reducing the dimensions of the wavefunction. The first step in this direction is the Born-Oppenheimer approximation \cite{56}, i.e. the separation of the wavefunctions of the nuclei and the electrons. In most practical cases the ions formed of the nuclei along with the core electrons of the closed shells are treated classically, characterized by their potential field, and only the wavefunction of the valence electrons is solved at a quantum-mechanical level. Still, without the introduction of single-particle wavefunctions $\phi_i(\mathbf{r}_i)$ that only depend on a single 3D spatial variable, the problem is impossibly large to handle. Most approaches in computational quantum chemistry and solid-state physics are therefore based on single-electron eigenvalue equations that correspond to the picture of each electron interacting only with the mean-field of all the others. In the Hartree-Fock method \cite{57} the many-electron wavefunction can be reconstructed from the single-electron functions as an antisymmetric product known as a Slater determinant. In the widely used density-functional theory \cite{2} the total wavefunction of the whole system is not even defined, instead the single-electron orbitals are deemed to bear physical meaning.

If one intends to study a solid-state device in this mean-field approximation, the core question is what happens to an electron that is injected into the simulation domain, giving rise to a flow of current? In the wavefunction formalism the single-particle stationary Schrödinger equation is modified to account for electrons entering/leaving the system through so-called open boundary conditions (OBCs) \cite{58}. Solving
the resulting equations reveals the probability for one electron injected at one side of the device to reach the other side. From this probability the electric current can be computed. However, when incoherent scattering mechanisms such as electron-phonon interactions are present in the considered systems, they can not be characterized with single-particle stationary wavefunctions any more. Instead, all the required information can be described by its Green’s functions within the non-equilibrium Green’s function (NEGF) formalism \cite{3 4}. In the NEGF approach one usually starts from a non-interacting (single-particle) Hamiltonian. Many-body effects are introduced in a perturbative way through scattering self-energy terms \cite{59 60}.

The most critical point is the choice of the non-interacting Hamiltonian entering the NEGF equations. This will depend on the desired accuracy and on the available computational resources. The simplest approach is the effective mass approximation (EMA) \cite{61}, where the wave equation of a crystal electron is replaced by the expression for a free particle with a different mass. A better description of the electronic properties of a given system is obtained by using a tight-binding (TB) Hamiltonian. In the conventional Slater-Koster TB method \cite{62} the electron wavefunctions are expanded in terms of atomic orbitals centered on the ions. The interactions of those orbitals with their nearest-neighbors are characterized by a set of material-dependent parameters.

The main advantage of this method is that it employs only short-range, e.g. nearest-neighbor interactions so that the Hamiltonian matrix describing typical devices is block-tridiagonal, allowing for the simulation of large atomic structures. Tight-binding accounts for more than one bands, for non-parabolicity effects, and supports band-to-band tunneling, contrary to the EMA. Its drawback resides in the mandatory parametrization of each new material that one wants to study. This task is accomplished by adjusting the material-dependent parameters mentioned above till selected band structure properties are accurately reproduced by the TB model. The targets may come from experimental data or from higher order band structure calculations using first principles \textit{(ab initio)} methods as provided by density-functional theory (DFT).

When exploring novel materials with no or very few experimental data, DFT is the only way to generate meaningful targets. Note that
even when enough data is available, TB might not offer a description of the band structure as reliable as the DFT, especially in systems with extended electron states. Increasing the number of atomic basis functions or the number of interactions can improve the results, but at the expense of the Hamiltonian size. Besides that, the fitting of 30 to 40 additional parameters is also not a trivial task.

The strength of DFT calculations is that they do not require any empirical parameters. In principle it would be possible to create the Hamiltonian of an entire device using DFT, if \textit{ab initio} calculations were not computationally so demanding as compared to TB ones. The most accurate DFT tools use plane-wave basis that yields full Hamiltonian matrices, which sets a practical limit for the system sizes at around a thousand of atoms. To simulate an homogeneous crystal, however, one only needs to compute the wavefunction of a representative single unit cell and apply periodic boundary conditions. Such calculations are feasible for most materials of interest. The electronic band structure is therefore relatively straightforward to obtain.

Instead of fitting a pre-defined TB model to a band structure of interest, a sparse Hamiltonian can be directly produced from DFT wavefunctions. With an approach based on maximally localized Wannier functions (MLWF) \cite{5} plane waves can be converted into an orthonormal localized basis set using a unitary transformation matrix. Due to the strong localization of the Wannier functions the strength of the interactions between two orbitals decays quickly with their distance. Truncating the matrix elements beyond a certain interaction range is therefore well justified and leads to a sparse block-tridiagonal TB-like matrix. Hence Wannier functions can serve as the basis for quantum transport simulations. The Hamiltonian matrix of realistic devices composed of up to hundred thousands of atoms can be assembled from the matrix elements of small supercells and efficiently solved on modern supercomputers.

It will be shown in this work that completely neglecting the ionic movements may result in qualitatively wrong predictions. This issue can be elegantly solved within the framework of the NEGF, whose main strength is its ability to incorporate various scattering mechanisms such as electron-phonon interactions without modifying the core of the simulation tool. Since the necessary electron-phonon coupling
parameters and phonon modes can also be extracted from *ab initio* density-functional perturbation theory (DFPT) \[63\] \[64\] calculations, we ended up with a simulation environment dedicated to nano-devices that does not rely on a single empirical material parameter and possesses therefore a great predictive power.

In Section 2.2 the basics of numerical quantum transport simulations will be reviewed through a simple example using single-particle wavefunctions. In Section 2.3 the NEGF formalism will be introduced, focusing on two-dimensional electron transport with electron-phonon scattering. Density-functional theory will be discussed in Section 2.4 and the MLWF transformation in Section 2.5. The complete simulation process is summarized in Section 2.6. Furthermore, a detailed step-by-step example on the preparation of the Hamiltonian matrix of a MoS$_2$ field-effect transistor will be presented in Appendix A.

### 2.2 Numerical quantum transport simulation basics

In this Section a simple technique for numerically solving the stationary Schrödinger equation in an open system is presented. As an example a 1D effective mass Hamiltonian is used. The basic concepts encountered in transport problems will be introduced and it will be shown how physical quantities such as the charge density, current, and electrostatic potential can be determined in the investigated system.

#### 2.2.1 The Schrödinger equation as a matrix equation

Let us first review how the time-independent Schrödinger equation transforms into a matrix equation when it is discretized on a real-space grid. For simplicity the single-band effective mass approximation (EMA) will be applied to a one-dimensional simulation domain. In the EMA the influence of the periodic crystal potential is accounted for through a modified effective electron mass $m_e^*$. Allowing $m_e^*(x)$ to spatially vary as in heterostructures, the problem to be solved becomes \[65\]
CHAPTER 2. THEORY OF NANOSCALE DEVICES

\[
\left[ -\frac{\hbar^2}{2} \frac{d}{dx} \frac{1}{m_e^*(x)} \frac{d}{dx} + V(x) \right] \Psi(x) = E \Psi(x). \tag{2.2}
\]

Here, \( V(x) \) is the external potential that excludes the crystal potential. The wavefunction \( \Psi(x) \) is discretized on a homogeneous grid of \( N \) points \( \{x_i\} \) with \( i = 1 \ldots N \) and a spacing \( \Delta x \) between them. It is represented as a vector of the function values at each grid point

\[
\Psi = \{\Psi(x_i)\} = \{\Psi_i\}. \tag{2.3}
\]

The numerical derivative \( \frac{d}{dx} \Psi(x) \) at the \( x_i \) point can be expressed then e.g. as

\[
\left. \frac{d}{dx} \Psi(x) \right|_{x_i} = \frac{\Psi(x_{i+1}) - \Psi(x_{i-1})}{2\Delta x}. \tag{2.4}
\]

It clearly appears in Eq. (2.4) that if we write the derivation operator as a matrix, all its \((i, i \pm 1)\) entries have a value \( \pm 1/2\Delta x \) equal to the coefficients multiplying the \( i \pm 1 \) components of the vector \( \Psi, \Psi_{i-1} \) and \( \Psi_{i+1} \) respectively. Using the same approach to express second derivatives on a discrete grid, Eq. (2.2) can be written after straightforward algebraic steps as

\[
(E - H_{ii}) \Psi_i - H_{ii+1} \Psi_{i+1} - H_{ii-1} \Psi_{i-1} = 0, \tag{2.5}
\]

with

\[
H_{ii} = \frac{\hbar^2}{\Delta x^2} \left( \frac{1}{m_i + m_{i+1}} + \frac{1}{m_i + m_{i-1}} \right) + V(x_i),
\]

\[
H_{ii+1} = -\frac{\hbar^2}{\Delta x^2} \frac{1}{m_i + m_{i+1}},
\]

\[
H_{ii-1} = -\frac{\hbar^2}{\Delta x^2} \frac{1}{m_i + m_{i-1}}. \tag{2.6}
\]

In Eq. (2.6) \( m_i = m_e^*(x_i) \) denotes the effective mass at position \( x_i \). It can be cast into a matrix-vector form

\[
(E - H) \Psi = 0, \tag{2.7}
\]

with the diagonal matrix \( E \).
2.2. QUANTUM TRANSPORT SIMULATION BASICS

\[ E = \begin{bmatrix}
E & 0 & 0 & \cdots & 0 & 0 \\
0 & E & 0 & & & \\
0 & 0 & E & \ddots & & \\
\vdots & \ddots & \ddots & \ddots & 0 & 0 \\
0 & 0 & 0 & E & 0 & \\
0 & 0 & 0 & 0 & E
\end{bmatrix}, \]  
(2.8)

the tridiagonal matrix \( H \)

\[ H = \begin{bmatrix}
H_{11} & H_{12} & 0 & \cdots & 0 & 0 \\
H_{21} & H_{22} & H_{23} & & & \\
0 & H_{32} & H_{33} & \ddots & & \\
\vdots & \ddots & \ddots & \ddots & \ddots & 0 \\
0 & & H_{N-1,N-1} & H_{N-1,N} & & \\
0 & 0 & 0 & H_{N,N-1} & H_{N,N}
\end{bmatrix}, \]  
(2.9)

and the vector \( \Psi \)

\[ \Psi = \begin{pmatrix}
\Psi_1 \\
\Psi_2 \\
\Psi_3 \\
\vdots \\
\Psi_{N-2} \\
\Psi_{N-1} \\
\Psi_N
\end{pmatrix}. \]  
(2.10)

The size of \( E, H, \) and \( \Psi \) corresponds to \( N \), the number of discretization points in the grid.

### 2.2.2 Open vs. closed boundary conditions

Special attention must be paid to the first and last rows of the Hamiltonian matrix, since the simulation domain only extends between the
In Eqs. (2.5-2.6) each row of the Hamiltonian matrix contains three nonzero matrix elements that multiply three elements of the \( \Psi \) vector, except the first and last rows of \( H \) which make up of two terms only. This is exactly the same result as if \( \Psi_0 = 0 \) and \( \Psi_{N+1} \) were explicitly enforced in Eqs. (2.5-2.6), i.e. \( H_{10} = 0 \) or \( \Psi_0 \) lead to the same Hamiltonian matrix. Hence, the absence of the \( H_{10} \Psi_0 \) and \( H_{N,N+1} \Psi_{N+1} \) terms forces the wavefunction to vanish at the boundaries. In this case we talk about closed boundary conditions.

Equation (2.14) is therefore an eigenvalue problem of a closed structure characterized by the Hamiltonian \( H \). The eigenvalues \( \{ E_n \} \) and eigenvectors \( \{ \Psi_n \} \) are the discrete eigenenergies and corresponding eigenstates of the system that are to be determined.

It is important to realize that a closed system is not suited for transport calculations, where electrons may enter or leave a finite-size domain. The wavefunction of an electron in the resulting open system should not vanish at the boundaries, but rather be smoothly connected to the wavefunction outside the simulated region. In the standard approach of quantum transport the device under consideration is assumed to be connected to homogeneous semi-infinite leads or contacts, where the effective mass and electrostatic potential energy remain constant and equal to the value at \( x_1 \) and \( x_N \), i.e. \( H_{ii} = H_{11}, \ H_{i,i+1} = H_{12} \) if \( i \leq 0 \) and \( H_{ii} = H_{NN}, \ H_{i,i-1} = H_{N,N-1} \) if \( i \geq N + 1 \). The wavefunction in the leads can then be approximated as the sum of Bloch functions. In the 1D EMA picture it reduces to plane waves

\[
\begin{align*}
\Psi_c^L(x) &= a^L e^{ik^L(E)x} + b^L e^{-ik^L(E)x} & x < x_1 \\
\Psi_c^R(x) &= a^R e^{-ik^R(E)(x-x_N)} + b^R e^{ik^R(E)(x-x_N)} & x > x_N.
\end{align*}
\tag{2.11}
\]

In the equations above \( \Psi_c \) is the wavefunction of the contact, the superscripts \( L \) and \( R \) refer to the left- and right-side regions of the device, respectively, \( k^L/R(E) \) represents the electron wavevector in the leads and depends on the energy \( E \), finally the terms multiplied by \( a^L/R \) (\( b^L/R \)) indicate incident (exiting) waves. It is worthwhile noting that the amplitude of the incident term, \( a^L/R \) can be normalized to 1 without inducing any restriction on the problem itself. Plugging Eq. (2.11) into Eqs. (2.5-2.6) for \( i = \{-1, 0, 1\} \) results in a linear system of equations that can be summarized in the following form for the left contact.
2.2. QUANTUM TRANSPORT SIMULATION BASICS

\[ (E - H_{11} - \Sigma_{11}(E))\Psi_1 - H_{12}\Psi_2 = S^L_1(E), \]  
(2.12)

with

\[
\begin{align*}
    k^L(E) &= \frac{1}{\Delta x} \cos \left( \frac{E-H_{11}}{2H_{12}} \right), \\
    \Sigma_{11}(E) &= H_{12}e^{ik^L(E)\Delta x}, \\
    S^L_1(E) &= H_{12}(1 - e^{2ik^L(E)\Delta x}).
\end{align*}
\]  
(2.13)

If we assume that electrons are only injected into the device from the left contact, \( a^R \) can be set to 0 and after applying the same procedure as before to the \( i = \{N - 1, N, N + 1\} \) indices we find that

\[ (E - H_{NN} - \Sigma_{NN}(E))\Psi_N - H_{N,N-1}\Psi_{N-1} = 0, \]  
(2.14)

with

\[
\begin{align*}
    k^R(E) &= \frac{1}{\Delta x} \cos \left( \frac{E-H_{NN}}{2H_{N,N-1}} \right), \\
    \Sigma_{NN}(E) &= H_{N,N-1}e^{ik^R(E)\Delta x}.
\end{align*}
\]  
(2.15)

Combining Eqs. (2.12) and (2.14) with Eq. (2.5) for the indices \( 2 \leq i \leq N - 1 \), a Schrödinger equation with open boundary conditions is obtained

\[ (E - H - \Sigma(E))\Psi^L = S^L(E), \]  
(2.16)

with the \( \Sigma(E) \) matrix having only two nonzero elements, the \( \Sigma_{11}(E) \) and \( \Sigma_{NN}(E) \) corner entries, while the \( S^L \) vector has only one, \( S^L_1(E) \), the first index. The matrix \( \Sigma(E) \) is called boundary self-energy, \( S^L \) the left injection vector. Both of them have a size of \( N \). The superscript in \( \Psi^L \) denotes that this wavefunction is the solution of the Schrödinger equation when injection only occurs at the left contact. Equation (2.16) is no longer an eigenvalue problem due to the nonzero right-hand-side and the fact that \( \Sigma \) and \( S^L \) depend on the energy \( E \). It becomes a sparse linear system of equations that must be solved for any injection energy \( E \), practically only those having a physical meaning, i.e. those corresponding to propagating and not decaying modes in the contacts.
CHAPTER 2. THEORY OF NANOSCALE DEVICES

There is however a second problem to tackle where the injection takes place in the right instead of the left contact. By setting $a^R = 1$ and $a^L = 0$, the following system of equations arise

$$\begin{align*}
(E - H - \Sigma(E))\Psi^R &= S^R(E), \\
\text{with the right injection vector } S^R \text{ having only one nonzero entry at the last index}
\end{align*}$$

$$S^R_N(E) = H_{N,N-1}(1 - e^{2i k^R(E) \Delta x}).$$

By defining $\Psi = \{\Psi^L, \Psi^R\}$ and $S(E) = \{S^L(E), S^R(E)\}$, both matrices of size $N \times 2$, Eqs. (2.16) and (2.17) can be cast into a single matrix equation

$$\begin{align*}
(E - H - \Sigma(E))\Psi &= S(E).
\end{align*}$$

It is important to emphasize that the boundary self-energy is the same regardless of the injection side. This simplifies the solution of Eq. (2.19) since the matrix $E - H - \Sigma(E)$ must only be factorized once per energy $E$. This operation is followed by two back-substitutions, one for $\Psi^L$ and one for $\Psi^R$.

### 2.2.3 Observables in an open system

Since $\Psi^L$ and $\Psi^R$ are uncorrelated waves, the electron density at a given point $x_i$ is the sum of two distinct carrier populations, one that depends on the $\Psi^L_i \Psi^L_i^*$ probability of an electron injected from left and one that is a function of $\Psi^R_i \Psi^R_i^*$. Besides the spatial distribution of the left- and right-injected states, their energy-dependent occupancy should be taken into account as well.

In order to get the electron density induced by the injection from the left or right side, the probability of finding an electron at $x_i$ has to be summed up over all possible $k$-states in the contact that model the lateral dimensions of the device along $y$ and $z$, and multiplied with the probability of those states to be occupied. As a consequence, the total 3-dimensional wavefunction can be expanded according to Bloch’s theorem as
\[ \Psi_{k_t}(x, r_t) = \frac{1}{\sqrt{A}} e^{i k_t \cdot r_t} \Psi(x), \quad (2.20) \]

where \( A \) is the area in the \( y - z \) plane, while \( k_t \) and \( r_t \) refer to the transverse momentum and the position in the plane perpendicular to the transport direction. In the effective mass approximation the total energy of electrons is given by

\[ E_{\text{tot}} = E + E(k_t) = E + \frac{\hbar^2 |k_t|^2}{2m^*}. \quad (2.21) \]

Since the electrons are in thermal equilibrium in the leads, the occupancy of the states obey the Fermi-Dirac distribution function \( f(E, E_f) \), where \( E_f \) is the Fermi energy. The charge density is therefore defined as

\[ n(x_i) = \int dE \sum_C g^C(E, x_i) \frac{1}{A} \sum_{k_t} f(E + E(k_t), E_f^C). \quad (2.22) \]

The superscript \( C \) denotes “contact” \((L/R)\), and \( g^C(E, x_i) \) is the density-of-states at energy \( E \) and position \( x_i \) resulting from the waves injected from contact \( C \)

\[ g^C(E, x_i) = \frac{1}{2\pi} \left| \Psi^C_{E,i} \right|^2 \left| \frac{dE(k_C)}{dk_C} \right|^{-1}_{E(k_c) = E}. \quad (2.23) \]

In Eq. \((2.23)\) \( \Psi^C_{E,i} \) is the \( i \)-th element of the solution of Eq. \((2.19)\) at the incident electron energy \( E \) and \( \frac{dE}{dk_C} \) is the derivative of the energy dispersion in the contact \( C \). By introducing the function \( F(E, E_f) \) as the normalized sum of the Fermi function \( f(E + E(k_t), E_f) \) over the transverse momenta \( k_t \)

\[ F(E, E_f^C) = \frac{1}{A} \sum_{k_t} f(E + E(k_t), E_f^C) \]

\[ = \frac{m^* k_B T}{\pi \hbar^2} \ln \left( 1 + \exp\left\{ (E_f - E)/k_B T \right\} \right), \quad (2.24) \]
where \( k_B \) is the Boltzmann constant and \( T \) the temperature, the electron density can be written in a more compact form

\[
n(x_i) = \sum_C \int_{C_{BM}}^{\infty} \frac{dE}{2\pi} \frac{\Psi^C_{E,i} \Psi^{C*}_{E,i}}{\left| \frac{dE}{dk^C} \right|^{-1}} F(E, E_f^C).
\] (2.25)

The integration in Eq. (2.25) goes from the conduction band minimum (CBM) of the appropriate lead. The determination of the physical quantities in an open system hence requires Eq. (2.19) to be evaluated at many electron energies such that the integration in Eq. (2.25) can be numerically approximated by a finite sum at those points.

It is crucial not to forget that the Schrödinger equation (2.2) contains an external potential energy term \( V(x) = V_H(x) + V_{mat}(x) \), whose Hartree component \( V_H(x) \) is yet to be determined. The material-dependent term \( V_{mat}(x) \) describes band offsets between different regions. The charge in the device gives rise to an electrostatic potential \( \Phi \) through Poisson’s equation

\[
\nabla^2 \Phi(r) = -\frac{\rho(r)}{\varepsilon(r)}.
\] (2.26)

The variable \( \varepsilon(r) \) is the dielectric permittivity at position \( r \), \( \rho(r) \) the total charge density including the contributions from both the electrons injected into the device \( n(r) \) and the position-dependent donor doping density \( N_D(r) \)

\[
\rho(r) = q(N_D(r) - n(r)).
\] (2.27)

In Eq. (2.27) \( q \) denotes the elementary charge. If holes and acceptors are also present and labeled \( p(r) \) and \( N_A(r) \), respectively, the expression modifies to

\[
\rho(r) = q(N_D(r) - N_A(r) + p(r) - n(r)).
\] (2.28)

The electrostatic potential energy \( V_H(r) \) of an electron is derived from \( \Phi(r) \) by multiplying it with \(-q\) so that

\[
V_H(r) = -q\Phi(r).
\] (2.29)
After $n(r)$ is determined, Eq. (2.26) can be numerically solved on a finite difference or finite element grid in a similar way as the Schrödinger equation. At this stage gate contact(s) can be introduced into the simulation domain of transistors through the appropriate Dirichlet boundary conditions. Since $n(r)$ depends on $V(r)$ through the Schrödinger equation and $V(r)$ is in turn a function of $n(r)$ through Poisson’s equation, these two quantities must be iteratively and self-consistently solved till convergence is reached.

Once this is achieved, the net current flowing through the device from left to right can be determined by the famous Landauer-Buttiker formula \[ I = -\frac{2e}{\hbar} \int_{C_B M}^{\infty} \frac{dE}{2\pi} T(E)(F(E, E_F^L) - F(E, E_F^R)), \] (2.30)

with $T(E)$ being the transmission probability per spin at energy $E$ defined as

$$T(E) = \frac{\Psi_{E,N}^L}{\Psi_{E,1}^L} \frac{v_R(E)}{v_L(E)}.$$ (2.31)

The transmission probability is built from the ratio of the last and first entries of the left-injected wavefunction times the ratio of the electron velocities in the right and left contacts. Replacing $L$ by $R$, $R$ by $L$, and swapping $1$ and $N$ gives exactly the same result.

### 2.3 Transport in the NEGF formalism

Although the use of steady-state single-particle wavefunctions represents an attractive way of approaching quantum transport problems, it suffers from severe limitations. When particles undergo inelastic scattering within the simulation domain, coherent wavefunctions are of no use anymore. Instead, one of the most widely spread techniques for the investigation of transport with scattering is the Non-equilibrium Green’s function (NEGF) formalism. Here, the Green’s function of a ballistic system expressed in the effective mass approximation is first introduced as a reformulation of the wavefunction equation. Then
the part of the NEGF theory that is relevant to this work will be
summarized, focusing on dissipative transport in 2D systems.

2.3.1 Green’s functions in the ballistic limit

Although Eq. (2.19) was derived in a 1D EMA scheme, the Schrödinger
equation with open boundary conditions keeps the same form within
any ballistic model where the Hamiltonian operator is expanded in
an orthogonal localized basis set and exhibits a block-tridiagonal pat-
tern. For brevity the energy-dependence of the self-energy and of the
injection vector will no more be marked from now on, which leaves

\[(E - H - \Sigma)\Psi = S\]  \hspace{1cm} (2.32)

as our starting point. In general \(\Sigma\) does not only include two
nonzero elements in the upper left and lower right corners, but two
diagonal blocks of the same size as the number of basis components
per discretization point. This quantity is labeled \(t_b\). Similarly, the
first and last \(t_b\) entries of the \(S\) vector may differ from zero. A
computational approach to determine these non-zero entries in the
tight-binding model is provided in Ref. [69].

The linear system of equations (2.32) can be reformulated in a
way that makes it straightforward to incorporate inelastic scatteri-
ing events into the picture. For this purpose we recall the definition of
the Green’s function \(G(x, x')\) associated with the operator \(L(x)\)

\[L(x)G(x, x') = \delta(x - x'),\] \hspace{1cm} (2.33)

where \(\delta(x - x')\) is the delta function. It follows from this equation
that once the Green’s function of the operator \(L(x)\) is known, the
solution \(f(x)\) of the inhomogeneous differential equation

\[L(x)f(x) = s(x)\] \hspace{1cm} (2.34)

can be simply calculated as

\[f(x) = \int G(x, x')s(x')dx'.\] \hspace{1cm} (2.35)

The Green’s function of the operator in Eq. (2.32) is called the
retarded Green’s function \(G^R\) and is defined by
where \( \mathbf{I} \) stands for the identity matrix, which is the discrete representation of the delta function. The self-energy is labeled now as \( \Sigma^R \) (retarded self-energy) for consistency in the notations, but it remains equal to the expression introduced before. The matrix \( \mathbf{G}^R \) can then be calculated as

\[
\mathbf{G}^R = (\mathbf{E} - \mathbf{H} - \Sigma^R)^{-1},
\]

and the wavefunction \( \Psi \) as

\[
\Psi = \mathbf{G}^R \mathbf{S}.
\]

By inserting this expression of \( \Psi \) into Eqs. (2.25) and (2.31), the charge and the current can be determined without the explicit knowledge of the wavefunction. More compact expressions can be derived with the introduction of three more quantities, (i) an advanced Green’s function \( \mathbf{G}^A \), defined as the conjugate transpose of the retarded Green’s function

\[
\mathbf{G}^A = \mathbf{G}^R \dagger,
\]

(ii) a lesser self-energy \( \Sigma^< \)

\[
\Sigma^< = i \sum_C \mathbf{S}^C \mathbf{S}^{C+} \left| \frac{dE}{dk^C} \right|^{-1} F(E, E_f^C) \Delta x,
\]

and (iii) a lesser Green’s function \( \mathbf{G}^< \)

\[
\mathbf{G}^< = \mathbf{G}^R \Sigma^< \mathbf{G}^A.
\]

With straightforward algebra Eq. (2.25) can be rewritten as

\[
n(x) = -\frac{i}{\Delta x} \int \frac{dE}{2\pi} \text{diag} \{ \mathbf{G}^<(E) \}.
\]

The electron density is therefore directly proportional to the diagonal elements of the lesser Green’s function.
The difference between the wavefunction and the Green’s function formalism is purely formal in the considered ballistic case. Equations (2.32) and (2.37) coupled with (2.41) involve the same physics and approximations. They just differ with respect to their computational costs. In large device structures it is advantageous to work in the wavefunction formalism that requires orders of magnitude less floating point operations to obtain exactly the same results. The strength of the Green’s function technique resides in its possible extension beyond the ballistic limit of transport, towards situations where inelastic scattering plays an important role and stationary single-particle wavefunctions fail. The NEGF formalism is discussed in details in the following subsection.

2.3.2 The NEGF equations in two dimensions

Over the last 20 years the non-equilibrium Green’s function method has established itself as one of the most popular techniques to determine the expectation values of observables in nanostructures driven out-of-equilibrium. The Green’s functions represent the response of a system to external perturbations. Similarly to wavefunctions, they contain all the required microscopic information, but many-body effects can be more easily incorporated into the governing equations through perturbation theory [70] and self-energy terms.

The single-particle non-equilibrium Green’s function introduced by Kadanoff and Baym [3] and Keldysh [4] is defined as the non-equilibrium ensemble average of the contour ordered field operators

$$G(11') = -\frac{i}{\hbar} \left\langle \hat{T}_C \left\{ \hat{\Psi}_H(1)\hat{\Psi}_H^\dagger(1') \right\} \right\rangle$$

$$= -\frac{i}{\hbar} \left( \Theta_C(1,1') \left\langle \hat{\Psi}_H(1)\hat{\Psi}_H^\dagger(1') \right\rangle \mp \Theta_C(1',1) \left\langle \hat{\Psi}_H^\dagger(1')\hat{\Psi}_H(1) \right\rangle \right).$$

(2.43)

In the equation above the argument (1) is the shorthand notation for \((r_1, t_1, \sigma_1)\), where \(\sigma\) is the spin index, \(r_1\) the position, and \(t_1\) the time. The upper sign (-) applies to fermions, the lower one (+) to bosons. The single-particle field operator \(\hat{\Psi}_H(1)\) is expressed in the Heisenberg picture, while \(\hat{T}_C\) is the contour-ordering operator.
2.3. TRANSPORT IN THE NEGF FORMALISM

Figure 2.1: Contour C along the time axis for an evaluation of the operator expectation value

\[ \hat{T}_C \left\{ \hat{A}(t_1) \hat{B}(t_2) \right\} = \Theta_C(t_1, t_2) \hat{A}(t_1) \hat{B}(t_2) \pm \Theta_C(t_2, t_1) \hat{B}(t_2) \hat{A}(t_1), \]

(2.44)

with \( \Theta_C(t_1, t_2) \) defined on the time contour C shown in Fig. 2.1

\[ \Theta_C(t_1, t_2) = \begin{cases} 1 & \text{if } t_1 \text{ is later than } t_2 \text{ on } C, \\ 0 & \text{otherwise.} \end{cases} \]  

(2.45)

The \( G(11') \) quantity has been proved very useful in describing the time-evolution of interacting systems by expanding the non-equilibrium ensemble averages in terms of the non-interacting Hamiltonian and by adding scattering in a perturbative way.

To compute \( G(11') \) in Eq. (2.43) its equations of motion must be derived with respect to the times \( t_1 \) and \( t_2 \) and solved in the time domain. This procedure is simplified by introducing four real-time Green’s functions, \( G^A(11') \) (advanced), \( G^R(11') \) (retarded), \( G^< (11') \) (lesser), and \( G^> (11') \) (greater), which are all equal to \( G(11') \), depending on the position of \( t_1 \) and \( t_1' \) on the complex contour of Fig. 2.1. Similarly, there exist four self-energies, \( \Sigma^A(11') \), \( \Sigma^R(11') \), \( \Sigma^< (11') \), \( \Sigma^> (11') \) that obey the same rules as their Green’s function counterpart. Here, we restrict ourselves to steady-state situations \[3 \quad 4 \quad 71\ \ 72\ \ 73\ \ 74\], where all the presented quantities are functions of \( t_1 - t_1' \). For convenience, this time difference can be Fourier transformed according to

\[ G^T(r_1 r_1'; E) = \int d(t_1 - t_1') e^{iE(t_1 - t_1')} G^T(11'), \]

(2.46)

with the energy \( E \) and \( T = \{ A, R, <, > \} \). From the initial 8 variables \( (x_1, y_1, z_1, t_1, x'_1, y'_1, z'_1, t_1') \) one could be eliminated, leaving
CHAPTER 2. THEORY OF NANOSCALE DEVICES

us with 7 to take care of. All Green’s functions and self-energies have been converted from real-time to energy-dependent variables.

To further reduce the computationally complexity, the Green’s functions as well as the self-energies can be expanded into a basis formed of \{\phi_n(r)\} components with

$$G^T(r_1r'_1; E) = \sum_{nm} G^T_{nm}(E) \phi^*_n(r_1) \phi_m(r'_1), \quad (2.47)$$

where the indices run over all the basis functions. To summarize, staring from the definition of the Green’s function in Eq. (2.43), by deriving its equations of motion, applying Langreth theorem to make the different Green’s function (A, R, <, >) appear [74], Fourier transforming the time differences, and expanding \(G^T\) and \(\Sigma^T\) in a basis set, we end up with the following system of coupled equations

$$EG^R_{nm}(E) - \sum_l (H_{nl} + \Sigma^R_{nl}) G^R_{lm}(E) = \delta_{nm} \quad (2.48a)$$

$$G^A_{nm}(E) = [G^R_{nm}(E)]^\dagger \quad (2.48b)$$

$$G^<_R_{nm}(E) = \sum_{l,v} G^R_{nl}(E) \Sigma^<_l(E) G^A_{vm}(E) \quad (2.48c)$$

$$\Sigma^R_{nm}(E) = \frac{1}{2} \left( \Sigma^>_nm(E) - \Sigma^<_nm(E) \right) - iP \int \frac{dE'}{2\pi} \frac{\Sigma^>_nm(E') - \Sigma^<_nm(E')}{E - E'} \quad (2.48d)$$

$$G^R_{nm}(E) - G^A_{nm}(E) = G^>_nm(E) - G^<_nm(E), \quad (2.48e)$$

with \(H_{nm}\) being the \((n, m)\)-th matrixelement of the unperturbed Hamiltonian \(\hat{H}\). The operator \(P\) denotes the principal value integral. The self-energies describe both the perturbations on the system and the coupling to its surrounding environment through open boundary self-energies. The charge density is again directly related to the lesser Green’s function by

$$n(r) = -i \sum_{n,m} \int \frac{dE}{2\pi} G^<_nm(E) \phi_n(r) \phi_m(r)^*.$$

(2.49)
Electron-electron interactions are modeled by the Hamiltonian

$$\hat{H}_{ee} = \frac{1}{2} \int dr \int dr' \hat{\Psi}^\dagger(r) \hat{V}_{ee}(r - r') \hat{\Psi}(r') \hat{\Psi}(r),$$

(2.50)

where $\hat{V}_{ee}(r - r')$ represents the Coulomb potential. It is treated perturbatively either through Wick’s decomposition [70], Feynman diagrams [59], or a variational approach [60], which leads to an infinite hierarchy of self-energies, the first one being the Hartree self-energy. It can be proved that the Hartree self-energy can be directly cast into the non-interacting Hamiltonian by introducing a potential energy $V(r) = -q\Phi(r)$, where $\Phi(r)$ is the solution of Poisson’s equation [75]. The Hamiltonian is, therefore, simply replaced in Eq. (2.48) by

$$H(r) \rightarrow H(r) + V(r).$$

(2.51)

Apart from the Hartree potential, the self-energies that we will include to characterize the perturbations of our quantum transport problems are the boundary and the electron-phonon scattering ones

$$\Sigma = \Sigma^B + \Sigma^S.$$  

(2.52)

In order to solve the coupled NEGF and Poisson’s equations, it is convenient to assume that the basis functions $\{\phi_n(r)\}$ are well localized functions around their $r_n$ centers, i.e $\phi_n(r) = f_n(r - r_n)$. The charge can then be represented as a point charge on the $r_n$ gridpoints that usually correspond to the atom positions. If more than one basis function is centered on a gridpoint, the $G^T_{nm}$ and $\Sigma^T_{nm}$ quantities in Eq. (2.48) become matrices whose indices indicate the atom they belong to and whose size is equal to $p \times q$, where $p$ is the number of basis functions centered on the $n$-th atomic gridpoint and $q$ is the number of basis functions centered on the $m$-th point. The charge on the $l$-th gridpoint is obtained by applying the trace operator to the $l$-th block of the lesser Green’s function

$$n(r_l) = -i \int \frac{dE}{2\pi} \text{tr} \, G^<_{ll}(E).$$

(2.53)

Since this work focuses on transport through two-dimensional devices, one can take advantage of the periodicity along the out-of-plane
Figure 2.2: (a) Side view of a double-layer MoS\textsubscript{2} structure where \( x \) denotes the transport direction and \( y \) the direction of confinement. (b) Top view of the same system as in (a). The \( z \) direction is assumed to be periodic and modeled through a set of \( k_x \) points. The device Hamiltonian \( H(k_z) \) is computed from the \( H_0 \) Hamiltonian of an MoS\textsubscript{2} strip with width \( \Delta \) and the \( H_- \) and \( H_+ \) couplings as

\[
H(k_z) = H_0 + H_- \cdot e^{-ik_z\Delta} + H_+ \cdot e^{ik_z\Delta}.
\]

The shaded areas mark the primitive hexagonal unit cell of MoS\textsubscript{2} used in the electronic structure calculations, while the rectangular one is utilized in the transport calculations.

direction. The transport axis will be denoted as \( x \), the direction of confinement \( y \), and the periodic direction \( z \). As an illustration of these principles the geometry of a double-layer MoS\textsubscript{2} is sketched in Fig. 2.2. The device dimensions are finite along \( x \) and \( y \), but periodicity along \( z \) is assumed. Due to the localization of the chosen basis functions, an \( r_{max} \) cutoff distance can always be defined such that atomic interactions beyond this range can be safely neglected. A rectangular unit cell with dimensions larger than \( r_{max} \) must be determined for transport calculations. It has a width \( \Delta \) in the \( z \) direction. Concretely, this means that the basis functions \( \phi_n(\mathbf{r}) \) should be made periodic along \( z \), which can be achieved through Bloch’s theorem.
\[ \phi_n(x, y, z) \rightarrow \phi_{n,k_z}(x, y, z) = f_n(x - x_n, y - y_n) \cdot e^{ik_z z}, \quad (2.54) \]

where \( f_n(x - x_n, y - y_n) \) denotes a basis function localized in the \( x - z \) plane and \( k_z \) is the transverse momentum that models the periodicity of the system. As a consequence of the periodic assumption, the total Hamiltonian matrix \( H(k_z) \) shows a \( k_z \)-dependence if Eq. (2.54) is plugged into Eq. (2.47). It can be decomposed into three components: \( H_0 \) (interactions within a strip of width \( \Delta \)) and \( H_{\pm} \) (interactions with a neighbor strip along the \( +z \) or \( -z \) direction). This gives

\[ H(k_z) = H_0 + H_- \cdot e^{-ik_z \Delta} + H_+ \cdot e^{ik_z \Delta}. \quad (2.55) \]

Considering the \( k_z \) dependence of the Hamiltonian, the equations for the retarded-, and the lesser/greater Green’s functions modify as

\[
\sum_l \{ [E_n - V(R_n)] \delta_{l,n} - H_{nl}(k_z) - \Sigma_{nl}^{RB}(E, k_z) 
- \Sigma_{nl}^{RS}(E, k_z) \} \quad G_{lm}^{R}(E, k_z) = \delta_{nm}, \quad (2.56)

\]

\[
G_{nm}^{\geq}(E, k_z) = \sum_{l_1, l_2} G_{nl_1}^{R}(E, k_z) \cdot \left[ \Sigma_{l_1l_2}^{\geq B}(E, k_z) + \Sigma_{l_1l_2}^{\geq S}(E, k_z) \right] \cdot G_{l_2m}^{A}(E, k_z). \quad (2.57)
\]

The indices \( l, m, \) and \( n \) refer to the positions \( R_l, R_m, \) and \( R_n \) that coincide with the center of the localized basis functions, usually atom positions. The matrices \( E_n \) and \( V(R_n) \) are diagonal and contain the electron energy \( E \) and the self-consistent electrostatic potential \( V \) at \( R_n \), respectively. They are both of size \( p \times p \), where \( p \) is the number of basis functions centered on the \( n \)-th atom. The Hamiltonian block \( H_{nm}(k_z) \) contains the matrix elements between the atoms with index \( n \) and \( m \), and is of size \( p \times q \). The boundary \( \Sigma_{nm}^{B}(E, k_z) \) and scattering \( \Sigma_{nm}^{S}(E, k_z) \) self-energies as well as the lesser, \( G_{nm}^{<}(E, k_z) \), greater, \( G_{nm}^{>}(E, k_z) \), retarded, \( G_{nm}^{R}(E, k_z) \), and advanced, \( G_{nm}^{A}(E, k_z) = G_{nm}^{R\dagger}(E, k_z) \) Green’s functions are of size \( p \times q \).
too and depend on the electron energy $E$ and its transverse momentum $k_z$.

Each $k_z$-dependent charge contribution within the first 1D Brillouin zone must be summed up to obtain the total charge density

$$n(r) = -i \sum_j \int_{-\pi}^{\pi} \frac{k_z}{2\pi} \int \frac{dE}{2\pi} \text{tr} \, G_{jj}^< (E, k_z) \delta(r - R_j). \quad (2.58)$$

Here the reciprocal lattice vector is assumed to be normalized to $2\pi$. The net electric current can be derived from its time-dependent definition

$$J(r, t) = \frac{\hbar^2}{2m_0} \lim_{r' \to r} (\nabla_{r'} - \nabla_r) \, G^< (r, t, r', t'), \quad (2.59)$$

or using the commutator relation $[\hat{r}, \hat{H}] = \frac{\hbar^2}{m} \nabla_r$

$$J(r, t) = \frac{1}{2} \lim_{r' \to r} \left( [\hat{r}, \hat{H}] - [\hat{r}', \hat{H}] \right) G^< (r, t, r', t'). \quad (2.60)$$

By taking the Fourier transform of this expression, integrating over all possible energies, and approximating the $r' \to r$ limes on the discrete atomic grid, the electric current is given as

$$J(r) = \frac{q}{\hbar} \sum_{m,n} \int_{-\pi}^{\pi} \frac{k_z}{2\pi} \int \frac{dE}{2\pi} \text{tr} \left( H_{mn}(k_z) G^<_{nm}(E, k_z) \right.
\left. - G^<_{mn}(E, k_z) H_{nm}(k_z) \right) (R_n - R_m) \delta(r - R_m). \quad (2.61)$$

To simplify the calculations, atoms may be grouped in unit cells such that each of them is only connected to its nearest-neighbors along the transport axis $x$. Following this definition, the total current flowing from the $i$-th unit cell to the next one can be evaluated with

$$J_d(x_i) = \frac{q}{\hbar} \int_{-\pi}^{\pi} \frac{k_z}{2\pi} \int \frac{dE}{2\pi} \text{tr} \left( H_{i,i+1}(k_z) G^<_{i+1,i}(E, k_z) \right.
\left. - G^<_{i,i+1}(E, k_z) H_{i+1,i}(k_z) \right). \quad (2.62)$$
2.3. TRANSPORT IN THE NEGF FORMALISM

The indices \( i \) and \( i + 1 \) run over all the basis functions situated within the \( i \)-th and \( i + 1 \)-th unit cells, respectively.

### 2.3.3 Electron-phonon coupling

Two distinct electron-phonon interaction mechanisms are considered in this work: scattering on deformation potentials and polar-optical scattering through the Fröhlich interaction when simulating polar materials. The scattering self-energies are the sum of the contributions associated with each of these two mechanisms

\[
\Sigma^S_{nl}(E, k_z) = \Sigma^{S, \text{def}}_{nl}(E, k_z) + \Sigma^{S, \text{Fröhl}}_{nl}(E, k_z),
\]

with the superscript 'def' and 'Fröhl' denoting scattering on deformation potentials and through Fröhlich interactions, respectively. In our approach the lesser/greater self-energies describing electron-phonon scattering caused by deformation potentials take the following form after expanding the Hamiltonian operator in a Taylor series with respect to the variations of the atom positions [76]

\[
\begin{align*}
\Sigma^{\geq S, \text{def}}_{nn}(E, k_z) &= \sum_{l,i,j} \int_{-\pi}^{\pi} \frac{dq_z}{2\pi} \sum_{\omega_{PH}} \nu_{nlln}^{ij}(\omega_{PH}, q_z) \cdot \nabla_i H_{nl}(k_z) \cdot \nabla_j H_{ln}(k_z), \\
\left\{ N^{PH}(\omega_{PH}) G_{ll}^{\geq}(E \pm \hbar \omega_{PH}, k_z - q_z) + \\
[N^{PH}(\omega_{PH}) + 1] \cdot G_{ll}^{\geq}(E \mp \hbar \omega_{PH}, k_z - q_z) \right\} \\
\cdot \nabla_j H_{ln}(k_z),
\end{align*}
\]

Only the diagonal entries of \( \Sigma^{R/\geq S, \text{def}}(E, k_z) \) are kept for computational reasons. The consequences of this are discussed in Ref. [77]. In Eq. (2.64) \( N^{PH}(\omega_{PH}) \) is the phonon occupancy (Bose-Einstein distribution) at frequency \( \omega_{PH} \) and \( \nabla_i H_{nl} \) is the derivative of the \((n,l)\)-th block of the Hamiltonian along the \( i \)-th Cartesian coordinate. The coupling factor \( \nu \) is computed by integrating the electron-phonon matrix elements \( V_{nlln}^{ij} [\omega_{\lambda}(q_x, q_z)] \) over those \( q_x \) points in the Brillouin zone where a phonon eigenstate with frequency \( \omega_{\lambda}(q_x, q_z) \) exists within a range \( \Delta E/\hbar \) around \( \omega_{PH} \).
\[ v_{nlln}^{ij}(\omega_{PH}, q_z) = \sum_{\lambda} \int_{-\pi}^{\pi} \frac{dq_x}{2\pi} \cdot V_{nlln}^{ij}[\omega_{\lambda}(q_x, q_z)], \quad (2.65) \]

with the condition

\[ \omega_{PH} - \Delta E/\hbar \leq \omega_{\lambda}(q_x, q_z) \leq \omega_{PH} + \Delta E/\hbar. \quad (2.66) \]

The electron-phonon matrix elements are defined as

\[ V_{nlln}^{ij}[\omega_{\lambda}(q_x, q_z)] = \frac{\hbar}{2\omega_{\lambda}(q_x, q_z)} \cdot \left[ \frac{f^i_{\lambda}(R_l, q_x, q_z)}{\sqrt{M_l}} - \frac{f^i_{\lambda}(R_n, q_x, q_z)}{\sqrt{M_n}} \right] \cdot \left[ \frac{f^j_{\lambda}*(R_n, q_x, q_z)}{\sqrt{M_n}} - \frac{f^j_{\lambda}*(R_l, q_x, q_z)}{\sqrt{M_l}} \right], \quad (2.67) \]

where \( f^i_{\lambda}(R_l, q_x, q_z) \) is the \( i^{th} \) Cartesian component of the \( l^{th} \) ion’s displacement in the normalized phonon eigenmode corresponding to wavevector \( (q_x, q_z) \) and mode \( \lambda \). The variable \( M_l \) refers to the mass of the ion situated at \( R_l \).

The scattering self-energy due to interactions with polar optical phonons has a form similar to Eq. (2.64), except that the electron-phonon coupling is different

\[ \Sigma_{nn}^{\text{Fr"oh}}(E, k_z) = \int_{-\pi}^{\pi} \frac{dq_x}{2\pi} \int_{-\pi}^{\pi} \frac{dq_z}{2\pi} |g_{\text{Fr"oh}}(q_x, q_z)|^2 \cdot \left\{ N^{PH}(\omega_{opt})G_{nn}^>(E \pm \hbar\omega_{opt}, k_z - q_z) + \left[ N^{opt}(\omega_{opt}) + 1 \right] \cdot G_{nn}^<(E \mp \hbar\omega_{opt}, k_z - q_z) \right\}. \quad (2.68) \]

Accurate estimation of the Fröhlich coupling constant \( |g_{\text{Fr"oh}}(q_x, q_z)| \) is essential to evaluate Eq. (2.68), that only involves the optical phonon frequency \( \omega_{opt} \).

Finally, note that the expression for the retarded self-energy in Eq. (2.48) the principal integral term is neglected so that
\[ \Sigma_{nn}^{R,S,\text{def/Fröh}}(E, k_z) \approx \frac{1}{2} \left( \Sigma_{nn}^{>\text{def/Fröh}}(E, k_z) - \Sigma_{nn}^{<\text{def/Fröh}}(E, k_z) \right). \] (2.69)

The single-particle Hamiltonian, the derivatives of the Hamiltonian matrix elements, the Fröhlich interaction constants, the phonon frequencies, and the phonon modes can be determined from density-functional theory and density-functional perturbation theory calculations, as shown later.

### 2.3.4 Mobility calculations

Self-consistently solving the NEGF and Poisson’s equations in the presence of dopants, gate contacts separated from the channel by oxide layers, and at various drain-to-source \( V_{ds} \) and gate-to-source \( V_{gs} \) voltages allows us to simulate the operation of field-effect transistors. Calculating the electron- and hole-mobilities in a material is, however, computationally less demanding since it does not require the consideration of many bias points. The mobility values are derived from the calculation of the drain current \( I_d(L) \) and channel resistance \( R(L) = V_{ds}/\Delta I_d(L) \) of 2D semiconductor samples with different lengths \( L \). The required simulations are performed in structures with a homogeneous electron density \( n_{2D} \), a small bias difference \( V_{ds} = 10^{-3} \) V between the two source and drain contacts, and in the presence of electron-phonon scattering. Due to the diffusive nature of electron transport under these conditions, the phonon-limited electron/hole mobility \( \mu_{ph} \) can be calculated with the so-called ”dR/dL” method \[78\], starting with the extraction of the channel resistivity \( \rho \)

\[ \rho = \frac{dR}{dL} = \frac{V_{ds}}{\Delta L} \left( \frac{1}{I_d(L)} - \frac{1}{I_d(L+\Delta L)} \right). \] (2.70)

In this work the initial sample length \( L \) is set to 40 nm and \( \Delta L \) to 20 nm. To verify that the channel resistance \( R(L) \) indeed linearly increases with the sample length, we have also considered the case \( L = 60 \) nm, \( \Delta L = 20 \) nm and demonstrated that the mobility value remains the same. The mobility \( \mu_{ph} \) is finally obtained as

\[ \mu_{ph} = \frac{1}{q \cdot n_{2D} \cdot \rho}, \] (2.71)
where \( q \) denotes the elementary charge. The expression for the hole mobility is identical, the hole concentration \( p_{2D} \) replacing \( n_{2D} \).

## 2.4 Density-functional theory

One of the key ingredients of the NEGF equations is the Hamiltonian, for which we have made only one assumption so far, it should be expressed in a localized basis. Tight-binding fulfills this requirement, but its empirical character represents a severe limitation in many applications. Density-functional theory (DFT) within a localized basis such as a linear combination of atomic orbitals \([79]\) or contracted Gaussian functions \([80]\) might do the trick, but at high computational cost due to the relatively large spreads of the basis components. Here, we have opted for an hybrid scheme that relies on plane-wave DFT and a transformation of the output into a localized Hamiltonian. With DFT a formally non-interacting Hamiltonian can be created for any atomic system, that already contains the equilibrium electron-electron and electron-ion interactions. In this Section the foundations of the theory will be briefly revised.

### 2.4.1 The Kohn-Sham equations

After applying the Born-Oppenheimer approximation to the time-dependent Schrödinger equation, the Hamiltonian \( \hat{H} \) that acts on the electronic wavefunction \( \Psi(\mathbf{r}_1, \ldots \mathbf{r}_N) \)

\[
\hat{H}\Psi(\mathbf{r}_1, \ldots \mathbf{r}_N) = E\Psi(\mathbf{r}_1, \ldots \mathbf{r}_N)
\]  

(2.72)

can be written in the following form using atomic units

\[
\hat{H} = \sum_i^N -\frac{1}{2} \nabla_i^2 + \sum_i^N V(\mathbf{r}_i) + \sum_{i<j}^N U(\mathbf{r}_i, \mathbf{r}_j).
\]  

(2.73)

Here, \( V \) is the static external potential field generated by the fixed nuclei and \( N \) the number of electrons. The three terms present in the definition of the Hamiltonian are usually denoted as \( \hat{T} \) (kinetic energy), \( \hat{V} \) (external potential), and \( \hat{U} \) (electron-electron interactions),
respectively. For a fixed number of particles it is only the potential $\hat{V}$ that is system-dependent.

The basic idea behind density-functional theory (DFT) is that in order to determine the ground state energy and every other observables of a given structure, the knowledge of the many-particle wavefunction is not necessary. It is sufficient to determine the spatial electron density

$$n(r) = N \int dr_2 \ldots \int dr_N \Psi^*(r, r_2 \ldots r_N) \Psi(r, r_2 \ldots r_N). \tag{2.74}$$

In Eq. (2.74) $\int dr_i$ refers to the integration over the coordinates of the $i$-th variable of the many-particle wavefunction $\Psi$. It is clear from this equation that $n(r)$ is a functional of $\Psi(r_1, \ldots r_N)$. For non-degenerate ground-states the wavefunction also uniquely depends on the potential $\hat{V}$ that defines the Hamiltonian corresponding to a fixed number of electrons. Therefore, the $\hat{V} \rightarrow n(r)$ relationship is unequivocal. Hohenberg and Kohn proved [81] that the reverse holds too, i.e. there exists exactly one $\hat{V}$, apart from an additive constant, that produces the prescribed charge density. Consequently, the ground-state wavefunction and energy are also unique functionals of the density since they are determined through the Schrödinger equation. Taking all these features into account the ground-state energy $E_{0,V}$ of a system with a specific potential $\hat{V}$ can be written in a functional form

$$E_{0,V} = E_V[n_0(r)] = \langle \Psi[n_0(r)]|\hat{T} + \hat{U} + \hat{V} |\Psi[n_0(r)]\rangle, \tag{2.75}$$

where $n_0(r)$ is the ground-state density. The more general energy functional

$$E_V[n(r)] = \langle \Psi[n(r)]|\hat{T} + \hat{U} + \hat{V} |\Psi[n(r)]\rangle \tag{2.76}$$

can be expressed as

$$E_V[n(r)] = F[n(r)] + \langle \Psi[n(r)]|\hat{V} |\Psi[n(r)]\rangle, \tag{2.77}$$

where $F[n(r)]$ is defined as
The variable $F$ is called universal functional since it does not depend on the specific system. Using Eq. (2.74), the expectation value of the potential can be explicitly expressed as

$$\langle \Psi[n(r)] | \hat{V} | \Psi[n(r)] \rangle = \int dV(r)n(r).$$  \hspace{1cm} (2.79)

The Hohenberg-Kohn theorem [81] also states that the energy functional reaches its minimum when the density $n(r)$ equals to the ground-state density $n_0(r)$, which can therefore be obtained by minimizing the energy functional

$$E_V[n(r)] = F[n(r)] + \int dV(r)n(r).$$  \hspace{1cm} (2.80)

This many-body problem can be mapped onto a single-body problem using an auxiliary system of $N$ non-interacting electrons that has the same density as the original one with interacting electrons. The energy functional is then expressed using the kinetic part of the non-interacting system $T_s$, its external potential, and its Coulomb-interactions. All differences between the real and the artificial system are cast into an exchange-correlation energy term $E_{xc}[n(r)]$

$$E_V[n(r)] = T_s[n(r)] + \int dV(r)n(r) + \frac{1}{2} \int drdr' \frac{n(r)n(r')}{|r-r'|} + E_{xc}[n(r)],$$  \hspace{1cm} (2.81)

with

$$E_{xc}[n(r)] = T[n(r)] - T_s[n(r)] + U[n(r)] - \frac{1}{2} \int drdr' \frac{n(r)n(r')}{|r-r'|}.$$  \hspace{1cm} (2.82)

The $\delta E_V[n] = 0$ minimum criteria along with the orthonormality constraint on the single-particle wavefunctions $\langle \phi_i | \phi_j \rangle = \delta_{ij}$ leads to the following effective Schrödinger equations

$$\left(-\frac{1}{2} \nabla^2 + V_s(r)\right) \phi_i(r) = \varepsilon_i \phi_i(r),$$  \hspace{1cm} (2.83)
with the single-particle effective potential $V_s$ defined as

$$V_s(r) = V(r) + \int d\mathbf{r}' \frac{n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} + V_{xc}(\mathbf{r}).$$ \hfill (2.84)

In Eq. (2.84) $V_{xc}(\mathbf{r})$ is the functional derivative of the exchange-correlation energy

$$V_{xc}(\mathbf{r}) = \frac{\delta E_{xc}[n(\mathbf{r})]}{\delta n(\mathbf{r})}. \hfill (2.85)$$

Equations (2.83 2.85) are called the Kohn-Sham equations \cite{2}. Due to the locality of the potential $V_s$, the computational time associated with solving these equations scales only with the third power of the system size $N$, contrary to the $N^4$-dependence of the Hartree-Fock method, where two-center integrals must be computed.

The charge density can be easily reconstructed from the orthogonal Kohn-Sham orbitals

$$n(\mathbf{r}) = \sum_i^n |\phi_i(\mathbf{r})|^2. \hfill (2.86)$$

Since the single-particle effective potential depends on the charge, Eqs. (2.83 2.86) form a self-consistent system of equations that has to be solved iteratively, as illustrated in Fig. 2.3.

\subsection*{2.4.2 The exchange-correlation functional}

Although the derivation of the Kohn-Sham equations above does not require any approximation, the exact form of the exchange-correlation functional is only known in the case of slowly varying densities or high densities, where it can be approximated as

$$E_{xc}^{LDA}[n(\mathbf{r})] = \int d\mathbf{r} n(\mathbf{r}) \epsilon_{xc}[n|_{\mathbf{r}}], \hfill (2.87)$$

with $\epsilon_{xc}[n|_{\mathbf{r}}]$ at a given $\mathbf{r}$ point as the exchange-correlation energy density of the uniform electron gas with density $n(\mathbf{r})$. This is the local density approximation (LDA). The exchange-correlation energy can be further split into an exchange- and a correlation part
Figure 2.3: Flowchart for iteratively solving the self-consistent equations of density-functional theory. \( \{ \phi_i^{(0)}(r) \}_i^N \) is the set of initial single-particle wavefunctions, \( n^{(i)}(r) \) the density at the \( i \)-th step, \( V_s \) the effective potential that depends on \( n^{(0)}(r) \), and \( \delta_{tol} \) the tolerance parameter that defines the stopping criteria for the iteration.

\[
\epsilon_{xc}[n] = \epsilon_x[n] + \epsilon_c[n]. \quad (2.88)
\]

The exchange energy density of a homogeneous electron gas can be analytically expressed employing the Thomas-Fermi model.
2.4. DENSITY-FUNCTIONAL THEORY

\[ \epsilon_x[n] = -\frac{3}{4} \left( \frac{3n}{\pi} \right)^{1/3}. \]  

(2.89)

No such analytical expression exists for the correlation part, but it can be calculated numerically with high accuracy using Monte-Carlo simulations. Ceperley and Alder performed such calculations for a wide range of densities [82] and their results were parametrized in an analytic form by Perdew and Zunger [83] as

\[ \epsilon_c[r_s] = \begin{cases} 
-0.1423/(1 + 1.09529\sqrt{r_s} + 0.3334r_s) & r_s > 1, \\
-0.048 + 0.0311 \ln r_s - 0.0116r_s + 0.002r_s \ln r_s & r_s \leq 1,
\end{cases} \]

(2.90)

with

\[ r_s = \left( \frac{3}{4\pi n} \right)^{1/3}. \]

(2.91)

While LDA has been very successfully applied to many quantum chemistry problems, it significantly underestimates the band gap of most semiconductors, in some cases even by 50%, which renders the physical meaning of the Kohn-Sham orbitals questionable. The accuracy of the method can be improved by taking into account the gradient of the electron density beside the local density itself. In the generalized gradient approximation (GGA) the exchange-correlation functional is expressed as the spatial integral of a function that depends both on the density and on the gradient of the density at a given point. Further accounting for the spin components \( \alpha \) and \( \beta \) the GGA functional has the following form

\[ E^{GGA}_{xc}[n_\alpha(r), n_\beta(r)] = \int d\mathbf{r} f(n_\alpha|\mathbf{r}, n_\beta|\mathbf{r}, \nabla n_\alpha|\mathbf{r}, \nabla n_\beta|\mathbf{r}). \]  

(2.92)

The GGA functional parametrized by Perdew, Burke, and Ernzerhof [84] (PBE) provided significant improvements in the single-electron orbital energies, without the introduction of any empirical parameters.
It is possible to reproduce the band gap of most semiconductors even more accurately with the use of hybrid functionals, where the exchange part of the functional is calculated as a linear combination of the Hartree-Fock exchange and the GGA exchange

$$E_{xc}^{\text{hybrid}} = aE_{x}^{HF} + (1 - a)E_{x}^{\text{GGA}} + E_{c}^{\text{GGA}}.$$ (2.93)

Here the mixing tag $a$ is a fitting parameter. The Coulomb interaction is usually further divided into a short- and a long-range part. The mixing with the Hartree-Fock exchange is only applied to the short-range part. The parametrization of the HSE06 – the most widely used hybrid functional – was performed by Heyd, Scuseria, and Ernzerhof [85]. The HSE06 functional was demonstrated to reproduce the formation enthalpies of an extensive range of molecules, as well as the lattice constants and band gaps of many metals, insulators, and semiconductors with a previously unattainable precision [86]. The two-center integrals of the Hartree-Fock method, however, severely limit the scalability of hybrid functional simulations.

### 2.5 Maximally localized Wannier functions

Most DFT codes use plane-waves to expand the single-particle Kohn-Sham wavefunctions. Although this basis set enables the calculation of electronic states with high accuracy, it does not naturally lend itself to transport simulations, where localized basis functions are highly desired. The localized, real-space representation of Bloch states introduced by Wannier in the 1930’s [87] can circumvent this problem. More recently, Marzari and Vanderbilt developed a formally and computationally attractive method for the construction of such Wannier functions (WFs) with a maximal localization criteria [5]. It was later implemented in the Wannier90 code [88]. The transformation of the Bloch Hamiltonian into the basis spanned by the maximally localized Wannier functions (MLWFs) yields a “maximally sparse” block-tridiagonal matrix that is essential for efficient atomic-level simulations of devices with realistic physical sizes. The interested reader is referred to the exhaustive 2012 paper of Marzari et al. in Rev. Mod. Phys. [89]. Here, only the basic theory behind MLWFs that is essential for the understanding of this work will be reviewed.
Furthermore, in Sec. 2.5.4 a simple technique for the extraction of the derivatives of the MLWF Hamiltonian matrix elements with respect to the ionic displacements is introduced.

### 2.5.1 The Wannier transformation

In periodic systems the eigenstates $\Psi$ of the Hamiltonian $H$ are Bloch functions formed by a product of a cell-periodic $u(r)$, and an envelope-function $e^{ikr}$

$$
\Psi_{nk}(r) = u_{nk}(r)e^{ik\cdot r},
$$

(2.94)

where $k$ is a wave-vector and $n$ a band index. It is obvious from Eq. (2.94) that Bloch functions are delocalized and spread over the entire space. It is possible, however, to construct a localized wave-packet out of the Bloch functions by superposing the contributions from different $k$ values. The simplest possible solution is to assign equal weights to each point in the Brillouin zone. The resulting function

$$
w_n(r) = \frac{V}{(2\pi)^3} \int_{BZ} dk \Psi_{nk}(r)
$$

(2.95)

is no more periodic, but localized around the origin. The $V/(2\pi)^3$ factor in Eq. (2.95) is a normalization constant with $V$ as the volume of the unit cell. The integration is carried out over the entire Brillouin Zone (BZ). Additionally, a phase factor $e^{-ik\cdot R}$ can be inserted into the integrand, which shift $w_n(r)$ by $R$, with $R$ being a real-space lattice vector [87]. In bra-ket notation $|R_n\rangle$ refers to the $n$-th Wannier function (WF) shifted to the unit cell located at $R$

$$
|R_n\rangle = \frac{V}{(2\pi)^3} \int_{BZ} dk e^{-ik\cdot R} |\Psi_{nk}\rangle.
$$

(2.96)

Equation (2.96) is formally a Fourier-transformation, whose inverse is defined as

$$
|\Psi_{nk}\rangle = \sum_R e^{ik\cdot R} |R_n\rangle.
$$

(2.97)
Since Eqs. (2.96) and (2.97) are unitary transformations, the set of Bloch functions and the set of WFs constitute an equally valid description of the \( n \)-th band’s subspace, with the trace of any one-particle operator being identical in both bases.

The Wannier functions are not unique, because of the gauge freedom in the definition of the Bloch functions. That is, replacing the \( |\Psi_{nk}\rangle \) Bloch functions by

\[
|\tilde{\Psi}_{nk}\rangle = e^{i\phi_n(k)} |\Psi_{nk}\rangle,
\]

(2.98)

describes the exact same physical system if \( \phi_n(k) \) is a real and periodic function in reciprocal space. This non-uniqueness of the Bloch functions appears in the Wannier functions too, so that different gauges will result in WFs with different shapes. More generally, if there are \( J \) bands that are separated from the lower and higher energy bands everywhere in the Brillouin zone, even with internal degeneracies and crossings among each other, any unitary transformation \( U(k) \) of size \( J \times J \) still gives a physically identical description of the selected subspace of \( J \) bands

\[
|\tilde{\Psi}_{nk}\rangle = \sum_{m} U_{mn}(k) |\Psi_{mk}\rangle.
\]

(2.99)

Hence the transformation to Wannier functions from the Bloch space has a general form

\[
|\mathbf{R}n\rangle = \frac{V}{(2\pi)^3} \int_{BZ} d\mathbf{k} e^{-i\mathbf{k} \cdot \mathbf{R}} \sum_{m} U_{mn}(k) |\Psi_{mk}\rangle.
\]

(2.100)

The \( |\mathbf{R}n\rangle \) WFs form an orthonormal basis, as the Bloch functions do. The Bloch Hamiltonian \( H(k) \) can therefore be transformed into the WF basis with

\[
H_{ij}(\mathbf{R}) = \sum_{l,m} \frac{V}{(2\pi)^3} \int_{BZ} d\mathbf{k} e^{-i\mathbf{k} \cdot \mathbf{R}} U_{il}(k)^\dagger H_{lm}(k) U_{mj}(k).
\]

(2.101)

In Eq. (2.101) \( H(\mathbf{R})_{ij} \) is the matrix element between the \( i \)-th Wannier function in the unit cell at the origin and the \( j \)-th Wannier
function in the unit cell at \( \mathbf{R} \), in other words \( H_{ij}(\mathbf{R}) = \langle 0i|H|R,j \rangle \). Since the Wannier functions in different unit cells, but with the same in-cell indices, are periodic replicas of each other, only the difference of two unit cell vectors is meaningful. Hence, a single lattice index is sufficient for describing the WF Hamiltonian. Also, note that the Bloch Hamiltonian \( H(\mathbf{k}) \) is a diagonal matrix in the Bloch basis with the eigenenergies of the Bloch orbitals retained for the Wannier transformation as non-zero elements, since the Bloch wavefunctions are the eigenvectors of \( H(\mathbf{k}) \).

The freedom in choosing the Wannier functions resides in the non-uniqueness of the unitary gauge transformation \( \mathbf{U}(\mathbf{k}) \). For transport simulations the desired outcome of this transformation should be as localized as possible Wannier functions, and consequently the rapid decay of the Hamiltonian matrix elements with the separation between two WFs.

### 2.5.2 Maximal localization

The localization of the Wannier functions in real-space is related to the smoothness of the Bloch functions in the reciprocal space characterized by the derivative of the Bloch function with respect to its \( \mathbf{k} \) vector \( \nabla_{\mathbf{k}} |\tilde{\Psi}_{nk}\rangle \). One way to find a smooth gauge is to start from a set of already localized trial orbitals \( g_n(\mathbf{r}) \), e.g. atomic orbitals, and project them onto the Bloch manifold

\[
|\Phi_{nk}\rangle = \sum_{m}^{J} |\Psi_{mk}\rangle \langle \Psi_{mk}|g_{n}\rangle . \tag{2.102}
\]

The resulting \( |\Phi_{nk}\rangle \) functions are typically smooth in \( \mathbf{k} \)-space, but they are not orthonormal. This issue can be fixed by applying Löwdin orthonormalization technique \[90\]. By introducing the matrix of the inner products

\[
A(\mathbf{k})_{mn} = \langle \Psi_{mk}|g_{n}\rangle , \tag{2.103}
\]

the overlap integral over the unit cell volume \( S_{mn}(\mathbf{k}) = \langle \Phi_{mk}|\Phi_{nk}\rangle \) can be expressed as \( S_{mn} = (A(k)^\dagger A(k))_{mn} \), and the Löwdin orthonormalized states are given by
\[ |\tilde{\psi}_{nk}\rangle = \sum_m^J \left(S^{-1/2}(k)\right)_{mn} |\Phi_{mk}\rangle, \quad (2.104) \]

so that the gauge transformation matrix \( U(k) = A(k)S(k)^{-1/2} \).

It is important to note that the above method is not the same as projecting the Bloch states onto atomic orbitals. The latter is not a unitary transformation, hence the resulting states will not give the exact same description of the original system. Instead, projecting some atomic orbitals onto the Bloch manifold will produce localized Wannier functions in most cases after the Löwdin-orthonormalization and a Fourier transformation.

The Wannier functions obtained by projection are often sufficiently localized, but in general they are not maximally localized. In order to find the most localized states, a well-defined localization criteria is needed first. The most widely used criteria is the one introduced by Marzari and Vanderbilt \[5\], where the localization is measured by the sum of the quadratic spreads of the Wannier functions within the unit cell. The problem is then rendered to finding the gauge transformation matrix \( U(k) \) that minimizes the localization functional

\[ \Omega = \sum_n \left[ \langle 0n|r^2|0n\rangle - \langle 0n|r|0n\rangle^2 \right] = \sum_n \left[ \langle r^2 \rangle_n - \bar{r}^2_n \right]. \quad (2.105) \]

This functional can be further split into a gauge-invariant \( \Omega_I \) and a gauge-dependent \( \tilde{\Omega} \) part

\[ \Omega = \Omega_I + \tilde{\Omega}, \quad (2.106) \]

with

\[ \Omega_I = \sum_n \left[ \langle 0n|r^2|0n\rangle - \sum_{Rm} |\langle Rm|r|0n\rangle|^2 \right], \quad (2.107) \]

and

\[ \tilde{\Omega} = \sum_n \sum_{Rm \neq 0n} |\langle Rm|r|0n\rangle|^2. \quad (2.108) \]
2.5. MAXIMALLY LOCALIZED WANNIER FUNCTIONS

The minimization of $\Omega$ with respect to $U(k)$ therefore corresponds to the minimization of the gauge-dependent term $\tilde{\Omega}$ only. Since density-functional theory codes produce Bloch functions, the equations governing $\tilde{\Omega}$ have to be connected to the Bloch formalism. In reciprocal space the matrix elements of $r$ and $r^2$ between WFs take the following form \[91\]

$$
\langle R_n | r | 0_m \rangle = i V \left( \frac{2\pi}{\lambda} \right)^3 \int d\mathbf{k} e^{i\mathbf{k} \cdot \mathbf{R}} \langle u_{nk} | \nabla_{k} | u_{mk} \rangle ,
$$

(2.109)

and

$$
\langle R_n | r^2 | 0_m \rangle = - i V \left( \frac{2\pi}{\lambda} \right)^3 \int d\mathbf{k} e^{i\mathbf{k} \cdot \mathbf{R}} \langle u_{nk} | \nabla_{k}^2 | u_{mk} \rangle .
$$

(2.110)

The expectation values between $u_{nk}$ and $u_{mk}$ involving $\nabla_{k}$ and $\nabla_{k}^2$ can be computed using finite differences. For any $f(k)$ that is a smooth function of $k$ the gradient $\nabla f(k)$ can be written as

$$
\nabla f(k) = \sum_{b} w_b b [f(k + b) - f(k)] + O(b^2),
$$

(2.111)

and its square

$$
|\nabla f(k)|^2 = \sum_{b} w_b b [f(k + b) - f(k)]^2 + O(b^3),
$$

(2.112)

with $b$ being a vector connecting $k$ with one of its neighbor and $w_b$ an appropriate geometric factor, depending on the actual $k$ grid. In Wannier90 the procedure is implemented for regular Monkhorst-Pack grids \[92\]. The method for choosing the set of $b$ vectors and $w_b$ values is discussed in Ref. \[88\]. Using these finite differences expressions, all the required quantities to calculate the localization functional $\Omega$ through Eqs. (2.109) and (2.110) are represented by the $M_{mn}^{(k,b)}$ overlaps between Bloch orbitals at neighbor $k$ points

$$
M_{mn}^{(k,b)} = \langle u_{mk} | u_{n,k+b} \rangle .
$$

(2.113)
The minimization of the spread functional $\Omega$ is performed by considering an infinitesimal gauge transformation matrix $U(k)_{mn} = \delta_{mn} + dW(k)_{mn}$, where $dW$ is an infinitesimal anti-Hermitian matrix. The gradient $G_{mn} = d\Omega/dW_{nm}$ is calculated by first defining

$$q^{k,b}_n = \text{Im} \ln M^{(k,b)}_{mn} + b \cdot \bar{r}_n,$$

(2.114)

together with the operators

$$A[B] = (B - B^\dagger)/2,$$

(2.115)

and

$$S[B] = (B + B^\dagger)/2i.$$

(2.116)

Putting all these together, the final expression for the gradient of the spread functional is found

$$G(k) = 4 \sum_b w_b \left( A \left[ R^{(k,b)} \right] - S \left[ T^{(k,b)} \right] \right),$$

(2.117)

where

$$R^{(k,b)}_{mn} = M^{(k,b)}_{mn} M^{(k,b)*}_{nn},$$

(2.118)

and

$$T^{(k,b)}_{mn} = q^{(k,b)}_n M^{(k,b)}_{mn} / M^{(k,b)}_{nn}.$$

(2.119)

In the simplest implementation the gauge transformation matrix $U(k)$ is changed by small steps in the direction opposite to the gradient until the minimum of $\Omega$ is reached. It is common practice to start the minimization procedure from a reasonable guess for the MLWFs, i.e. from an $U(k)$ that is defined by projecting atomic orbitals onto the Bloch manifold and performing Löwdin orthonormalization, as in Eq. (2.104).

The overlap matrices $M^{(k,b)}$ also need to be updated after each step as

$$M^{(k,b)} = U(k)\dagger M^{(0)(k,b)} U(k + b),$$

(2.120)
with $\mathbf{M}^{(0)}(k,b)$ denoting the initial Bloch orbital overlaps supplied by any density-functional theory code. Therefore, once the calculation of the initial trial orbital projections $\mathbf{A}(k)$ and initial overlaps $\mathbf{M}^{(0)}(k,b)$ are done, the minimization of the spread of the Wannier functions (“Wannierization”) becomes independent of the DFT code that was utilized to generate the initial Bloch states.

2.5.3 Entangled bands

When the bands of interest do not form an isolated group, i.e. when they are not separated by a finite energy gap from all the other lower and higher energy bands at each $k$ point, the procedure for obtaining MLWFs has to be modified. We talk in this case about entangled bands. To handle this situation an energy interval can be set such that it contains all the bands selected for the Wannier transformation, plus some additional bands whose number is not necessarily constant and may depends on the position in the reciprocal space. The total number of bands $\mathcal{J}(k)$ is therefore larger or equal than the number of target Wannier functions $J$ at each $k$ point. The method of trial orbital projections is still a viable way to obtain a smooth gauge, but the number of bands in Eq. 2.102 has to be changed

$$|\Phi_{nk}\rangle = \sum_{m}^{\mathcal{J}(k)} |\Psi_{mk}\rangle \langle \Psi_{mk}|g_{n}\rangle. \quad (2.121)$$

The inner product matrices $\mathbf{A}(k)$ in Eq. 2.103 become rectangular with dimensions $\mathcal{J}(k) \times J$. The overlap matrix $\mathbf{S}(k)$ remains formally the same, $S_{mn} = (A(k)^{\dagger} A(k))_{mn}$, only with a modified size of $\mathbf{A}(k)$. The Löwdin-orthonormalized Bloch-like orbitals

$$|\tilde{\Psi}_{nk}\rangle = \sum_{m}^{J} (S^{-1/2}(k))_{mn} |\Phi_{mk}\rangle. \quad (2.122)$$

usually produce localized Wannier functions in the $J$-dimensional subspace of the bands of interest after the Fourier transformation. This gauge transformation can be used again as a starting point in the minimization of $\tilde{\Omega}$. However, in general the resulting WFs are not maximally localized. This is due to the fact that the gauge-invariant
part $\Omega_I$ of the spread functional depends on the subspace selection method and it is not minimized by the orbital projections, despite providing in many cases a reasonably small gauge-independent spread.

The gauge-independent spread functional $\Omega_I$ can also be minimized in an iterative way, similarly to $\tilde{\Omega}$. Since $\Omega_I$ does not depend on the unitary mixing of the Bloch orbitals, it measures the smoothness of the set of $J$ bands that are selected from the $J(k)$ bands in the considered energy interval, also called disentanglement window, rather than the smoothness of the individual bands. As demonstrated by Souza et al. [93], the minimization of $\Omega_I$, i.e. the selection of the set of $\{\tilde{u}_{nk}\}$ states such that $\delta\Omega_I(\{\tilde{u}_{nk}\}) = 0$ with orthonormality conditions, is equivalent to solving the following linear system of equations

$$
\left[ \sum_b w_b P_{k+b} \right] |\tilde{u}_{nk}\rangle = \lambda_{nk} |\tilde{u}_{nk}\rangle ,
$$

(2.123)

with $P_k$ being the gauge-invariant orbital projection operator at wavevector $k$

$$
P_k = \sum_n^J |\tilde{u}_{nk}\rangle \langle \tilde{u}_{nk}| .
$$

(2.124)

The system of equations (2.123) connects different $k$ points, which makes an iterative solution more efficient. The details of the algorithm implemented in Wannier90 are discussed in Ref. [93]. An important feature of the method is that it is easy to include further restrictions in the subspace selection. In general, due to the mixing of the $J(k)$ states, the resulting $J$ Wannier bands will not be identical to the initial $J$ Bloch bands selected for the transformation. It is, however, important to preserve the characteristics of those states close to the valence and conduction band edges in semiconductors and insulators, or close to the Fermi energy in metals. The problem is circumvented by introducing a second energy interval, an inner- or “frozen-” window that contains a maximum of $J$ bands everywhere in the Brillouin zone. The states within that interval are enforced to remain unchanged. Typically, for transport simulations a frozen window that spans from
2.5. **MAXIMALLY LOCALIZED WANNIER FUNCTIONS**

Figure 2.4: Band structure of single-layer MoS$_2$ (left) and graphene (right). The black lines are the Bloch bands obtained from VASP, the red circles are the Wannier-interpolated states. The Fermi energy is at 0 eV. In the MoS$_2$ case the top 7 valence and bottom 4 conduction bands are separated from the lower- and higher lying bands. The method implemented for a set of isolated bands can therefore be applied for the Wannier transformation of these states. However, in the case of graphene the bottom 5 bands need to be disentangled from the higher energy bands. The horizontal blue lines and the arrow mark the inner- “frozen” energy window utilized for the determination of the optimally connected (smoothest) subspace of the bottom 5 graphene bands.

1-2 eV below the valence band edge to 1-2 eV above the conduction band edge is necessary to preserve the important properties of the system. Figure 2.4 illustrates the difference between the case of composite- and entangled bands. A small group of bands around the Fermi energy of MoS$_2$ are isolated from the other bands, while in the case of graphene the bands of interest are entangled with the higher energy states.

Since $\Omega_I$ is gauge-independent, the minimization of the spread functional can be divided into two separate tasks: (i) the selection of the smoothest subspace of $J$ bands from the manifold at each $\mathbf{k}$ point, and (ii) the minimization of $\hat{\Omega}$ as described in Sec. 2.5.2.
The typical procedure is to start with the trial orbital projection method to obtain an initial guess for the disentanglement transformation \( U_{\text{dis}}^{(0)}(k) = A(k)S(k)^{-1/2} \), minimize \( \Omega_I \), then proceed with the minimization of \( \tilde{\Omega} \), and finally obtain \( U(k) \). The matrix required for the transformation of the Hamiltonian into the MLWF basis in Eq. (2.101) will be the product of \( U_{\text{dis}}(k)U(k) \) instead of just \( U(k) \) when the disentanglement method is applied.

\[
H_{ij}(R) = \sum_{l,m,p,q} \frac{V}{(2\pi)^3} \int_{BZ} dk e^{-ikR} \left( \Delta H_{lm}(k)U_{\text{dis},mq}(k)U_{qj}(k) \right).
\]

(2.125)

2.5.4 Derivatives of the MLWF Hamiltonian

In order to evaluate the electron-phonon scattering self-energies in Eq. (2.64), apart from the Hamiltonian \( H \), its derivatives \( \nabla_i H_{nl} \) with respect to the Cartesian coordinate \( i \) and along the “bond” joining the Wannier centers \( n \) and \( l \) are also required. The derivatives of the real-space MLWF Hamiltonian are computed by applying a small hydrostatic stress to the investigated material. This is achieved by changing the lattice vectors of the unit cell uniformly and calculating the resulting Bloch Hamiltonian with the selected DFT code. The difference between the strained \( H_s(k) \) and unstrained \( H_0(k) \) Hamiltonians \( \Delta H(k) = H_s(k) - H_0(k) \) can then be transformed into the MLWF basis with the same unitary transformation matrix \( U \) that was determined by Wannier90 in the unstrained case

\[
\Delta H_{ij}(R) = \sum_{l,m} \frac{V}{(2\pi)^3} \int_{BZ} dk e^{-ikR} U_{il}(k)\Delta H_{lm}(k)U_{mj}(k).
\]

(2.126)

If the disentanglement method was used, then \( U_{\text{dis}}(k)U(k) \) should take the place of \( U(k) \) in Eq. (2.126). Due to the linearity of the Wannier transformation, this definition of \( \Delta H(R) \) is equivalent to

\[
\Delta H(R) = H_s(R) - H_0(R),
\]

(2.127)
2.5. **MAXIMALLY LOCALIZED WANNIER FUNCTIONS**

Figure 2.5: Difference between the MLWF Hamiltonian matrix elements with- and without strain. The couplings between one Wannier function situated on a Mo atom and three different WFs on one of its S neighbors are shown for a (a) double-layer MoS$_2$ with the selected Wannier function on the molybdenum having a $d_{z^2}$ orbital character and (b) a triple-layer MoS$_2$ with a $d_{xz}$-like orbital on the Mo atom. The green down-pointing triangles, red up-pointing triangles, and blue squares correspond to $p_x$, $p_y$, and $p_z$-like MLWFs on the sulfur atom. As expected, $\Delta H$ linearly depends on the applied relative strain.

with $H_s(R)$ and $H_0(R)$ being the strained and unstrained MLWF Hamiltonians produced from $H_s(k)$ and $H_0(k)$ with the transformation defined in Eq. (2.101), but with one single transformation matrix $U$ (or $U_{\text{dis}} U$). With two distinct $U$’s, the MLWF basis sets might be different in the strained and unstrained cases so that the derivatives $\nabla_i H_{nl}$ have no meaning any more. This trick is indeed needed because the MLWF transformation is not unique and it may strongly depend on small variations of the atomic positions.

The $\nabla_i H_{nl}$ derivatives are evaluated along the bonds $r_{R_{ij}}$ that connect the $i$-th Wannier function center in the first unit cell with the $j$-th Wannier function in the unit cell with coordinate $R$. 
\[
\frac{d}{dr} H_{ij}(\bf{R}) = \frac{\Delta H_{ij}(\bf{R})}{\Delta |r_{Rij}|}.
\] (2.128)

Here, \(\Delta |r_{Rij}|\) is the length variation of the bond \(r_{Rij}\) when hydrostatic stress is applied. In practice the derivatives in Eq. (2.128) are calculated as the slope of a linear fit on the \(\Delta H_{ij}(\bf{R})\) vs. \(\Delta |r_{Rij}|\) functions evaluated at \(-10^{-4}, -10^{-5}, -10^{-6}, 10^{-6}, 10^{-5}, 10^{-4}\) relative strain values. The \(\Delta H\) function at these points follow a straight line with less than 0.1 % variations from a perfect linear fit. The linearity of the MLWF Hamiltonian around its equilibrium position validates our approach, as demonstrated in Fig. 2.5 for single- and triple-layer MoS\(_2\). The derivatives along the Cartesian directions are calculated by multiplying \(\frac{d}{dr} H_{ij}(\bf{R})\) with the corresponding Cartesian components of \(\frac{r_{Rij}}{|r_{Rij}|}\).

Alternative methods exist where \(\nabla_i H_{nl}\) is directly computed. Instead of applying hydrostatic strain, the unit cell vectors of an orthorhombic lattice can be independently scaled. In this way the derivatives along the three Cartesian directions are separately obtained from different DFT calculations. This approach is expected to provide a more realistic approximation of the \(\nabla_i H_{nl}\) values, especially in the case of anisotropic materials. In our experience this procedure results in \(\Delta H_{ij}(\bf{R})\) linearly depending on \(|r_{Rij}|\) only when the symmetry group of the strained and the unstrained systems are identical. It is likely due to the splitting of degeneracies that might change the order of the bands. If the energy of the \(n\)-th state in the unstrained system increases above the \(n+1\)-th one due to the strain, the difference of the two Hamiltonian matrices becomes meaningless since the same indices are associated with different states.

The accuracy is expected to improve when, instead of the lattice vectors, the atom positions are varied independently. With this method the influence of the undesirable changes in the periodic images can be attenuated by using sufficiently large supercells. However, it greatly increases the computational cost associated with this task.
2.6 Summary of the workflow

The procedure for the \textit{ab initio} simulation of transistors made of a homogeneous channel material is summarized in Fig. 2.6. It starts with the self-consistent calculation of the Kohn-Sham states in the relaxed, periodic crystal on a regular mesh of \( \mathbf{k} \) points. The Bloch Hamiltonian \( \mathbf{H}(\mathbf{k}) \) is a diagonal matrix of the single-particle eigenenergies. Most of the DFT tools that are interfaced with Wannier90, e.g. VASP or Quantum Espresso can produce the overlap matrices \( M^{(k,b)} \) in Eq. \( (2.113) \) and the initial projections onto trial orbitals \( A(\mathbf{k}) \) in Eq. \( (2.103) \). With these inputs Wannier90 finds the gauge transformation \( U(\mathbf{k}) \) and \( U_{\text{dis}}(\mathbf{k}) \) if the disentanglement method is used that minimizes the spread of the Fourier-transformed real-space wavefunctions. Transforming the Bloch Hamiltonian into this MLWF basis gives the matrix elements \( \langle 0_i | \mathbf{H} | R_j \rangle \) between each pair of Wannier functions. The latter are used for constructing the Hamiltonian \( H_{ij} \) of the finite-size device. This last step is discussed in more details in Appendix A.

The eigenenergies of the system with a small strain are determined by the same DFT tool and with the same input parameters, except for the lattice vectors that are uniformly changed by the same ratio. The difference \( \Delta \mathbf{H}(\mathbf{k}) \) between the strained \( \mathbf{H}_s(\mathbf{k}) \) and unstrained \( \mathbf{H}_0(\mathbf{k}) \) Bloch Hamiltonians are transformed into the MLFW basis with the same matrix as the one used to generate the MLWF Hamiltonian matrix elements. The resulting \( \frac{d}{dr} \mathbf{H}(\mathbf{R}) \) derivatives are projected along the \( x, y, \) and \( z \) coordinates to yield the expression required in Eq. \( (2.64) \).

The calculation of the electron-phonon couplings is independent of the MLWF Hamiltonian and its derivatives. It requires computing the dynamical matrix of the considered system, either by evaluating the set of forces acting on each atom displaced by a small finite value along all symmetrically non-equivalent directions or by using density-functional perturbation theory (DFPT). This method allows for obtaining the second derivatives of the energy directly, within one simulation. The phonon frequencies \( \omega \) and modes \( f_\lambda \) are then calculated on a regular phonon wavevector \( q \) grid and the electron-phonon coupling factors \( \nu(\omega, q_z) \) are determined according to Eqs. \( (2.65, 2.67) \).
The NEGF quantum transport solver takes the real-space Hamiltonian, its derivatives, and the electron-phonon coupling factors as inputs, along with the parameters defining the geometry of the device, the dopings, the gate- and drain biases, and the dielectric constants of the oxide and channel materials. It constructs and updates iteratively the boundary and scattering self-energies and the Green’s functions, and solves Poisson’s equation until the charge-density $\rho(r)$ and the electrostatic potential energy $V(r)$ reach self-consistency. Thereafter, the drain current $I_d$ is computed from the Hamiltonian $H$ and the lesser Green’s function $G^<$ as in Eq. (2.62).

With this method the properties of realistic nano-devices can be predicted and analyzed without the need for any empirical parametrization. It allows one to compare the performance limits of transistors based on novel channel materials where experimental data is hardly available or where it is difficult to separate the effects arising from fabrication imperfections and the intrinsic features of the material. Such an accurate simulation approach can give insight into the underlying microscopic physical processes that govern the operation of nanoscale devices and support the on-going experimental activity. In the subsequent chapters its application to the study of two-dimensional transistors based on van der Waals materials will be presented.
2.6. **SUMMARY OF THE WORKFLOW**

Ab-initio DFT code

```
initial projections, overlaps, Bloch Hamiltonian
```

```
A(k), M^{(k,b)}
```

```
Wannier90
```

```
U(k)
```

```
MLWF Hamiltonian matrix elements
```

```
device Hamiltonian
```

```
<0|i|H|Rj>
```

```
\Delta H(k)
```

```
\Sigma_{R}, \Sigma_{G}, G_{R}, G_{R}, n, V, I_d
```

```
phonon frequencies and modes
```

```
\omega(q), f_\lambda(q)
```

```
electron-phonon coupling
```

```
v(\omega, q_z)
```

```
U(k)
```

```
H_0(k)
```

```
H_s(k)
```

```
\Sigma_{R}, \Sigma_{G}
```

```
dynamical matrix
```

```
strain Bloch Hamiltonian
```

```
U(k)
```

```
Wannier transformation matrix
```

Figure 2.6: Flowchart of the *ab initio* transport simulation procedure. The inputs for the NEGF solver are determined by density-functional theory and density-functional perturbation theory before the quantities are transformed into a MLWF basis.
Chapter 3

Simulation of MoS$_2$ FETs

3.1 Introduction

Molybdenum disulfide (MoS$_2$) has been the first layered semiconductor from which an atomically thin sheet could be successfully turned into a working transistor [14]. So far, it has received the largest attention among the many existing van der Waals materials. We therefore started our study of 2D transistors with MoS$_2$.

Field-effect transistors (FETs) made of single-layer MoS$_2$ – the most widely studied transition metal dichalcogenide (TMD) – have been investigated at different levels of approximation. Studies have been performed in the ballistic limit of transport [94] and in the presence of electron-phonon scattering [95], but in both cases within the effective mass approximation. Full-band calculations have been undertaken too, but without considering scattering [96]. A full-band dissipative approach has been also proposed, but only to evaluate the electron mobility of MoS$_2$ with the linearized Boltzmann Transport Equation [97]. In two recent works ab initio quantum transport simulations of monolayer, single-gate MoS$_2$ FETs have been carried out, taking electron-phonon scattering into account [98] [99]. It has been suggested that the relatively high ON-current and negative differential
resistance observed in previous studies \cite{96} may be artifacts of the applied approximations inherent to each model. In order to obtain physically meaningful results a full-band model in conjunction with electron-phonon scattering should be employed to explore the design space of MoS$_2$ transistors.

Therefore, in this chapter single-, double-, and triple-layer MoS$_2$ FETs with single- and double-gate configurations are investigated with the state-of-the-art simulation approach described in Chapter 2.

### 3.2 Computational details

The required \textit{ab initio} components (Hamiltonian and electron-phonon coupling) are determined with the help of the density-functional theory (DFT) package VASP \cite{100}. It has been demonstrated earlier \cite{101} that the generalized gradient approximation (GGA) of Perdew, Burke, and Ernzerhof (PBE) produces a band gap very close to the experimental one ($E_{g,exp} = 1.8$ eV) in the trigonal-prismatic form of single-layer MoS$_2$. The best match is obtained when the structure of the single-layer is built up using the measured bulk lattice parameters without geometry relaxation. The same approach is employed here with $a = 3.16$ Å as lattice constant and $b = 3.172$ Å as the distance between the sulfur layers within the MoS$_2$ sheet. The multilayer structures are constructed considering the “2H” symmetry (AB-stacking) with an out-of-plane lattice parameter $c = 12.29$ Å \cite{102}. Van der Waals interactions are included through the DFT-D2 method of Grimme \cite{103}. A $25 \times 1 \times 25$ Monkhorst-Pack $k$-point grid and a 500 eV plane-wave cutoff energy are used in the electronic structure calculations. Spin-orbit coupling is neglected. The convergence criteria is set to less than $10^{-3}$ eV total energy difference between two subsequent iterations.

The seven highest valence bands and four lowest conduction bands of single-layer MoS$_2$ are captured with the Wannier transformation by considering three $p$-like Wannier functions centered on each sulfur atoms and five $d$-like orbitals on each molybdenum. The method developed for composite bands is employed, as detailed in Sec. 2.5.2, excluding the lower and higher lying bands that are separated by a
finite gap from the bands of interest. In the case of double- and triple-layer MoS$_2$ the same projection scheme was applied for the atoms in the second and third layers.

The Hamiltonian matrix in the MLWF basis has a tight-binding-like sparsity pattern with the hopping parameters quickly decaying as the distance between the Wannier function centers increases. The device Hamiltonian is constructed from these matrix elements after truncating the terms corresponding to interactions beyond $6^{th}$ nearest-neighbors. Figure 3.1 (a) shows how the band gap, the energy separation of the $K$ and $\Sigma$ valleys, and the effective masses obtained from the MLWF Hamiltonian converge to the DFT values as the number of neighbor interactions included in the model increases, justifying the selected interaction range. The accuracy of the method is also demonstrated in Fig. 3.1 (b)-(d), where the MLWF band structures of single- double- and triple-layer MoS$_2$ are compared to the DFT ones.

In order to construct the scattering self-energies the phonon frequencies $\omega_\lambda(q)$ and eigenvectors $f_\lambda$ have to be determined. Using density-functional perturbation theory (DFPT) VASP can compute the dynamical matrix of MoS$_2$, from which the phonon frequencies and oscillation amplitudes are extracted [104]. This operation requires first to fully relax the ions in the chosen MoS$_2$ unit cell with high accuracy, which is achieved by a convergence criteria of less than $10^{-8}$ eV/Å force acting on each ions. The phonon band structure of the single-layer and double-layer MoS$_2$ extracted from a $3 \times 1 \times 3$ supercell calculation is shown in Fig. 3.2 (a) and (b), respectively. For the generation of the scattering self-energies the phonon eigenmodes and frequencies are conveniently calculated in the rectangular unit cell shown in Fig. 2.2 (b), corresponding to a supercell with 48 atoms per MoS$_2$ layer discretized with a $21 \times 1 \times 21$ phonon wavevector $q$ grid.

Interactions with polar optical phonons are accounted for through the Fröhlich Hamiltonian parametrized in Ref. [105] for MoS$_2$, i.e. the electron-phonon matrix element of Ref. [105] is inserted into Eq. (2.68).

Regardless of the channel length, carrier concentration, or number of layers, the following parameters have been used in all our device simulations. The phonon frequency $\omega_{PH}$ in Eqs. (2.64), (2.65), and (2.66) is sampled by grouping all the phonon energies into 10 distinct
Figure 3.1: (a) Comparison of the band structure features of single-layer MoS$_2$ as obtained with the MLWF Hamiltonian and the DFT one. The plot shows the relative deviations of the band gap ($E_G = 1.79$ eV), energy separation of the $\Sigma$ and $K$ valleys in the conduction band ($E_{K\Sigma} = 0.048$ eV), electron effective mass at the $K$ point ($m^*_{K} = 0.52m_0$) and at the $\Sigma$ point ($m^*_{\Sigma} = 0.62m_0$) along the $K - \Gamma$ line as a function of the number of neighbor interactions included in the MLWF Hamiltonian. (b) Single-layer, (c) double-layer, and (d) triple-layer MoS$_2$ band structures obtained with DFT and with the MLWF Hamiltonian including up to 6$^{th}$ nearest neighbor interactions. The zero energy level is set to the midgap energy.
3.3. PHONON-LIMITED ELECTRON MOBILITIES

Figure 3.2: (a) Single-layer and (b) double-layer MoS$_2$ phonon band structure calculated with VASP.

intervals with a uniform energy width. The derivatives of the MLWF Hamiltonian in Eq. (2.64) were neglected for index pairs corresponding to atoms beyond nearest-neighbor distances. The Green’s functions and the self-energies are sampled on an energy grid with 1 meV step sizes, which yields approximately 1000 points in the energy range considered for transport. The periodic direction $z$ is modeled via 21 $k_z$ points. The convergence criteria of the self-consistent iterations between two subsequent steps is set to a less than $10^{-3}$ relative difference in the electrostatic potential (with respect to the $L^1$ norm) that is computed by solving the Poisson equation with the charge density obtained from the previous step. The simulations are performed at room temperature $T = 300$ K.

Two critical approximations have been made here: (i) the experimental lattice constants have been used to create the systems from which the MLWFs were extracted and (ii) no beyond-nearest-neighbor Hamiltonian derivatives were considered. In Appendix C, these unnecessary simplifications will be alleviated and their influence on the results presented here will be discussed.
3.3 Phonon-limited electron mobilities

As a first step the phonon-limited electron mobility $\mu_{ph}$ of single-, double-, and triple-layer MoS$_2$ is determined with the “dR/dL“ method and Eq. (2.70). Figure 3.3 (a) shows the channel resistance $R(L)$ as a function of the sample size at an electron concentration $n_{2D} = 10^{11}$ cm$^{-2}$. The resistance linearly increases with the sample length in each case, confirming the diffusive character of the transport. The mobility values as a function of the carrier concentration are plotted in Fig. 3.3 (b). At low charge densities the mobility of single-layer MoS$_2$ reaches a maximum of $\mu_{ph} = 220$ cm$^2$/V$s$, which is larger than the experimentally extracted value (15-60 cm$^2$/V$s$) [106]. This overestimation is due to the neglect of charged impurity scattering (CIS) [107]. Including this scattering mechanism is out-of-the-scope of this study where the intrinsic properties of MoS$_2$ shall be investigated, i.e. the properties that would be obtained if the impurity and defect concentration of MoS$_2$ are reduced to the minimum. Our mobility value is, however, smaller than the 410 cm$^2$/V$s$ predicted in Ref [97].
using the linearized Boltzmann transport theory. This discrepancy might come from the mobility extraction technique (Boltzmann vs. dR/dL method), but it most likely originates from the different DFT approaches that were used (geometry relaxation vs. no relaxation), which leads to a smaller energy separations between the two lowest conduction band valleys in our case. This will be confirmed in Appendix C. The mobility increases with the number of layers per MoS$_2$ sample, in agreement with experiments [108] [109]. So far it has been believed that this increase resulted from a reduced influence of CIS and long-range disorder in thicker MoS$_2$ flakes [110] [111] [112]. Although these effects play an important role in MoS$_2$ without a doubt, here we show that the higher mobility is an intrinsic property of multilayer MoS$_2$. Furthermore, in all three MoS$_2$ layer configurations a decrease in the mobility is observed at high carrier concentrations, with a more pronounced drop for the single-layer structure. The overall mobility degradation with increasing charge results from the simultaneous rising of the Fermi level. As a consequence, states situated at higher energies become filled. They can more efficiently relax than low-energy states through optical phonon emission, thus inducing an additional scattering source and a reduction of the current magnitude and mobility. To better understand the difference between the behavior of single-layer and multilayer samples their respective band structure must be carefully analyzed.

The key feature is the energy separation between the $K$ and $\Sigma$ valleys. Its role is explained in Fig. 3.4. By increasing the carrier concentration in single-layer MoS$_2$ the $\Sigma$ valley that is situated 40 meV above the $K$ valley in our calculations becomes more and more populated due to the rising Fermi level. Electrons in the $\Sigma$ valley have a heavier effective mass $m^*_K = 0.52$ $m_0$ than those in the $K$ one $m^*_\Sigma = 0.52$ $m_0$. Hence, filling this satellite valley through thermionic emission or the absorption of a phonon induces a reduction of the mobility. In the double-layer case the $\Sigma$ valley is situated 180 meV below the $K$ valley. Although electrons in the $\Sigma$ valley have a higher effective mass than those in the $K$ valley, the electron-phonon scattering rate diminishes because the $K - \Sigma$ separation is larger than the maximum phonon energy. This leads to a higher mobility. When going from double- to triple-layer the valley separation further increases, while the effective mass in the $\Sigma$ valley decreases. Since
Figure 3.4: Close-up of the (a) single-, (c) double-, and (d) triple-layer MoS$_2$ band structures showing the energy separations between the $K$ and $\Sigma$ valleys. At the $K$ and $\Sigma$ points only the highest transport effective masses of the degenerate valleys are shown. (b) Energy-resolved electron density in single-layer MoS$_2$ at different charge concentrations. The height of the first peak at $E = 0.91$ eV ($K$ valley) is normalized to 1. As the electron population increases, the satellite $\Sigma$ valley situated at $E = 0.95$ eV along the $K - \Gamma$ direction becomes more and more filled.

Both evolutions positively affect the average electron velocity, a larger mobility increase than when going from single- to double-layer is observed.
3.3. PHONON-LIMITED ELECTRON MOBILITIES

The influence of the different electron-phonon scattering mechanisms has been studied too. The mobility values reported above have been obtained by including electron-phonon scattering caused by acoustic and optical deformations as well as by polar optical interactions through Fröhlich’s theory. Calculations without Fröhlich interactions have been performed for single-layer MoS₂. It has been found that polar optical phonon scattering reduces the mobility by less than 10% at all investigated charge densities. This indicates that the main phonon scattering mechanism in MoS₂ is the one caused by deformation potentials.

The temperature dependence of \( \mu_{ph}(T) \) has also been examined. It follows a \( T^{-\gamma} \) law with \( \gamma = 2.2 \), a value smaller than in bulk (\( \gamma = 2.6 \)), but larger than in experiments \[106,113\] and in other theoretical studies \[97,107\] (\( 1.0 \leq \gamma \leq 1.9 \)). Since the latter only considers the \( K \) valley but not the satellite \( \Sigma \), it can be estimated that the difference in the \( \gamma \) values mainly come from our low \( K-\Sigma \) energy separation. By increasing the temperature from 200 to 300 K, the carrier concentration in the satellite valleys goes from 30% up to 45% of the total population, thus drastically reducing the mobility. With a larger \( K \)-satellite splitting this effect would be less significant and \( \gamma \) smaller. We have therefore performed a GW calculation for monolayer MoS₂ because this approach gives more accurate quasiparticle energies. It has been demonstrated earlier that the \( G_1W_0 \) level produces an optical band gap very close to the experimental one \[114\]. We have followed the same procedure with a \( 12 \times 1 \times 12 \) \( k \)-point grid, an energy cutoff \( E_S = 450 \) eV, and 360 bands included in the simulation. The transport Hamiltonian has been created in the same way as in the DFT-PBE case, after transforming the quasiparticle states into a set of maximally localized Wannier functions. The \( \Sigma-K \) valley separation is equal to 0.165 eV in the Wannier-interpolated GW band structure, reducing \( \gamma \), as compared to the previous results. However the obtained mobility values are unrealistically high (> 1000 cm²/Vs), because we neglected Hamiltonian derivatives that go beyond nearest-neighbors, as already discussed above. Note that possible screening of the band gap by the underlying substrate \[115\] is missing in our GW approach, while it is implicitly included in the DFT-PBE scheme through the known band gap underestimation. Therefore we have uses the Hamiltonians obtained by DFT-PBE in the following device simulations.
3.4 Ballistic device simulations

The intrinsic performance (no contact resistance) of the single-layer MoS$_2$ FET depicted in Fig. 3.5 (a) is investigated first within the ballistic limit of transport. The gate length is set to $L_g = 10.7$ nm according to the industry specifications for 2020 by the International Technology Roadmap for Semiconductors (ITRS) 2013 [116]. The source and drain extensions measure $L_s = L_d = 15$ nm each and are doped with a donor concentration $N_D = 6 \times 10^{13}$ cm$^{-2}$. The channel is covered by a 3 nm thick HfO$_2$ layer with $\varepsilon_r = 20$, which corresponds to an equivalent oxide thickness EOT=0.58 nm. The transfer characteristics $I_d - V_{gs}$ of the device are shown in Fig. 3.6 at $V_{ds} = 0.05$ and 0.68 V.

The current exhibits an unusual negative differential resistance (NDR) behavior that can be observed at high gate voltages. The ON-current $I_{on}$ drops from 520 $\mu$A/$\mu$m to 275 $\mu$A/$\mu$m as the source-to-drain voltage $V_{ds}$ increases from 0.05 to 0.68 V. This NDR has been already reported in an earlier theoretical work [96]. Here, we show that it is non-physical and an artifact of the ballistic model; the phenomenon vanishes in the presence of electron-phonon scattering.

Figures 3.7 (a) and (b) report the conduction band structure of single-layer MoS$_2$ along the high symmetry lines in the hexagonal

Figure 3.5: (a) Schematic view of a single-gate monolayer MoS$_2$ field-effect transistor (FET), where the source, drain, and gate regions measure $L_s = 15$ nm, $L_d = 15$ nm, and $L_g = 10.7$ nm, respectively. (b) Schematics of a double-gate FET with triple-layer MoS$_2$ as the channel material.
Brillouin zone of the primitive unit cell and along the $k_x$ line at 
$k_z = 0$ of the rectangular supercell used in the transport simulations,
respectively. Four sub-bands can be distinguished that start from the 
two lowest energy valleys and mostly contribute to electron transport. 
In the ballistic limit a state must be available in the source, channel, 
and drain region so that a reflectionless transmission is possible \[117\]. 
Figure 3.7 (c) shows the conduction band profile in the device at 
$V_{gs} = 0.35$ V and $V_{ds} = 0.68$ V, as well as the channels that the four 
available sub-bands form. At high $V_{ds}$ states in bands 1, 2, and 3 
cannot reach the drain side due to their narrow energy widths. Only 
states in band 4 can propagate from source to drain. However, at 
low $V_{ds}$ all the bands can carry the current, as long as the distance 
between the top of the barrier and the drain CB edge is smaller than 
the band widths. The current therefore increases with decreasing 
$V_{ds}$. This NDR has not been observed in room-temperature experiments 
and has no physical meaning. In reality electron-phonon scattering 
can connect different sub-bands and channels that would otherwise 
be unavailable in the ballistic limit start to become conductive. It 
can be concluded that the ballistic model is not suitable for realistic
Figure 3.7: (a) Conduction band structure of single-layer MoS$_2$ in its hexagonal primitive unit cell and (b) folded band structure in a rectangular unit cell along the transport direction $x$ at $k_z = 0$. Both plots highlight the four lowest sub-bands that contribute to electron transport. (c) Conduction band edge of the simulated device at $V_{gs} = 0.35$ V and $V_{ds} = 0.68$ V, including the channels that the bands in subplots (a-b) form. The arrow shows the maximum position of the electrostatic potential energy in the channel, known as Top-of-the-Barrier (ToB). A reflectionless transmission is only possible if a given channel is available in the source, channel, and drain region [117]. Due to their narrow energy width electrons in bands 1, 2, and 3 cannot propagate from source to drain at high $V_{ds}$, contrary to band 4. At low $V_{ds}$ states belonging to any of the four sub-bands can propagate, explaining the negative differential resistance observed in Fig. 3.6.

simulations of MoS$_2$ FETs. This finding hold for most few-layer TMDs and not only for MoS$_2$. A similar behavior has benn observed in MoTe$_2$, MoSe$_2$, WSe$_2$, and WS$_2$.

3.5 Dissipative device simulations

The monolayer MoS$_2$ FET is now simulated in the presence of electron-phonon scattering. The inclusion of this dissipative mechanism has
3.5. DISSIPATIVE DEVICE SIMULATIONS

Figure 3.8: Energy- and position-resolved current $I_d(E, x)$ of a single-gate monolayer MoS$_2$ FET at $V_{ds} = 0.68$ V and $V_{gs} = 0$ V in the case of (a) ballistic transport and (b) when electron-phonon scattering is turned on. Red indicates high current concentrations, green no current, the dashed line is the conduction band edge.

two opposite consequences: on one hand electrons flowing from source to drain may be scattered back to their origin, which reduces the current. On the other hand, electrons situated in the narrow bands 1, 2, and 3 in Fig. 3.7 can propagate from one side of the transistor to the other because phonon absorption and emission connects them, as illustrated in Fig. 3.8. In sub-plot (a) the spectral current from the ballistic simulation is reported. It remains constant along the device length, since the energy of the electrons injected into the simulation domain cannot vary in the absence of a dissipative scattering mechanism. The consideration of electron-phonon interactions profoundly modifies the current distribution, creating a high flow density on the drain side of the transistor, below the CB edge of the source, as shown in Fig. 3.8 (b). The emission/absorption of phonons allows states that would normally be reflected back to the source to continue their ways towards the drain.
Figure 3.9: Transfer characteristics $I_d - V_{gs}$ at $V_{ds} = 0.05$ (solid red lines) and 0.68 V (dashed blue lines) with electron-phonon scattering for the single-gate monolayer MoS$_2$ FET.

The resulting transfer characteristics with dissipative scattering are displayed in Fig. 3.9. At low $V_{ds}$ the current is reduced, as compared to the ballistic one, due to backscattering [118]. This is not the case at high $V_{ds}$, where the gain from connecting bands with a narrow width on the source and drain sides overcomes the decrease caused by backscattering. The NDR behavior disappears and the current follows the expected increase with respect to $V_{ds}$.

FETs made of double- and triple-layer MoS$_2$ have also been simulated using the same approach as in the single-layer case, but with a single- and a double-gate contact. The source-, drain-, and gate lengths and the oxide thicknesses remain the same as before. The OFF-state currents $I_{off}$ are set to 0.1 $\mu$A/$\mu$m by adjusting the gate work function. The transfer characteristics $I_d - V_{gs}$ are reported in Fig. 3.10 for the single-gate transistors at $V_{ds} = 0.68$ V. The subthreshold slope deteriorates with the increasing number of layers due to the weakening of the gate control over the channel: it goes from 78 mV/dec in the single-layer FET up to 88 mV/dec for the
Figure 3.10: Transfer characteristics of single- (solid blue lines), double- (dashed red lines), and triple-layer (dashed-dotted green lines) MoS$_2$ FETs at $V_{ds} = 0.68$ V with a single-gate geometry.

triple-layer structure. This comes from the significant screening of the electrostatics in MoS$_2$ caused by the high density-of-states of this material. Despite the poorer subthreshold slope the ON-current $I_{on}$ of the double-layer MoS$_2$ FET is higher than its single-layer counterpart as a result of the increased electron mobility. Although the mobility in triple-layer MoS$_2$ is even higher, this device performs worse than the double-layer one because the mobility gain does not compensate for the drastic deterioration of the electrostatic control occurring in thicker transistors with a single-gate contact of length $L_g = 10.7$ nm.

The electrostatic control characterizes how efficiently the potential in the channel is modulated by variations of the potential at the gate contact. Figure 3.11 (a) shows the evolution of the maximum of the electrostatic potential energy in the channel, known as Top-of-the-Barrier (ToB), as a function of the gate voltage in the single-, double-, and triple-layer MoS$_2$ FETs with a single-gate. In case of a perfect gate control the change in the electrostatic potential is identical to the change in the gate voltage in the subthreshold region. However,
Figure 3.11: Maximum of the electrostatic potential energy in the channel (ToB) as a function of the gate voltage in the (a) single-gate and (b) double-gate FETs at $V_{ds} = 0.68$ V.

even the single-layer MoS$_2$ FET only operates at around 75% of its ideal limit. This number is 65% for the triple-layer MoS$_2$ FET, giving rise to a 9 meV energy difference when comparing their ToB position in the ON-state.

A higher potential barrier, as in the triple-layer configuration, results in a lower charge accumulation in the channel, the ON-state charge density at the ToB, $n_{ToB}$ decreasing from $1.5 \times 10^{13}$ 1/cm$^2$ (single-layer) to $1.4 \times 10^{13}$ 1/cm$^2$ (triple-layer), as shown in Table 3.1. The electron injection velocity $v_{inj}$ follows an opposite trend and increases with the layer thickness ($v_{inj} = 2.6 \times 10^6$ cm/s in single-layer and $3.1 \times 10^6$ cm/s in triple-layer) due to the reduced inter-valley scattering in multilayers. This leads to a higher drain current. The calculated velocities at the ToB are significantly lower than the band structure-limited injection velocities $v_{max}$ that are $6.9 \times 10^6$ cm/s in single-layer MoS$_2$, $6.6 \times 10^6$ cm/s in double-layer, and $7.0 \times 10^6$ cm/s in triple-layer. The reduction in $v_{max}$ when going from single- to double-layer is attributed to the $\Sigma$ valley, which has a heavier effective mass than the $K$ one and represents the minimum of the conduction
3.5. DISSIPATIVE DEVICE SIMULATIONS

<table>
<thead>
<tr>
<th></th>
<th>Single-layer</th>
<th>Double-layer</th>
<th>Triple-layer</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n_{ToB} )  ( (1/cm^2) )</td>
<td>Single-gate</td>
<td>1.5 ( \times 10^{13} )</td>
<td>1.5 ( \times 10^{13} )</td>
</tr>
<tr>
<td></td>
<td>Double-gate</td>
<td>2.5 ( \times 10^{13} )</td>
<td>2.5 ( \times 10^{13} )</td>
</tr>
<tr>
<td>( v_{inj} ) ( (\text{cm/s}) )</td>
<td>Single-gate</td>
<td>2.6 ( \times 10^{6} )</td>
<td>3.1 ( \times 10^{6} )</td>
</tr>
<tr>
<td></td>
<td>Double-gate</td>
<td>3.1 ( \times 10^{6} )</td>
<td>4.4 ( \times 10^{6} )</td>
</tr>
</tbody>
</table>

Table 3.1: 2-D charge density \( (n_{ToB}) \) and electron injection velocity \( (v_{inj}) \) in the ON-state of the investigated MoS\(_2\) devices at the top of the potential barrier.

band in double-layer MoS\(_2\). The maximum injection velocity increases between a double- and triple-layer structure thanks to the reduction of the effective mass at \( \Sigma \).

Note that in ultra-scaled devices the injection velocity and the charge at the top of the potential barrier are not as well defined as in long-channel devices due to the lack of a plateau in the middle of the channel. The potential barrier is very narrow at high gate biases and it is located very close to the source. Extracting the above quantities just 0.3 nm (approximately a unit cell width) away from the maximum position of the potential results in a 30-50% deviation in their values. However, their behavior does not qualitatively change when comparing the different devices, no matter what interpolation technique is used to determine the charge and potential between the available discrete atomic positions.

To complete our analysis we have finally investigated the performance of double-gate MoS\(_2\) transistors, as depicted in Fig. 3.5 (b). In this geometry the bottom gate, as well as the top gate, is separated from the channel by a 3 nm thick HfO\(_2\) layer. The current vs. voltage characteristics of the single-, double-, and triple-layer devices are compared in Fig. 3.12. The subthreshold slope of the three switches
CHAPTER 3. SIMULATION OF MOS₂ FETS

Figure 3.12: Transfer characteristics of single- (solid blue lines), double- (dashed red lines), and triple-layer (dashed-dotted green lines) MoS₂ FETs at $V_{ds} = 0.68$ V with a double-gate geometry.

become very close to each other. It only varies from 67 mV/dec (single-layer) to 69 mV/dec (triple-layer). The improved electrostatic control has a significant impact on the performance of even the single-layer MoS₂ FET. The ON-state current (extracted at $V_{gs} = 0.68$ V) doubles from 600 to 1200 $\mu$A/$\mu$m. The $I_{on}$ of the double-layer transistor reaches 1670 $\mu$A/$\mu$m, while the boost in the triple-layer case is even more spectacular with $I_{on} = 1850$ $\mu$A/$\mu$m, thus outperforming both the single- and double-layer devices.

A superior electrostatic control directly manifests its presence when comparing the behavior of the Top-of-the-Barrier potentials in the single and double-gate devices in Fig. 3.11 (a) and (b). In the double-gate FETs approximately 85% of the gate voltage variations are reported to the channel potential in the subthreshold regime, regardless of the number of MoS₂ layers. As a consequence higher charge concentrations ($n_{ToB} = 2.5 \times 10^{13}$ 1/cm² in the single-layer and $2.4 \times 10^{13}$ 1/cm² in the triple-layer MoS₂ FET) and injection velocities ($v_{in,j} =$
3.6. **CONCLUSIONS**

3.1 \times 10^6 \text{ cm/s} in single-layer and 5.0 \times 10^6 \text{ cm/s} in triple-layer) are found at the Top-of-the-Barrier location.

The single-layer device simulations have been repeated using the MLWF Hamiltonian obtained from GW calculations. Although the current is probably overestimated by this method, as explained when discussing the temperature dependence of the mobility, one important point is verified: a negative differential resistance occurs in the ballistic limit of transport, which disappears when electron-phonon scattering is turned on. Hence this phenomenon does not depend on the band structure model.

### 3.6 Conclusions

We have investigated the performance of single- and few-layer MoS$_2$ transistors by means of 3D, atomistic, dissipative quantum transport simulations. Both the electronic Hamiltonian and the electron-phonon self-energies have been constructed based on physical quantities derived from *ab initio* density-functional theory calculations. The phonon-limited electron mobility has been found to increase with the number of layers in the MoS$_2$ sample and to drop at high carrier concentrations. We have shown that the ballistic limit is not suited for realistic simulations of MoS$_2$ FETs because of the artificial negative differential resistance it predicts. It has been demonstrated that at a gate length $L_g = 10.7$ nm the double-layer MoS$_2$ FET performs the best when only a single-gate contact is present. With an improved electrostatic control obtained through a double-gate configuration the performance of each device significantly improves and the triple-layer MoS$_2$ outperforms both the single- and double-layer transistors. The proposed simulation approach can be extended to account for charged impurity scattering. It is also ready to be applied to other 2D semiconductors such as MoTe$_2$, WS$_2$, or SnS$_2$. Most findings discussed here were published in Physical Review B [99].
Chapter 4

Black phosphorus: an emerging transistor alternative

4.1 Introduction

Few-layer black phosphorus (phosphorene) has recently stood out from the 2-D crystal family as a promising field-effect transistor (FET) candidate at the end of the semiconductor roadmap. Besides the excellent electrostatic control inherent to the 2-D technology, black phosphorus exhibits a direct band gap and significantly higher mobility values than most transition metal dichalcogenides (TMDs), making it a strong competitor in the race for ultimate logic devices.

The demonstration of a high drain current modulation and high field-effect mobility has aroused the interest of the scientific community for black phosphorus (BP). Despite already spectacular achievements it remains to determine whether phosphorene is capable of outperforming ultra-scaled Si or TMD-based switches. To answer this question computer simulations can assist the on-going experimental efforts and shed light on the characteristics of not-yet-fabricated devices in a time- and cost-efficient way. So far phosphorene FETs have been
Figure 4.1: (a) Perspective side view of multilayer black phosphorus. (b) Top view of single-layer phosphorene. The arrows show the X (armchair) and Y (zig-zag) directions. The dashed rectangle marks the primitive unit cell with 4 atoms. (c) Side view of phosphorene along the armchair edge.

simulated either within a simple Top-of-the-Barrier approach [119] or using a 4-band tight-binding model [120], in both cases in the ballistic limit of transport. The influence of electron-phonon scattering has also been investigated, but only with respect to the phonon-limited carrier mobility calculated with acoustic deformation potentials [43].

This chapter goes one step further by analyzing the performance of single-layer phosphorene FETs by means of ab initio quantum transport simulations including acoustic and optical electron-phonon scattering.

4.2 Simulation setup

The same ab initio scheme as in Chapter 3 is used to simulate black phosphorus samples. The Bloch-electron states of single-layer crystals are computed with the VASP density-functional theory (DFT) tool [100]. The GGA functional of Perdew-Burke-Ernzerhof is utilized because it implicitly accounts for the screening by the substrate through its known band gap underestimation. The phosphorene layer is separated from its periodic replica by a 20 Ångström vacuum region. The ions and the lattice vectors are fully relaxed. In 2-D systems volume relaxation would require selectively changing the in-plane lattice
4.2. SIMULATION SETUP

Figure 4.2: (a) Electronic band structure of single-layer phosphorene obtained by VASP (blue lines) and by the MLWF Hamiltonian utilized in the transport simulations (red circles). The inset shows the path through the high symmetry points in the Brillouin zone. (b) Phonon band structure of single-layer BP calculated by VASP using DFPT.

vectors, while keeping the out-of-plane one fixed. Since such a feature is not implemented in VASP, the ionic relaxations are performed in fixed unit cells with various lattice vectors. The equilibrium lattice constants are then determined as the minimum position of a quadratic fit to the total energies. A force of less than $10^{-3}$ eV/Å acting on each ion and a total energy difference smaller than $10^{-5}$ eV between two subsequent electronic steps are chosen as convergence criteria. A $15 \times 15 \times 1$ Monkhorst-Pack $k$-point grid and a 500 eV plane wave cutoff energy is applied in the geometric optimization process. Spin-orbit coupling is neglected.

The lattice constants are found to be $a_x = 4.38$ Å and $a_y = 3.31$ Å, with the subscript $x$ and $y$ referring to the armchair and the zig-zag directions, respectively. The atomic geometry of black phosphorus is illustrated in Fig. 4.1. After determining the relaxed geometry the electron states are computed on a $21 \times 21 \times 1$ $k$-point grid using a 550 eV cutoff, yielding a direct gap of 0.91 eV at $\Gamma$.

As initial trial orbitals for the Wannier transformation four sp$^3$ hybrids are placed on each of the four P atoms forming the unit cell. It gives rise to 10 valence and 6 conduction bands. Electron states within a 5 eV interval centered around the mid-gap energy are frozen
when disentangling the subspace of the 16 Wannier functions from the higher energy bands. The quadratic spread of each of the resulting MLWFs is less than $2 \, \text{Å}^2$ so that the resulting Hamiltonian matrix displays a tight-binding-like sparsity pattern. Hopping terms are kept within a 18.5 Å interaction range, producing an excellent agreement with the DFT data, as demonstrated in Fig. 4.2 (a). This range corresponds to a 96-atomic transport supercell of $4 \times 6 \times 1$ primitive unit cells interacting only with its nearest-neighbors.

The dynamical matrix of phosphorene is also determined with VASP in the transport unit cell using density-functional perturbation theory (DFPT). The phonon modes and frequencies are extracted with the help of the Phonopy package [104] onto a $21 \times 21 \times 1$ $q$ point grid. They are also computed in the primitive unit cell along the high symmetry lines and reported in Fig. 4.2 (b). The electron-phonon coupling constants are obtained by applying hydrostatic strain to the phosphorene layer in VASP and transforming the difference of the perturbed and unperturbed Hamiltonian into the same MLWF basis. Unlike MoS$_2$, black phosphorus is not a polar material, hence there is no need to include Fröhlich interactions in the calculations.

The transport simulation parameters that are not discussed here (e.g.: transverse $k$ points, energy grid, temperature, convergence criteria . . . ) are identical to those described in Chapter 3.

### 4.3 Investigation of electron- and hole mobilities

First, the phonon-limited electron and hole mobility of single-layer black phosphorus is studied. The linear dependence of the resistance on the channel length, as illustrated in Fig. 4.3, clearly demonstrates the diffusive nature of transport. The $dR/dL$ method [78] is again employed to compute the mobility of both carrier types, which exhibits a strong anisotropy. Figure 4.4 (a) shows that the hole mobility value along $X$ is about 18 times larger than along $Y$, reaching 3000 cm$^2$/Vs. In Ref. [43] where a single-parameter deformation potential model was used to evaluate the phonon-limited carrier mobilities, a hole mobility 16-38 times larger in the $Y$ direction than in the $X$ one was
4.3. INVESTIGATION OF CARRIER MOBILITIES

Figure 4.3: Channel resistance of \( n \)- and \( p \)-type single-layer black phosphorus samples as a function of their lengths. The labels X/Y in the legends refer to the transport direction (armchair/zig-zag), and e/h to the carrier type (electron/hole). The charge concentration is equal to \( 2 \times 10^{13} \) \( \text{1/cm}^2 \) in all cases. As expected with diffusive transport, the resistance linearly increases with the sample size.

predicted. This finding was in contradiction with few-layer samples reported in the same work, where significantly higher hole mobilities were observed in the \( X \) direction than in the \( Y \) one. We attribute this difference to the limits of the Takagi model [121] employed in Ref. [43], that is based on a single-band picture. Our assumption is that this technique fails in the case of an extremely small deformation potential.

Our multi-band simulation results are consistent with the valence band structure features depicted in Fig. 4.4 (b). The effective mass is unusually large in the \( Y \) direction (\( m^*_y \approx 7 m_0 \)) compared to the light holes in the \( X \) direction (\( m^*_x = 0.13 m_0 \)). This anisotropy is the source of the lower hole mobility along \( Y \). Note that the dispersion between the \( \Gamma \) and \( Y \) points shows very heavy non-parabolicity with an almost flat band edge, preventing a more accurate extraction of the effective mass. This non-parabolicity also manifests itself in the mobility increase following the charge density. In most cases a higher carrier concentration reduces the mobility because states away from
the band edge get filled by the shifting Fermi level. These states can then relax through optical phonon emissions, an additional scattering source that reduces the overall mobility. However, due to the heavy non-parabolicity of the phosphorene band in the zig-zag direction, a lowering Fermi level results in an increased average hole velocity that overcomes the effect of the additional scattering on the mobility.

The electron mobilities display an anisotropy similar to the hole ones. As Fig. 4.5 (a) shows, at low carrier concentrations the maximum along the X direction is at around 2500 cm$^2$/Vs, which is about 4 times larger than along Y. At higher concentrations these values reduce regardless of the transport direction, as expected, due to the absence of significant non-parabolicity in the conduction band structure around the band edge. The less important anisotropy of the electron mobility compared to holes is consistent with the smaller difference in the direction-dependent electron effective masses, as indicated in Fig. 4.5 (b). We have $m_y^* = 1.2\ m_0$ vs. $m_x^* = 0.15\ m_0$. Although this behavior of the electron mobilities qualitatively agrees with the findings of Ref. [43], they report a 14-fold improvement in favor of X, three times more than in our calculations. Also, note that
4.4 Analysis of phosphorene FETs

As a next step, the single-layer BP FETs depicted in Figure 4.6 (a) are simulated. The gate length is set to $L_g=10.5$ nm, while the source and drain regions extend over $L_s=L_d=15$ nm and are doped with a donor (acceptor) concentration $N_D=4\times 10^{13}$ cm$^{-2}$ ($N_A=4\times 10^{13}$ cm$^{-2}$) in the $n$-type ($p$-type) configuration. The equivalent oxide thickness in all transistors is equal to EOT=$0.58$ nm (3 nm of HfO$_2$ with $\epsilon_R=20$). The phosphorene channels are deposited on a 20 nm SiO$_2$ substrate. The gate workfunction is adjusted so that the OFF-state current $I_{off}$ =

Figure 4.5: (a) Phonon-limited electron mobility of monolayer black phosphorus as a function of the electron concentration ($1\times 10^{11} \leq n_{2D} \leq 2\times 10^{13}$ cm$^{-2}$) and transport direction ($X$: blue line with circles, $Y$: red line with squares). (b) Conduction band structure of phosphorene around the $\Gamma$ point in the $Y$ and $X$ directions (blue lines) with fitted parabolas (green dashed lines). The highest reported electron and hole mobilities both in Ref. [43] and in this work exceed the experimentally observed ones [31] [30]. This discrepancy arises from the omission of charged impurity scattering in the applied theoretical models and from the neglection of electron-phonon coupling terms beyond first nearest neighbor distances. This latter simplification will be eliminated and its implications discussed in Appendix C.

4.4 Analysis of phosphorene FETs

As a next step, the single-layer BP FETs depicted in Figure 4.6 (a) are simulated. The gate length is set to $L_g=10.5$ nm, while the source and drain regions extend over $L_s=L_d=15$ nm and are doped with a donor (acceptor) concentration $N_D=4\times 10^{13}$ cm$^{-2}$ ($N_A=4\times 10^{13}$ cm$^{-2}$) in the $n$-type ($p$-type) configuration. The equivalent oxide thickness in all transistors is equal to EOT=$0.58$ nm (3 nm of HfO$_2$ with $\epsilon_R=20$). The phosphorene channels are deposited on a 20 nm SiO$_2$ substrate. The gate workfunction is adjusted so that the OFF-state current $I_{off}$ =
Figure 4.6: (a) Schematic view of the single-layer, single-gate $n$- and $p$-type black phosphorus field-effect transistors considered in this work. (b) Ballistic spectral current (current as a function of the position and energy) flowing through an $n$-type black phosphorus FET with transport along $Y$. Red indicates high current concentrations, green no current. The blue line refers to the conduction band edge. (c) Same as in (b), but in the presence of electron-phonon scattering.

0.1 $\mu$A/$\mu$m at $V_{gs} = 0$ V and $V_{ds} = 0.67$ V. These are the same conditions as for the MoS$_2$ FETs in Chapter 3.

The analysis starts by investigating the effect of electron-phonon interactions on a BP FET. The device selected for this task is an $n$-type one with transport along the $Y$ direction. The spectral currents and the conduction band profiles with electron-phonon scattering turned off and on are compared at the same bias in Fig. 4.6 (b) and (c), respectively. It appears that interactions with phonons have little influence on the spectral distribution of the current and on the profile of the conduction band edge. With electron-phonon scattering the current reaches 90% of its ballistic limit. These findings indicate that the ballistic model captures reasonably well the physics of BP FETs, unlike in the MoS$_2$ case. Nevertheless, electron-phonon interactions are included in all the following simulations for the sake of accuracy.

The transfer characteristics $I_d - V_{gs}$ of $p$- and $n$-type devices at $V_{ds} = 0.67$ V are reported in Figs. 4.7 (a-b). The currents remain at around 90% of their ballistic limit and significantly depend on the channel orientation, reflecting the anisotropy of the mobilities and effective masses. The ON-current is about three times larger in the
4.4. ANALYSIS OF PHOSPHORENE FETS

Figure 4.7: (a) Transfer characteristics $I_d$-$V_{gs}$ at $V_{ds} = 0.67$ V of a $p$-type single-layer, single-gate black phosphorus FET as in Fig. 4.6 (a). Transport along the $X$ (blue lines) and $Y$ (red lines) crystal orientations is considered. A linear (dashed lines) and logarithmic (solid lines) scale of the current are provided. (b) Same as (a), but for the $n$-type component.

$X$-oriented $p$-FET ($I_{ON,x} = 3000 \, \mu A/\mu m$) than in the $Y$-oriented one ($I_{ON,y} = 1000 \, \mu A/\mu m$). It is more than twice higher in the $n$-FET with a channel direction along $X$ ($I_{ON,x} = 2600 \, \mu A/\mu m$) than in a transistor with transport along $Y$ ($I_{ON,y} = 1200 \, \mu A/\mu m$). The anisotropy in the ON-currents follows the same trend as in the mobility, although the correspondence is nowhere linear. The hole mobility displays a 14-fold increase in favor of the armchair direction, while the ON-current improvement of the $p$-type device is limited to a factor of 3. The smaller (4-fold) difference in the electron mobility results in a smaller (about 2-fold) disparity in the $n$-FET currents. The same is true for the magnitude of the ON-current when comparing the $n$-type FETs with the $p$-type ones with the same transport direction. In the case of the $X$-oriented devices the ON-current of the $n$-FET is smaller than the $p$-FET one, just as the electron mobility in the $X$ direction is smaller than the hole mobility. And with $Y$ as the transport direction it is larger for the $n$-FET than for the $p$-FET, in agreement with the electron mobility being larger than the hole one along $Y$. The most important device metrics that were calculated are also summarized in Table 4.1.
<table>
<thead>
<tr>
<th></th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( I_{ON} (\mu A/\mu m) )</td>
<td>( I_{ON} (\mu A/\mu m) )</td>
</tr>
<tr>
<td></td>
<td>2624</td>
<td>1220</td>
</tr>
<tr>
<td></td>
<td>( I_{OFF} (\mu A/\mu m) )</td>
<td>( I_{OFF} (\mu A/\mu m) )</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>( SS (mV/\text{dec}) )</td>
<td>( SS (mV/\text{dec}) )</td>
</tr>
<tr>
<td></td>
<td>87</td>
<td>77.5</td>
</tr>
<tr>
<td></td>
<td>Ballisticity (%)</td>
<td>Ballisticity (%)</td>
</tr>
<tr>
<td></td>
<td>93</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td>( m^*_{\text{trans}} (m_0) )</td>
<td>( m^*_{\text{trans}} (m_0) )</td>
</tr>
<tr>
<td></td>
<td>0.14</td>
<td>1.26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.13</td>
</tr>
</tbody>
</table>

Table 4.1: Summary of the ON-current (\( I_{ON} \)), OFF-current (\( I_{OFF} \)), sub-threshold slope (\( SS \)), ballisticity of the ON-current, and transport effective mass (\( m^*_{\text{trans}} \)) for the 4 single-layer, single-gate black-phosphorus FETs simulated here.

Still, the spectacularly high ON-currents in the devices with the armchair configuration (\( X \)) as the transport direction are not fully explained by the high mobility values and require further investigation. III-V devices, for example, also exhibit very high electron mobilities, but the benefit of the observed high carrier velocity is partly annihilated by the reduced carrier population due to the low density-of-states (DOS) [122]. Since the current \( I = q \cdot n_{TOB} \cdot v_{inj} \) is proportional to the charge density in the channel \( n_{TOB} \) times the injection velocity \( v_{inj} \), both quantities need to be simultaneously increased. To shed light on this issue, the carrier concentration in the channel is extracted and plotted versus the gate voltage in Fig. 4.8 for each investigated device. The slope of these curves above the threshold voltage gives the gate capacitances \( C_g = \frac{dn_{TOB}}{dV_{gs}} \), what is usually modeled as two capacitances in series, the oxide- \( C_{ox} \) and quantum- \( C_q \) one

\[
C_g = \frac{C_q C_{ox}}{C_q + C_{ox}}. \tag{4.1}
\]
Figure 4.8: Carrier concentration $n_{ToB}$ at the top of the potential barrier (ToB) separating the source and drain extensions of the 4 different single-layer, single-gate black phosphorus FETs investigated in this work ($n$- and $p$-type with transport along the $X$ or $Y$ crystal axis). The charge is reported with respect to the applied $V_{gs}$ at $V_{ds}=0.67$ V.

The fact that $C_g$ is insensitive to the transport orientation and to the $p/n$-type operation indicates that the transistors operate in the classical capacitance limit, where it can be approximated as the oxide capacitance $C_g \approx C_{ox}$. The influence of the larger $C_q$ can be neglected. If the single-layer BP FETs would operate in the quantum capacitance limit, the device with the lowest density-of-states would have a smaller $C_q$ and therefore $C_g$ than the others, which is not the case here. The source of the large $C_q$ is the anisotropy of the band structure. The DOS does not depend on the transport orientation, since it is given by the number of the states per unit energy in the whole Brillouin-zone. A high transverse effective mass hence increases the DOS without having any adverse influence on the injection velocity that instead
Figure 4.9: (a) Hole injection velocity extracted from the same p-type FETs as in Fig 4.8 as a function of the hole concentration at the ToB and channel orientation (X: blue line, Y: red line). (b) Same as (a), but for the electron injection velocity.

depends on the transport effective mass. This is a clear advantage over isotropic III-V materials with small effective masses that suffer from the so-called DOS-bottleneck effect.

As Fig. 4.9 demonstrates, both the hole- and electron injection velocities keep increasing with the increased charge concentrations, regardless of the transport direction. As a consequence the current is not limited by the carrier velocity either.

The scalability of the phosphorene FETs has also been investigated by calculating the sub-threshold slope (SS) of single-gate devices with various gate lengths ($L_g=5, 7.5, 10.5, 12.5$ and $15$ nm). The $SS$ is extracted from the simulation of two low $V_{gs}$ bias points at $V_{ds}=0.67$ V for each of the 4 device types ($p$- and $n$-FET with transport along $X$ and $Y$). The results are shown in Fig. 4.10. It can be seen that the small effective masses severely limit the gate length scalability of black phosphorus transistors due to the increased source-to-drain tunneling leakage. The sub-threshold slope of both $n$- and $p$-FETs deteriorates much faster if the channel is oriented along $X$. It reaches values above $200$ mV/dec at $5$ nm gate length, making such devices impractical. However, a double-gate design, as investigated in Ref. [120] might help
4.4. ANALYSIS OF PHOSPHORENE FETS

Figure 4.10: (a) Sub-threshold slope (\(SS\)) of the \(p\)-type single-layer, single-gate black phosphorus FET as a function of its gate length (\(L_g=5, 7.5, 10.5, 12.5\) and 15 nm) and channel orientation (\(X\): blue line, \(Y\): red line). (b) Same as in (a), but for the \(n\)-type FET.

The \(X\)-oriented devices maintain their advantage, even below 10 nm channel lengths.

TMD-based FETs with the same gate length \(L_g=10.5\) nm and equivalent oxide thickness EOT=0.58 nm as the BP transistors investigated in this chapter and depicted in Fig. 4.6 (a) have been simulated and compared to them. The following channel materials have been selected: MoS\(_2\), MoSe\(_2\), MoTe\(_2\), WS\(_2\), and WSe\(_2\). The OFF-current is set to \(I_{\text{off}} = 0.1\) \(\mu\)A/\(\mu\)m in each case. The transfer characteristics \(I_d - V_{gs}\) of the \(p\)- and \(n\)-type devices at \(V_{ds} = 0.67\) V are reported in Fig. 4.11 (a) and (b), respectively. The results indicate that MoS\(_2\) is not the best choice as FET material. Among the investigated devices the highest ON-current is delivered by WSe\(_2\) in the \(p\)-type configuration \((I_{on} = 1100\) \(\mu\)A/\(\mu\)m\) and by the WS\(_2\) when considering \(n\)-FETs \((I_{on} = 1500\) \(\mu\)A/\(\mu\)m\).

To complete this study, the best-performing TMD FETs have been compared to phosphorene. In Fig. 4.12 (a) the transfer characteristics of the \(p\)-type WSe\(_2\) transistor are shown along with the \(p\)-type, \(X\)- and \(Y\)-oriented phosphorene. The advantage of the BP FET with the armchair (\(X\)) direction of transport is obvious, the ON-currents being 3000 vs. 1100 \(\mu\)A/\(\mu\)m for WSe\(_2\). When the channel is aligned with
Figure 4.11: Transfer characteristics $I_d - V_{gs}$ at $V_{ds} = 0.67$ V of (a) $p$-type and (b) $n$-type TMD FETs with the same device parameters as the BP transistors investigated in this chapter on a logarithmic and linear scale. The material that provides the highest ON-current is marked with a red rectangle on the legends of both sub-plots.

the zig-zag ($Y$) direction, both materials perform almost identically, with WSe$_2$ just slightly surpassing BP. The superiority of the $X$-oriented BP FET over TMDs is also clear among the $n$-type devices, as demonstrated in Fig. 4.12 (b). The ON-current $I_{on} = 2500$ $\mu$A/$\mu$m far exceeds the 1500 $\mu$A/$\mu$m delivered by the WS$_2$ transistor. If the direction of transport is switched in BP, WS$_2$ becomes more favorable.

Hence, with a gate length of $L_g = 10.5$ nm phosphorene is more appealing for high-performance logic applications than transition-metal-dichalcogenides. However, at shorter channel lengths this advantage is no more clear since transport along the armchair direction induces high leakage currents in practical applications and the performance of zig-zag-oriented phosphorene FETs is comparable to the TMDs. With a double-gate design the superiority of the $X$-oriented phosphorene FET might be preserved for devices with gate lengths between 5-10 nm, but other materials with better tradeoff between mobility and effective mass might eventually be required. The mobility becomes less important in ultra-scaled devices, contrary to the transverse effective mass that can reduce leakage currents.
4.5 Conclusions

Black phosphorus FETS have been simulated at the \textit{ab initio} level in the presence of electron-phonon scattering. The phonon-limited electron and hole mobilities have been found to exhibit a high anisotropy with the best performance in favor of the armchair direction. This anisotropy is also observed in the transfer characteristics of \textit{p}- and \textit{n}-type transistors with different channel orientations. The armchair direction is favorable when the gate length is longer than 10 nm, but a zig-zag-oriented channel better withstands the short-scale effects arising in ultra-small devices. We have also shown that BP FETs do not suffer from the density-of-states bottleneck effect observed in III-V devices, despite a low transport effective mass. The extracted mobility and current values indicate that this material holds great promise down to 10 nm gate lengths, but might lose its advantages over transition-metal-dichalcogenides when further scaled. Most findings discussed in this Section were presented at the 2015 International Electron Devices Meeting and published in its proceedings [123].
Chapter 5

Van der Waals heterojunction tunneling FETs

5.1 Introduction

Band-to-band tunneling field-effect-transistors (TFETs) are promising candidates as building blocks for low power electronic circuits because their inverse subthreshold slope (iSS) can be reduced below 60 mV/dec at room temperature due to the injection of cold instead of hot electrons \[44\]. TFETs with a steep iSS and a low OFF-state current have already been fabricated \[47\], but they do not simultaneously deliver an ON-current \(I_{on}\) high enough for practical applications. Increasing \(I_{on}\) without deteriorating the subthreshold region remains a challenge in TFETs that van der Waals heterostructures (vdWHs) might solve.

With the emergence of atomically thin vdWHs \[15\] \[124\] it is nowadays possible to fabricate many type-II heterojunctions with high quality interfaces due to the absence of covalent bonds at the junctions. In atomically-thin vertical heterojunctions the tunneling length is reduced to its minimum, offering the possibility to reach higher currents than in monolayer metal dichalcogenide (MD) TFETs \[125\] \[126\]
and to achieve symmetric \( n \)- and \( p \)-type operations. Furthermore, the larger effective masses of individual MD layers, as compared to III-V, suppress direct source-to-drain tunneling, but do not affect band-to-band tunneling due to the possible formation of a staggered band alignment when they are put together.

TFETs with BN \([127]\) or WS\( _2 \) \([128]\) embedded between graphene layers have already been demonstrated. In these experiments, however, BN and WS\( _2 \) act as a tunneling barrier between the graphene electrodes, which limits the current. In turn, spectacular improvements have been achieved in the ON/OFF current ratio by depositing double-layer MoS\( _2 \) onto a germanium crystal \([129]\). In this configuration the tunneling length is indeed the smallest possible. Nevertheless, the bulk-like character of the germanium side restricts the efficiency of the electrostatic control. Further performance boost is expected if tunneling takes place between two different MD layers closely surrounded by two independently regulated gate contacts, provided that a favorable band edge alignment is obtained \([130]\).

In this Chapter material combinations that fulfill this condition and are suitable for TFET operations are investigated first. Thereafter, a device design with a performance competitive with state-of-the-art Si/InAs \([48]\) and GaAsSb/InGaAs \([49]\) heterojunctions is presented.

## 5.2 Material selection

The ideal ingredients of a tunneling heterojunction are single-layer semiconductors with relatively large individual band gaps and effective masses for preventing source-to-drain tunneling, and with a large band offset such that the valence band maximum (VBM) of the \( p \)-type component is closely aligned with the conduction band minimum (CBM) of the \( n \)-type one. A broken gap or a small staggered band gap configuration enables efficient switching when applying a vertical electric field perpendicular to the interface. If the equilibrium gap is close to zero, a small electrostatic energy disparity at the two different layers can shift the VBM of the \( p \)-type material above the CBM of the \( n \)-type one, hence opening a tunneling energy window. Similarly, it can close it if the polarity is changed. We start the pursuit for such
5.2. MATERIAL SELECTION

Figure 5.1: Band gaps and band alignments of selected single-layer metal dichalcogenides as calculated with density-functional theory (DFT) and PBE-SOC functional. Red columns indicate valence bands, blue conduction bands. The data is taken from Ref. [131].

As a first choice we opted for WT$_2$ as the $p$-type component since it appears to have the highest valence band maximum among the investigated single-layer MDs. As for the $n$-type side a low conduction band minimum is desired that the five materials on the right side of Fig. 5.1 (ZrS$_2$, ZrSe$_2$, HfS$_2$, HfSe$_2$, and SnS$_2$) almost equally well satisfy. Nevertheless, for practical reasons we have restricted ourselves to the best explored and easily accessible transition metal dichalcogenides, i.e. those with Mo or W as the transition metal atom. This limitation has led to the selection of MoS$_2$, the one with the lowest CBM in this subset.
A TFET made of a MoS\textsubscript{2}-WTe\textsubscript{2} heterojunction has then been simulated and analyzed \cite{130}. However, that study should be considered more as a proof-of-concept demonstration of the usability of the simulation method described in Chapter 2 for vdW heterojunction tunneling transistors rather than a proposal for a realistic device, because of the two following main reasons:

(i) An unrealistically high (approximately 7 \%) strain had to be applied to both materials with opposite signs to compensate for the difference in the lattice parameters, which significantly modified the band structures of the materials. Note on one hand that the lattice matching is not a physical constraint since vdW heterostructures can be assembled from any two MDs due to the lack of dangling bonds at the interface. The numerical analysis, on the other hand, must start with the simulation of a periodic system that requires a unit cell small enough such that the DFT-level calculations and the generation of the maximally localized Wannier functions (MLWFs) remain computationally feasible. To overcome e.g. a 10 \% lattice mismatch between two MD layers without introducing any artificial strain, a supercell containing $9 \times 9$ primitive unit cells of one material and $10 \times 10$ of the other is required, thus increasing the number of atoms in the primitive system from 6 to 543. The supercell method becomes more appealing at larger mismatches, when the ratio of the lattice parameters can be approximated by the ratio of two small integers.

(ii) Contrary to the claims of Gong et al. \cite{131}, the semiconducting allotrope of single-layer WTe\textsubscript{2} with a trigonal prismatic (1H) configuration is found to be metastable. Although DFT calculations might indeed suggest the opposite when only the 1H and the perfect octahedral (1T) coordinations are compared, the results are reversed when the distortion of the 1T structure is allowed through the zig-zag clusterization of the metal atoms, which is a well-known phenomenon in bulk WTe\textsubscript{2} \cite{132} \cite{133} \cite{134} \cite{135} \cite{136}.

Taking all these facts into account another material combination has been chosen for further analysis. MoTe\textsubscript{2} is selected as the $p$-type component since it has the second highest VBM among the monolayer MDs presented in Fig. 5.1. For the $n$-type one SnS\textsubscript{2} is the ideal candidate, not only due to its low CBM, but also because of its large band gap and a lattice constant that only differs by 4 \% from
5.3 Simulation approach

Similarly to the procedures described in the previous Chapters, we employ the density-functional theory (DFT) tool VASP \[100\] with the PBE functional \[84\] for the electronic structure simulations. A $25 \times 25 \times 1$ Monkhorst-Pack $k$-point grid and a 500 eV plane-wave cutoff energy is used without spin-orbit coupling. The convergence criteria is set to less than $10^{-3}$ eV/Å forces acting on each ion and a total energy difference smaller than $10^{-3}$ eV between two subsequent iterations. Van der Waals interactions are included through the DFT-D2 method of Grimme \[103\].

First, the equilibrium lattice constants of the trigonal prismatic (1H) MoTe$_2$ monolayer and the octahedral (1T) SnS$_2$ are determined, 3.55 Å and 3.696 Å, respectively. The heterostructure is constructed after applying strain to both materials so that they share a common lattice parameter $a_0 = 3.623$ Å. Two different configurations are considered, one where the Sn atoms are placed above the Mo’s, i.e. they share the same $x$-$z$ coordinates, and one where they are at the same in-plane positions as the S atoms. After performing ionic
relaxations within the fixed unit cells the former one is found to exhibit the lower total energy. We thus only consider this structure hereinafter. The relaxed geometry of the heterostructure is presented in Fig. 5.2.

In order to obtain well-localized Wannier functions (WFs) a set of atomic orbitals that give a reasonably good description of the system should be supplied to the DFT tool to generate the initial projection matrices $A_{mn}(\mathbf{k})$ of Eq. (2.103). This operation requires the chemical analysis of the bands of interest first. In the case of a heterojunction it is recommended to start by the study of the isolated individual components. Hence, the band structures of single-layer SnS$_2$ and MoTe$_2$ are computed and reported in Fig. 5.3.

Additionally, the wavefunctions are projected onto spherical harmonics centered at the atoms. The coefficient that gives the highest contribution is selected for each state and used for coloring that point.
5.3. SIMULATION APPROACH

![Graph showing band structure of the SnS$_2$-MoTe$_2$ heterostructure obtained with VASP (lines) and with the MLWF Hamiltonian (circles). The conduction band is formed of an SnS$_2$ band (blue), the valence bands of the MoTe$_2$ ones (red).](image)

Eg = 0.19 eV

Figure 5.4: Band structure of the SnS$_2$-MoTe$_2$ heterostructure obtained with VASP (lines) and with the MLWF Hamiltonian (circles). The conduction band is formed of an SnS$_2$ band (blue), the valence bands of the MoTe$_2$ ones (red).

The resulting plots provide easily comprehensible visual guides for choosing initial trial orbitals that fit the problem well. For SnS$_2$ it is sufficient to select the 6 $p$-orbitals lying on the two S atoms (3 each) in order to reproduce the lowest conduction- and the 5 highest valence bands. In the MoTe$_2$ case the 5 Mo:$d$ and 6 Te:$p$ atomic states are the ones most characterizing the 4 lowest conduction- and 7 highest valence bands.

Therefore, the S:$p$, Mo:$d$, and Te:$p$ orbitals are projected onto the Bloch manifold as the first step towards the generation of the maximally localized Wannier functions (MLWFs) of the combined heterostructure. Since the conduction bands of the heterostructure become entangled, the disentanglement procedure of Sec. 2.5.3 is applied with the top of the frozen energy window set 1 eV above the CBM. The low-lying $s$-bands of the Sn, S, and Te atoms are safely excluded from the calculations.
In the MLWF Hamiltonian hopping terms up to 6-th nearest-neighbor interactions are kept, thus reproducing the DFT band structure with high precision, as illustrated in Fig. 5.4. It exhibits an indirect band gap of $E_g = 0.19$ eV. Note that the 2% artificial stress that is applied to get rid of the lattice mismatch results in a 0.1 eV lowering of the MoTe$_2$ valence band and a 0.2 eV increase of the SnS$_2$ conduction band. The fully relaxed heterostructure would therefore either form a smaller staggered gap or an even more beneficial broken gap.

The device modeled here consists of an SnS$_2$ layer deposited on an MoTe$_2$ one with an overlap of the two components at the center. Both materials extend further in one or the other direction to form the source and drain regions. The localization of the Wannier functions and the weak inter-layer interactions allow us to construct the Hamiltonian of the stand-alone MoTe$_2$ and SnS$_2$ layers from the same set of MLWF matrix elements as obtained in the DFT simulation of the periodic heterostructure. This procedure is explained in details in Appendix B.1.

The impact of the inter-layer couplings on the band structures is demonstrated in Fig. 5.5. First, the shape of the heterostructure’s Hamiltonian is depicted in subplot (a). The band edges along the $k_z = 0$ line in the rectangular unit cell used for transport is provided in subplot (b) for the heterostructure and for the separated monolayers. The band structure of the isolated layers is computed by removing the off-diagonal blocks corresponding to the inter-layer couplings from the Hamiltonian matrix. Interactions between the different layers increase the band gap, causing offsets in the band edges at the beginning and at the end of the overlap region.

For the transport calculations the selected rectangular supercell contains 16 primitive hexagonal unit cells. The out-of-plane $z$ direction is assumed periodic and modeled via 11 $k_z$ points. All simulations are performed at room temperature $T = 300$ K. Although the band gap of the heterojunction is indirect and formed between the VBM of MoTe$_2$ situated at $K$ and the CBM of SnS$_2$ at $M$, by aligning the transport axis $x$ with the $M - K$ direction, the CBM and VBM are folded to the $k_z = 0$ line. Since momentum is only conserved in the periodic (transverse) $z$ direction, but not along the transport direction due to the varying electrostatic potential, the ballistic limit
5.4. DEVICE SIMULATIONS

Figure 5.5: (a) Schematics of the MLWF Hamiltonian. The diagonal blocks contain the intra-layer-, the off-diagonals the inter-layer interactions. (b) Band structure of the MoTe$_2$-SnS$_2$ heterostructure (solid lines) and of the isolated layers (dashed lines) along the $k_z = 0$ line in the rectangular unit cell utilized for transport. The electron energies in the monolayers are computed from the Hamiltonian of subplot (a) with the inter-layer couplings set to zero.

of transport is sufficient to capture the tunneling process. This will be confirmed by the subsequent simulations and by rotating the direction of transport by 90 degrees, in which case no tunneling is observed anymore.

5.4 Device simulations

The schematics of the investigated device is presented in Fig. 5.6 (a). There is a 20 nm long overlap region between the SnS$_2$ and MoTe$_2$ layers. The gates exceed this region by 20 nm on both sides to suppress leakage. The source is covered first with hexagonal BN that has a similar permittivity as MoTe$_2$ to ensure flat-band condition at the start of the overlap region. The gates are separated from the channel materials by 3nm of HfO$_2$, corresponding to an equivalent oxide thickness EOT=0.58 nm. The top and bottom gate contacts are made of different metals to turn off the transistor at $V_{gs} = 0$ V. In the case of a $n$-type device $\Phi_{M1} - \Phi_{M2} = 0.8$ eV and a voltage is only applied to
the top gate, the bottom one being grounded. In $p$-type operations the bottom gate voltage is swept and $\Phi_{M1} - \Phi_{M2} = 0.65$ eV with $\Phi_{M1}$ being 0.35 eV lower than in the $n$-type case. To model a device with low contact resistances the source and drain extensions are doped with a high acceptor and donor concentration $N_A = N_D = 10^{13}$ cm$^{-2}$, respectively. Nevertheless, the intrinsic device performance has been found insensitive to the doping concentrations. A low supply voltage $V_{dd} = 0.4$ V is applied in all simulations.

The interlayer interaction moves up the CBM of SnS$_2$ by 60 meV compared to the isolated SnS$_2$ monolayer, as shown in Fig. 5.5 (b). This induces an undesired band offset at the overlap-drain interface that might create a leakage path. It has been found that the electrostatics favorably reduces this leakage if the drain side is surrounded by HfO$_2$ layers, unlike the source. The device band diagram without considering self-consistency (flat band condition) is reported in

Figure 5.6: (a) Schematic view of the investigated double-gate MoTe$_2$-SnS$_2$ hetero-TFET. (b) Band profiles in the device along the transport axis without considering self-consistency (flat band condition).
Figure 5.7: (Top) Band diagrams and energy-resolved transmission probability in the $n$-type TFET at various gate voltages and at $k_z = 0$. (Bottom) Spatial current density distribution at $V_{gs} = 0.2$ V and at $V_{gs} = 0.4$ V. Red: high density, blue: low density. Current conservation is ensured during the simulation, but is not visible here since the absolute value of all the in- and out-flowing currents are plotted at each atomic position. The red circles indicate the regions where tunneling occurs, both in the current-density and band diagram plots.

Fig. 5.6 (b). Switching is achieved by creating an electric field between the fixed gate and the active one, shifting the CBM of SnS$_2$ and the VBM of MoTe$_2$ by different amounts.
Figure 5.8: Transfer characteristics $I_d - V_{gs}$ of the $p$- and $n$-type TFETs on linear and logarithmic axes at $|V_{ds}| = 0.4$ V.

This mechanism is illustrated for the $n$-type device in Fig. 5.7. At $V_{gs} = 0$ V no tunneling path exists because the VBM of MoTe$_2$ is located below the CBM of SnS$_2$ in the channel region. As the gate voltage increases, the CBM of SnS$_2$ is pushed down faster than the VBM of MoTe$_2$ since it is more efficiently modulated by the top gate. At $V_{gs} = 0.2$ V, though electrons propagating in the VB of MoTe$_2$ still cannot directly enter into the CB of SnS$_2$ at the center of the device, they can tunnel from the rightmost MoTe$_2$ cells into the drain extension, the CBM being lower in the monolayer than in the heterostructure.

In the device ON-state, at $V_{gs} = 0.4$ V, the MoTe$_2$ and SnS$_2$ bands are completely reversed in the overlap region. Electrons can flow through the whole interface area, as shown at the bottom of Fig. 5.7. This process delivers a high current since the tunneling length corresponds to the distance between two atomic layers. By fixing the OFF-state current of the device to $10^{-6}$ $\mu$A/$\mu$m through the gate work functions the ON-current reaches $75$ $\mu$A/$\mu$m. The corresponding $I_d - V_{gs}$ transfer characteristics are shown on the right side of Fig. 5.8. The inverse subthreshold slope is $25$ mV/dec at 0 V and it remains lower than $60$ mV/dec for more than 6 orders of magnitude.
5.5 Conclusions

We have shown that single-layer SnS$_2$ stacked on MoTe$_2$ forms a type-II heterojunction with a small staggered band gap, favorable for tunneling transistors. A double-gate device based on this heterostructure has been simulated and a promising steep-slope behavior has been observed. At a low supply voltage $V_{dd} = 0.4$ V and OFF-current
$I_{\text{off}} = 10^{-6} \, \mu A/\mu m$ an ON-state current $I_{\text{on}} = 75 \, \mu A/\mu m$ has been demonstrated for both $n$- and $p$-type devices. It appears that van der Waals heterostructures might offer new perspectives for the realization of TFETs with a steep inverse subthreshold slope, low OFF-, and high ON-current. The proposed device concept along with the analysis presented here were published in Electron Device Letters [137].
Chapter 6

Metal contacts in 2D semiconductors

6.1 Introduction

Despite exhibiting fascinating properties for digital logic applications, the performance of devices based on 2D materials is often limited by the contact resistance of the metal interconnects. Realizing transparent contacts represents one of the greatest challenges in the integration of few-layered materials into electrical circuits [50] [51]. As long as the physics of the transport mechanism through metal-semiconductor interfaces is not fully understood, as in the case of 2D systems, interpreting measurements on the investigated materials becomes increasingly difficult. Ab initio computer simulations can provide invaluable insights into the underlying phenomena in these situations.

Recent studies on this topic have mainly focused on the equilibrium properties of such interfaces, as the Schottky barrier height, tunneling length, or charge distribution [54] [55] [138]. However, theoretical [55] as well as experimental [139] works both indicate that there are several more issues to be addressed in order to understand the exact nature of these material combinations. It is still an open question what path the electric current takes and how the transport properties of the semiconductor change when a single-layer material is partly covered
by a metallic contact. Answering these questions might help design more efficient 2D-metal contacts.

Scandium has recently been proposed as an efficient, low resistance contact to MoS$_2$ \cite{42}. Therefore, in this Chapter scandium contacted to single-layer MoS$_2$ is investigated through quantum transport simulations based on first principles. Current paths, electrostatic potential energy profiles, and contact resistances are extracted in various device configurations under bias to shed light on the non-equilibrium transport properties of the system.

### 6.2 Simulation setup

The approach employed for simulating transport through an interface between a metal and a 2D semiconductor is detailed in greater depth in Appendix B.2. Here, a brief summary is presented along with the applied numerical parameters.

Six layers of scandium in a simple hexagonal lattice is modeled in a common unit cell with the MoS$_2$ monolayer. The in-plane lattice constant is set to $a_0 = 3.183$ Å, i.e. 4% compressive strain is applied to the metal. In the vertical direction $y$ a 10 Å vacuum region separates the periodic replicas. The ions are relaxed within the unit cell except for the top 4 Sc atoms that are kept at fixed positions. The geometric optimization is performed using VASP with the PBE functional. Van der Waals forces are included through the DFT-D2 method of Grimme. A $11 \times 1 \times 11$ Monkhorst-Pack $k$-point mesh and a convergence criteria of less than $10^{-3}$ eV/Å forces acting on each atom is chosen. The vertical separation between the metal and the semiconductor layers is consistently found to be 1.75 Å, regardless of the initial assumptions.

A rectangular unit cell is constructed from two optimized primitive hexagonal unit cells and repeated 40 times along the transport direction $x$. The metal atoms are removed from the last 15 replicas so that only MoS$_2$ is present in that region, as shown in Fig. 6.1. The system consists of 540 atoms in total. The Brillouin zone is sampled with 5 points in the $k_z$ direction at $k_x = k_y = 0$ when the charge density is self-consistently determined. It is then increased to 9 as the wavefunctions are recomputed with the charge fixed to the
6.2. SIMULATION SETUP

Figure 6.1: Atomic geometry of the scandium-MoS$_2$ interface simulated by VASP and schematics of the extraction of the MLWF Hamiltonian blocks utilized for constructing the Hamiltonian matrix of a device with realistic physical dimensions. Blue rectangles identify the regions from where interactions between cells containing only the metal atoms are taken, red the complete unit cells with the metal and semiconductor parts overlapping, and orange the semiconductor-only part. The green rectangle marks the central part that has to be connected to the left and right extensions.

pre-converged value. The reason for computing the wavefunctions at a larger number of $k$ points than what the self-consistency requires is that the number of neighbor cell interactions contained in the resulting real-space MLWF Hamiltonian is proportional to the size of the initial $k$-point set. The number of bands retained in the simulation is also increased (NBANDS=3600) so that a smoother subspace of bands can be disentangled from the Hilbert space. As initial projections Mo:1=2, S:1=1, Sc:1=2, and Sc:sp3-1 are supplied to Wannier90. The frozen energy window covers a 2 eV range around the Fermi level. The resulting Wannier functions are well localized around the atom centers with most of the quadratic spreads measuring between 1.5 and 2 Å$^2$. Even the largest one remains smaller than 5.5 Å$^2$.

From the MLWF Hamiltonian of this structure sub-matrices that describe interactions between different parts of the system (scandium only, scandium and MoS$_2$ overlapping, and MoS$_2$ only) are identified. They are used for creating a device Hamiltonian with extended, more realistic dimensions, as sketched in Fig. 6.2. It contains an 8 nm long Sc extension, an 54 nm overlapping region, and 50 nm of stand-alone
Figure 6.2: Schematics of a scandium top contact deposited on a MoS$_2$ strip with extended lateral dimensions. In the initial setup $L_M$ measures 8 nm, $L_O$ 54 nm, and $L_S$ 50 nm.

MoS$_2$. The width of the strip $L_z$ in the $z$ direction is also increased by a factor of three, making it 1.65 nm so that interactions beyond nearest neighbors can be neglected.

The Fermi level of the metal contact is determined by the charge neutrality condition in the left metallic contact and kept fixed during the transport simulations. The periodicity of the device is modeled by 7 $k_z$ points in the NEGF calculations, which has been found sufficient, the results being not affected by the inclusion of more points.

### 6.3 Investigation of the transfer length

First, the influence of the metal-semiconductor interface’s area on the device current is studied. The transmission probability $T(E)$ through the system presented in Fig. 6.2 is computed with our NEGF simulator as the function of the incident electron energy $E$ without applying any external electrostatic potential. The current $I_d$ is then evaluated through the Landauer-Büttiker formula

$$I_d = -\frac{2 \text{spin} e}{\hbar} \frac{1}{N_k} \sum_k \int_{C_{BM}}^{\infty} \frac{dE}{2\pi} T(E, k)(f(E, E_f^L) - f(E, E_f^R)).$$

In Eq. (6.1) the summation is carried over all the $k$ points indexed from 1 to $N_k$, while the integration goes from the conduction band minimum (CBM) of MoS$_2$ on the drain side up to infinity. The variable $e$ denotes the elementary charge, $E_f^L$ and $E_f^R$ the Fermi levels of the left and right contacts, respectively, and $f(E, E_f)$ is the Fermi-Dirac distribution. A source-to-drain bias $V_{ds} = 0.5$ V is set.
that determines the difference between the left and right Fermi levels
\[ E_f^R = E_f^L - eV_{ds}. \]

The same procedure is repeated for a device where the length of
the metal-semiconductor overlap region is reduced to \( L_O = 6.5 \text{ nm} \).
Despite the 8-fold reduction in the interface area the difference in the
total currents is less than 1 %. It indicates that the main transport
mechanism is an edge process that does not depend on the length
of the interface along the transport direction \( x \). The transmission
probabilities in the two different devices are compared to each other in
Fig. 6.3(a). The differences between the curves are negligible. Further
tests have been conducted where connections between neighbor MoS\(_2\)
cells below the metal contact have been removed from the Hamiltonian
matrix. The results are found to be insensitive to these changes.
These calculations support the claim that the least resistance path is
the one where the current stays as long as possible in the metal before
tunneling into MoS\(_2\).

This finding seems to contradict the conclusions of Refs. 52 and
53. In both studies currents flowing through single- or few-layer
MoS\(_2\) samples were measured between contacts with varying lengths,
and contact length dependences of the resistances were found. In the
first one a transfer length of more than 600 nm was extracted in single-
layer MoS\(_2\) at a back gate bias \( V_g = 100 \text{ V} \). In the second one two- and
six-layers MoS\(_2\) samples were investigated. The transfer length was
found to increase from about 20 to 70 nm in the double-layer device
in an almost linear manner as the gate voltage was swept from 0 to
70 V, and from about 40 to 170 nm in the 6-layers one.

Although the above experiments were performed with Ti as the
contact metal, the workfunctions of Ti and Sc deposited on MoS\(_2\)
are quite similar and both configurations result in the metal wave
functions significantly overlapping with the semiconductor orbitals,
i.e. penetrating into the band gap of MoS\(_2\). Thus, we believe that the
difference in the conclusions between these experiments and our work
does not originate from the choice of the contact material.

As a next step, we would like to understand why our contact
properties do not exhibit any length dependence. Our simulation
results predict a Schottky Barrier height of 215 meV in the Sc+MoS\(_2\)
system. This value can be extracted in Fig. 6.3(b), where the density-
of-states (DOS) originating from MoS\(_2\) only is plotted, i.e. the blocks
Figure 6.3: (a) Transmission probability from the metal into the MoS$_2$ as a function of the energy in a structure with 54 nm overlap length (solid blue line) and in one with 6.5 nm overlap (dashed red lines). (b) Density-of-states as a function of energy and spatial position in the MoS$_2$ layer where the metallic part of the Hamiltonian matrix was removed. Blue: no density, yellow: high density. The $x = 0$ position is set to the beginning of the MoS$_2$ layer. $E_f$ marks the Fermi level of the metal and $\Phi_{SB}$ the Schottky barrier height.

that correspond to Sc in the total Hamiltonian matrix have been removed before the DOS is computed. In the overlap region ($0 \leq x \leq 54$ nm), the conduction band minimum (CBM) is situated at an energy $E = 1.035$ eV, while the Fermi level lies at 0.82 eV. In the MoS$_2$ extension ($x > 54$ nm), the CBM has a lower energy $E = 0.816$ eV because there is no transfer of electrons from Sc to MoS$_2$ in this region. In effect, the wave function of the metal in the energy range around the
Fermi level can penetrate into the band gap of MoS$_2$ at $0 \leq x \leq 54$ nm, as shown in Fig. 6.4 (a). These additional electrons raise the CBM and pin the Fermi level below the metallic contact.

The energy-resolved spectral current in the investigated device is presented in Fig. 6.4 (b). As a consequence of the band diagram in Fig. 6.3 (b), the electrons that have the highest probability to tunnel from Sc into MoS$_2$ are those situated around $x = 54$ nm and at an energy close to the Fermi level since they have the shortest tunneling distance. This explains why we have a current flowing at the edge of the metal. According to Refs. [53] and [139], the length dependence of the contact should come from a variation of the electrostatic potential along the $x$ axis in the overlap region. This does not seem possible in single-layer structures because, as mentioned
above, the Fermi level is pinned by the metal-induced states there. The situation might be different in few-layer MoS$_2$ where the metal wave function cannot penetrate into the layers that are sufficiently far away from the contact, thus allowing for electrostatic potential modulations and possible length-dependent contact currents.

Note that a recent computational study based on DFT suggested that a Ti top contact can metallize MoS$_2$, which could explain the large transfer length [55]. We also observe non-zero partial density-of-states associated with the Mo or S atoms within the band gap of MoS$_2$ with Sc as the top contact, as shown in Fig. 6.4 (a). However, we disagree with their conclusion. As it is demonstrated in Fig. 6.3 (b), no mid-gap states are apparent in the MLWF basis if only the subspace with WF centered on Mo and S atoms is considered. In Ref. [55], the wavefunctions that contribute to the band gap DOS on Mo and S atomic orbitals are likely the more delocalized metal wavefunctions that overlap with the semiconductor. These decaying states into the MoS$_2$ layer do not participate in transport, similarly to the wavefunction at a metal-insulator interface that penetrates into the insulator, but does not induce conductive states in the insulator. The experimentally observed sheet resistance of MoS$_2$ where it is covered by metal [53] [139] also contradicts the hypothesis that the semiconductor states are metallized.

### 6.4 Contact resistance calculation

As it appears that the defect-free overlap area does not influence the transport properties of the device, what remains is to investigate the dependence of the contact resistance on the electrostatics. Varying the doping concentrations, source-to-drain and source-to-gate biases, oxide thicknesses, and oxide materials results in a rather large design space to explore. In order to minimize the computational burden associated with the self-consistent NEGF simulations, we divided this task into two parts. First, the shape of the electrostatic potential profile is analyzed at zero source-to-drain bias, then the resistance is computed using a set of different electrostatic potential functions.

Through self-consistent NEGF-Poisson simulations it has been determined that the electrostatic potential profile of the device around
6.4. CONTACT RESISTANCE CALCULATION

\[ \Phi = a + b \ e^{-x/\lambda} \]

Figure 6.5: Schematics of the MoS\(_2\) band profile in the device. In the region where the metal overlaps the MoS\(_2\) a constant \(\Phi_{SB} = 0.215\) V Schottky barrier is present between the MoS\(_2\) CBM and Sc Fermi level that remains insensitive to the back gate voltage. The source-to-drain bias \(V_{ds}\) does not influence the electrostatics in the back-gate controlled middle part where the potential can be approximated by a decaying exponential function with \(\lambda\) as characteristic length.

The source contact can be well approximated by a decaying exponential function with a characteristic length \(\lambda\) regardless of the oxide and doping properties, as illustrated in Fig. 6.5. In relatively long channel transistors the influence of the source-to-drain bias \(V_{ds}\) can be neglected in the source contact area. Hence, it is not directly accounted for in the simulations.

The \(\lambda\) parameter that primarily affects the resistance has been investigated in various configurations. First, its dependence on the oxide thickness has been analyzed when MoS\(_2\) is undoped and a back gate bias \(V_{gs} = 0.5\) V is set. Note that \(V_{gs} = 0\) V corresponds to the case where the potential in the middle of the channel is aligned with the source Fermi level. Secondly, the dependence of \(\lambda\) on the doping concentration has been evaluated without applying a gate bias. SiO\(_2\) is employed as the oxide material with a relative dielectric permittivity \(\varepsilon_{SiO_2} = 3.9\). The results are summarized in Fig. 6.6 (a). The characteristic length \(\lambda\) quickly decreases as the doping concentration increases, allowing for more efficient injection of electrons into MoS\(_2\).
Figure 6.6: (a) Dependence of the characteristic length $\lambda$ of the exponentially decaying potential profile as a function of the doping concentration (blue circles) at zero gate bias and as a function of the oxide thickness (red triangles) without doping. (b) Transfer characteristics $I_d - V_{ds}$ of the Sc-contacted MoS$_2$ at various $\lambda$ values and at a gate bias $V_{gs} = 0.5$ V.

Transport through the Sc+MoS$_2$ contact structure is then simulated with the electrostatic potential represented as an exponentially decaying function with various characteristic lengths $\lambda$. Figure 6.6 (b) shows the drain current $I_d$ vs. source-to-drain bias $V_{ds}$ curves at $\lambda = 1, 2, 5$, and 10 nm. The influence of $V_{ds}$ is simply accounted for by shifting the Fermi level of the drain contact in Eq. (6.1). The shortest $\lambda$ results in the highest current due to the reduced injection barrier width. The contact resistances are extracted from the linear regime of the curves at small bias. All simulations have been performed through a narrower barrier. The $\lambda$ vs. oxide thickness relationship is approximately linear. A smaller oxide thickness leads to a better gate control that again decreases the width of the injection barrier.
in the ballistic limit of transport, where the semiconductor does not exhibit length-dependent resistance due to the absence of scattering mechanisms. The inverse of the $I_d - V_{ds}$ slope is therefore the contact resistance $R_c$ itself. It is found to be $R_c = 25.9$, $54.4$, $111.2$, and $203.5 \ \Omega \mu m$ at $\lambda = 1$, $2$, $5$, and $10$ nm, respectively.

In Ref. [140] an interface resistance of $R_c = 650 \ \Omega \mu m$ is reported for Scandium contacts on MoS$_2$, which is in a relatively close range to our calculations, especially knowing that, typically, $R_c$’s of several $k\Omega \mu m$ are experimentally measured. The difference is likely coming from impurities or defects at the metal-semiconductor interface that are not modeled in our work. They might create a current path through the MoS$_2$ layer below the metal, but also increase the sheet resistance and extend the width of the Schottky barrier. The investigation of these effects are, however, out of the scope of this work. It is also apparent that the electrostatics of the device plays a major role in achieving efficient electron injection from the metal into the semiconductor. A suitable doping density and appropriate gate control can reduce the contact resistance by at least an order of magnitude.

6.5 Conclusions

We have simulated transport through a Sc-MoS$_2$ contact interface. It has been found that in case of a defect-free surface the electric current flows along the metal and enters into the MoS$_2$ layer at the edge of the metal contact only. Such trajectories are due to the absence of conductive mid-gap states and to the Fermi-level pinning caused by charge transfer from the metal. The electrostatic potential profile has been analyzed in various device settings: a single exponential function accurately fits its behavior as a function of the doping concentration or oxide thickness. Contact resistances have been evaluated at various Schottky barrier widths that might depend on several factors, e.g. gate biases, doping densities, and oxide materials.
Chapter 7

Conclusion and outlook

The simulation framework developed for the study of 2D semiconductor devices and demonstrated in the previous chapters of this thesis is ready to be applied to other problems as well. Extending the method to properly model bulk 3D materials or 1D systems requires only minor modifications of the available codes. The on-going works utilizing this approach at the Integrated Systems Laboratory of ETH Zürich include the investigation of Si and InAs nanowire FETs, Li-ion batteries, MgO-Fe spintronics devices, and hBN-Au antennas. To conclude this work, however, we stay in the realm of 2D materials. In order to complete the analyses, we have systematically investigated the properties of the most widely studied transition metal dichalcogenides (TMDs) in addition to what has already been presented. We have computed the hole- and electron mobilities of single-layer MoSe$_2$, MoTe$_2$, WS$_2$, and WSe$_2$ as a function of their carrier concentrations.

In Appendix C improvements of the simulation methodology are introduced. They mainly concern the calculation of the electron-phonon scattering self-energy for which important contributions coming from beyond nearest-neighbor atomic interactions had been neglected as a first approximation. The carrier mobilities of the MoS$_2$ and black phosphorus (BP) monolayers have been re-evaluated in light of these corrections. Consequently, the phonon-limited electron mobility of MoS$_2$ at low carrier concentrations increases to 495 cm/Vs, a value close to what other theoretical groups reported [97].
In the following calculations the same approach as in Appendix C is used for all considered TMD monolayers. The lattices and the ions are fully relaxed using the PBE functional and the derivatives of the Hamiltonian matrix elements are retained up to third nearest-neighbor distances in the evaluation of the scattering self-energy term in Eq. (2.64). The results are presented in Fig. 7.1 along with the mobilities of MoS$_2$ and BP obtained by this more accurate procedure.

Although the hole mobility of BP is significantly reduced now compared to what was claimed in Chapter 4, it still surpasses all the investigated TMDs when transport occurs along its armchair direction. The properties of TMDs are rather isotropic, thus no distinctions are made between their different crystal orientations. The electron mobility of the armchair-oriented BP underwent an even more severe
Table 7.1: Summary of the phonon-limited hole- ($\mu_h$) and electron mobilities ($\mu_e$) at low carrier concentrations and hole- ($m_h^*$) and electron effective masses ($m_e^*$) of the single-layer semiconductors considered in this work.

<table>
<thead>
<tr>
<th></th>
<th>$\mu_h$ (cm$^2$/Vs)</th>
<th>$m_h^*$ ($m_0$)</th>
<th>$\mu_e$ (cm$^2$/Vs)</th>
<th>$m_e^*$ ($m_0$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MoS$_2$</td>
<td>74.6</td>
<td>0.59</td>
<td>494.7</td>
<td>0.46</td>
</tr>
<tr>
<td>MoSe$_2$</td>
<td>195.2</td>
<td>0.61</td>
<td>294.7</td>
<td>0.54</td>
</tr>
<tr>
<td>MoTe$_2$</td>
<td>121.5</td>
<td>0.85</td>
<td>235.2</td>
<td>0.54</td>
</tr>
<tr>
<td>WS$_2$</td>
<td>364.1</td>
<td>0.38</td>
<td>1060</td>
<td>0.30</td>
</tr>
<tr>
<td>WSe$_2$</td>
<td>400.1</td>
<td>0.49</td>
<td>639.7</td>
<td>0.33</td>
</tr>
<tr>
<td>BP:AC</td>
<td>641.6</td>
<td>0.15</td>
<td>311.6</td>
<td>0.18</td>
</tr>
<tr>
<td>BP:ZZ</td>
<td>47.7</td>
<td>2.5</td>
<td>67.5</td>
<td>1.2</td>
</tr>
</tbody>
</table>

reduction due to the inclusion of beyond nearest-neighbor electron-phonon coupling terms. The fully converged simulation results now predict a lower electron mobility for BP than for MoS$_2$.

It is also clear from Fig. 7.1 that the phonon-limited carrier mobilities are higher in those TMDs that are composed of tungsten rather than molybdenum as the metal atom type. This phenomenon is attributed on one hand to the heavier atomic mass of W that scales the electron-phonon matrix elements according to Eq. (2.67) and on the other hand to the lower carrier effective masses of tungsten-based components. The highest electron mobility is observed in WS$_2$ that also possesses the lowest electron effective mass value $m_e^* = 0.3$ $m_0$. However, the hole mobility of WS$_2$ is lower than that of WSe$_2$, despite the smaller hole effective mass $m_h^* = 0.38$ $m_0$ of the former versus $m_h^* = 0.49$ $m_0$ for the latter. Besides the effective mass value at the valence band maximum ($K$ point) the hole transport properties of TMDs also strongly depend on the exact value of the small energy separation between the $K$ and $\Gamma$ valleys and therefore the effective mass at $\Gamma$.

The results are also summarized in Table 7.1. Our calculations suggest that at gate lengths where source-to-drain tunneling is not yet the most dominant factor determining the performance of a device,
WS$_2$ $n$-FETs and armchair-oriented BP $p$-FETs might be the most promising choices for high-performance 2D logic switches.
Appendix A

Assembling the device Hamiltonian

In this Appendix the process of creating the Hamiltonian matrix of a two-dimensional transistor is detailed through a step-by-step example consisting of a single-layer MoS$_2$. The density-functional theory (DFT) simulations are performed using VASP. The input files along with the scripts utilized for processing the data are available on the website of the Nano-TCAD group of ETH Zürich at https://www.nano-tcad.ethz.ch. On this subject only a few remarks specific to 2D systems are made along with some useful hints for preparing the inputs for Wannier90.

It is worthwhile noting that depending on the DFT tool the volume relaxation routines might fail when one lattice vector is supposed to be fixed. This is the case of 2D materials where a large out-of-plane lattice vector models the separation of the monolayer from its periodic replicas by a vacuum region. Therefore, it is better to perform ionic relaxations within fixed unit cells with various lattice constants $a_0$. Example calculations of this kind are provided by the VASP developer group at http://cms.mpi.univie.ac.at/wiki/index.php/Fcc_Si. For improved accuracy a parabola can be fitted to the discrete total energies. The minimum position determines the value of the optimal $a_0$. It is found to be 3.183 Å for single-layer MoS$_2$ using the PBE
APPENDIX A. ASSEMBLING THE HAMILTONIAN

Figure A.1: Fit of a quadratic curve onto the total energy of the MoS$_2$ monolayer versus the lattice constant. The minimum point is determined as $a_{0,\text{optimal}} = -\frac{p_2}{2p_1} = 3.183$ Å.

After determining the lattice parameter an ionic relaxation with a higher precision is performed in the optimized unit cell. Once the geometry is relaxed and the charge density is computed, the process continues by calculating the band structure along the high-symmetry lines in the Brillouin zone. In order to obtain useful information about the chemical nature of the bands of interest, the wavefunctions are projected onto spherical harmonics centered around the ions. In VASP this is achieved by setting the LORBIT=10 parameter in the INCAR file. The coefficient that gives the highest contribution is selected for each state and used for coloring that point when plotting the band structure, as shown in Fig. A.2. A MATLAB script is provided among the accompanying files for performing this task.

In the case of MoS$_2$ the two lowest energy sulfur: $s$ bands can be safely excluded from the calculation of the maximally localized Wannier functions (MLWFs). Furthermore, each state above the fourth conduction band is also separated by a finite energy gap from the 11 entangled bands that are composed of a mixture of molybdenum:$d$
Figure A.2: Band structure of single-layer MoS$_2$. Each state is projected onto spherical harmonics centered at the atoms and colored according to the highest contribution. The vacuum level is set to 0 eV. The dashed horizontal line marks the mid-gap energy. The 4 lowest conduction bands and the 7 highest valence bands of the MoS$_2$ are primarily of molybdenum:d and tellurium:p character (orange and purple lines, respectively).

and sulfur:p orbitals. The overlap matrices $M_{mn}(\mathbf{k})$ in Eq. (2.113) and the initial projections $A_{mn}(\mathbf{k})$ in Eq. (2.103) are computed with the NBANDS=13 setting in the VASP INCAR file and exclude_bands=1-2 in wannier90.win with Mo:1=2 and S:1=1 as trial orbitals. The method developed for composite bands is employed, as detailed in Sec. 2.5.2. The hr.plot = T and write.xyz = T flags are necessary for generating and saving the MLWF Hamiltonian matrix elements into the wannier90.hr.dat file, and for writing the coordinates of the Wannier function centers along with the atom coordinates in the unit cell into the wannier90.centres.xyz file.

Since the MLWFs are well localized around the atom centers, they will be represented as point charges at the atom positions when Poisson’s equation is solved on a finite element grid. As a first approximation, the MLWFs in the immediate vicinity of an atom are grouped together and form one single point charge. All inputs for this task are contained in wannier90.centres.xyz. From now on the atoms
are defined by their coordinates in the primitive unit cell used in the DFT simulation and by a set of MLWF indices.

The MLWF Hamiltonian matrix elements decay quickly with the distance between the Wannier functions. Consequently, a cutoff distance $r_{cut}$ can always be set such that interactions beyond this range are negligible. This parameter is determined by analyzing the variations in the band structure features with respect to the maximum interaction range, as in Fig. 3.1. This can be done using Wannier90, which is capable of calculating the band structure at various $r_{cut}$ values.

The next step is the selection of a rectangular unit cell for the transport simulations. The dimensions of the supercell are defined in terms of the primitive lattice vectors with the requirement that the lengths of the new lattice vectors must exceed the value of $r_{cut}$. Hence, all retained interactions are confined within two neighbor cells at most.

The confinement of inter-atomic interactions is an important requirement since the periodicity of the device along the $z$ direction is modeled by a $k_z$-dependence of the Hamiltonian $H(k_z) = H_0 + H_+ \cdot e^{-i k_z \Delta} + H_- \cdot e^{i k_z \Delta}$, as discussed in Section 2.3.2 and shown in Fig. [A.3](a). Therefore, second nearest-neighbor supercell interactions in the $z$ direction will not be included in the device Hamiltonian. Furthermore, the $H_0$ and $H_{\pm}$ matrices themselves can be built up from blocks corresponding to nearest-neighbor supercell interactions if the transport unit cell is also sufficiently long along the $x$ direction.

A supercell that is large enough to fulfill the above condition for interactions up to 6-th nearest-neighbor atoms is sketched in subplot (a). Its lattice vectors $\mathbf{A}_{x/z}$ in terms of the primitive lattice vectors $\mathbf{a}_1$ and $\mathbf{a}_2$ shown on subplot (b) are given as $\mathbf{A}_x = (4, -4)$ and $\mathbf{A}_z = (2, 2)$.

An array of all primitive unit cells contained within the supercell and holding at least one atom is then created. It stores their lattice vector indices and the indices of the atoms within the primitive cell. This operation is achieved by looping over primitive unit cells with increasing $(i, j)$ lattice vector indices and over each atom within the unit cell with $\mathbf{r}_k$ coordinates. The $\{i, j, k\}$ combinations that satisfy $0 \leq R_x < L_x$ and $0 \leq R_z < L_z$ are saved until the total number of atoms in the array matches the number of unit cell atoms times the number of primitive cells composing the supercell. The variables $R_x$
Figure A.3: (a) Top view of single-layer MoS$_2$. The $z$ direction is assumed to be periodic and modeled through a set of $k_z$ points. The device Hamiltonian $H(k_z)$ is computed from the $H_0$ Hamiltonian of an MoS$_2$ strip with width $\Delta$ and the $H_-$ and $H_+$ couplings as $H(k_z) = H_0 + H_- \cdot e^{-ik_z\Delta} + H_+ \cdot e^{ik_z\Delta}$. The shaded areas mark the primitive hexagonal unit cell of MoS$_2$ used in the electronic structure calculations, while the rectangular one is utilized in the transport calculations. (b) A possible tiling of the rectangular supercell with the hexagonal unit cells.

and $R_z$ refer to the $x$ and $z$ components of the atom coordinates in the supercell $\mathbf{R} = i\mathbf{a}_1 + j\mathbf{a}_2 + \mathbf{r}_k$, while $L_x$ and $L_z$ are the lengths of the supercell lattice vectors $\mathbf{A}_x$ and $\mathbf{A}_z$. If e.g. the top left corner of Fig. A.3 (b) is chosen as the origin, then there will be only a single Mo atom (red ball) in the $(0, 0)$ primitive cell, while two S and one Mo atom (each yellow ball represents two overlapping S atoms) in the $(1, 0)$ primitive cell, which is complete. In the indicated $(2, 2)$ primitive cell (bottom left corner) there are only two S atoms.

With this setup it is straightforward to compose the $H_{\text{B}_{\text{center}}}$ block of the Hamiltonian matrix containing the interactions within a supercell. From Wannier90 we have got the $H_{ijmn}$ MLWF matrix elements describing the couplings between the $m$-th Wannier function in the $(0, 0)$ primitive unit cell and the $n$-th WF in the one with indices $(i, j)$. The $(w_1, w_2)$ entries of the supercell Hamiltonian $H_{B_{\text{center}}^{ew_1w_2}}$ are
created by taking $H_{ijmn}$ at the appropriate indices that correspond to the same WFs in the primitive cells as $w_1$ and $w_2$ and to the same lattice vector as the one pointing from the center of the unit cell of $w_1$ to the one containing $w_2$. We therefore have 6 nested loops in total: $w_1$ is increased one by one as we loop over all unit cells, all atoms within each unit cell, and all WFs centered around each atom. For each $w_1$ we also have to run $w_2$ through the same 3 loops. The difference of the lattice vectors of the two selected unit cells identify $i$ and $j$, while the internal indices of $w_1$ and $w_2$ in their respective unit cells give $m$ and $n$. If the distance of the atoms is smaller than $r_{cut}$, we fill up the matrix as $HB_{w_1w_2}^{\text{center}} = H_{ijmn}$.

In order to get the nearest-neighbor supercell couplings, the same procedure is repeated with the only difference that a supercell lattice vector is added to the position of the second atom before evaluating the distance of the atoms and determining the $(i, j)$ relative lattice vector indices. For consistency in the notations, $HB_{\text{center}}$ is denoted $H_{22}$ from now on, $H_-$ as $H_1$, $H_0$ as $H_2$, and $H_+$ as $H_3$. There are 9 blocks $H_{ij}$ containing all interactions required to create $H_1$, $H_2$, and $H_3$, as shown in Fig. [A.4] The first index $i$ of $H_{ij}$ refers to the relative $z$ position of any neighbor supercell with respect to the central one,
$L_z$, 0, and $L_z$ for $i = 1, 2$, and 3, respectively. The second index $j$ marks the relative $x$ positions: $-L_x$, 0, and $L_x$ for $j = 1, 2$, and 3.

The matrices $H_i$ are assembled from the $H_{ij}$ blocks placed in a block-tridiagonal fashion. The diagonal blocks are the replicas of $H_{i2}$, the lower off-diagonals of $H_{i1}$, and the upper off-diagonals of $H_{i3}$, as sketched in subplot (b). The number of the diagonal blocks must match the desired number of supercells that the device consists of. The size of the matrices $H_i$ is then $N_{\text{cells}} \times N_{W\text{Fs,per.cell}}$, with $N_{\text{cells}}$ denoting the number of supercells in the device and $N_{W\text{Fs,per.cell}}$ the number of Wannier functions in a supercell. Our NEGF simulator takes $H_1$, $H_2$, and $H_3$ as the inputs to calculate the $H(k_z)$ matrices for a number of $N_{k_z}$ different wavevector values $k_z$ defined at runtime.
Appendix B

Modeling of interfaces

B.1 Van der Waals heterojunctions

As demonstrated in Chapter 5, transport through van der Waals heterojunctions can be accurately simulated using the Wannier functions obtained from DFT calculations in the unit cell of the periodic compound system. This approximation relies on the weak second order interactions that hold the layers together. They have negligible influence on the on-site energies and on the stronger intra-layer couplings. Therefore, the sub-space of the heterostructure’s Hamiltonian associated with one or the other semiconductor layer gives a reliable description of the isolated monolayers.

Nevertheless, the modeling of the system with periodic boundary conditions can be cumbersome itself. When the lattice constants of the components differ, there are two options to confine the problem into a single unit cell. The first one is to apply strain with opposite signs to the individual layers so that they share a common lattice parameter. This is only viable if the mismatch is small enough so that the required strain value does not modify the relevant transport properties of the materials.

The other one is to create a supercell that consists of different number of replicas of the two materials. If the lattice constant of the first one is e.g. 50 % larger than the second one, then a supercell
Figure B.1: Unit cell of the (a) MoTe$_2$-SnS$_2$ and (b) MoS$_2$-MoTe$_2$ heterostructures. Cyan balls represent Mo atoms, purple Te, gray Sn, and yellow S.

made of $2 \times 2$ replicas of the first one and $3 \times 3$ of the second one represents the unit cell of the heterostructure. In general, however, the lattice mismatch is not the ratio of two small integers. Hence, in order to keep the size of the system in the computationally feasible range of DFT-level calculations, the two approach should be mixed. The smallest supercell that consists of integer numbers of the ingredients’ replicas with a predefined maximum strain value has to be found.

Figure B.1 (a) shows how the MoTe$_2$-SnS$_2$ heterostructure investigated in Chapter 5 can be modeled by a 6-atomic unit cell. The lattice parameter of MoTe$_2$ is $a_0 = 3.55$ Å, while it is 3.69 Å for SnS$_2$. The 4% mismatch is cleared off by applying 2% compressive- and 2% tensile strains to the SnS$_2$ and MoTe$_2$, respectively. On subplot (b) the unit cell of a MoS$_2$-MoTe$_2$ heterostructure is represented. The lattice constant of MoS$_2$ is $a_0 = 3.183$ Å, that is 11.5% smaller than the MoTe$_2$ one. The difference is overcome by a 0.5% strain on both layers and by the construction of a supercell with $9 \times 9$ replicas of MoS$_2$ and $8 \times 8$ of MoTe$_2$.

The algorithm to create the MLWF Hamiltonian matrix of a heterojunction device is insensitive to the number of primitive unit cells the individual layers in the supercell consists of. The procedure described in Appendix A does not depend on the shape or size of the unit cell utilized in VASP and Wannier90 and the number of atoms
in it. The difference between assembling the Hamiltonian matrix of a MOSFET with a homogeneous channel material and of a heterojunction tunneling FET resides in the distinction of regions where the two single-layer materials overlap and where either one or the other is present only.

In Fig. B.2 (a) a MoTe$_2$-SnS$_2$ heterojunction is sketched as an illustration. It consists of only 6 transport unit cells whose dimensions are determined according to Appendix A by the criterion of confining
all interactions between maximum 6-th nearest-neighbor atoms within two neighbor cells at most. The green rectangles at the center of the structure mark the complete unit cells where the two materials overlap. On the left side the red rectangles indicate the unit cells with only MoTe$_2$ in them, while the yellow rectangles on the right refer to the truncated unit cells consisting of Sn and S atoms.

In our approach the Hamiltonian matrix of this heterojunction is the same as if selected rows and columns of a device Hamiltonian that is made of 6 complete unit cells had been removed, i.e those corresponding to the Wannier functions centered on Sn or S atoms in the first 2 unit cells and on Mo or Te atoms in the last 2 cells. The sparsity pattern of the $H_0$ matrix containing the interactions within a strip that is periodic in the out-of-plane $z$ direction is shown in Fig. B.2 (b) with the blue area marking the non-zero entries. The red, green, and yellow squares indicate the blocks that contain the interactions within the corresponding unit cells in subplot (a). The black rectangles represent the interactions between neighbor cells. Since there are 5 WFs centered on each Mo atoms and 3 on the Te, while 3 on S and none on the Sn, the various block types have different sizes.

In practice the matrix is not created by truncating a larger one. Instead the required block types are generated first by selecting the appropriate subspaces from the complete blocks containing interactions within or between the complete unit cells. They are then placed in a block-tridiagonal fashion similarly to the example presented in Appendix A but accounting for the varying block sizes too.

This approximation neglects possible surface effects caused by dangling bonds at the beginning and at the end of the overlap region, but accurately captures the physics of both the interface and the monolayers. To justify our approach the band structures of the isolated single-layer materials obtained from DFT simulations are compared to the source- and drain band structures calculated from the device Hamiltonian. The results are presented in Fig. B.3. It can be seen that the band edges that play the most important role in the electron transport are accurately reproduced. Remarkable deviations from the DFT band structure appear only at the satellite valleys.
B.2. Metal-semiconductor interfaces

The method described above only works if the coupling between the different materials is weak enough such that the appropriate subspaces of the heterostructure’s Hamiltonian approximate well the states of the isolated components. This is not the case when metals are connected to semiconductors due to the charge transfer at the interface that can significantly modify the properties of the semiconductor in the vicinity of the metal. To simulate transport through this heterojunction a larger system has to be treated at the DFT level that includes a sizeable volume around the interface in both directions.

The procedure for metal-semiconductor contacts starts again with the modeling of the interfaced materials in a common unit cell, similarly to the van der Waals heterojunctions. In Chapter 6 a scandium top contact is simulated on single-layer MoS₂. Scandium crystallizes...
APPENDIX B. MODELING OF INTERFACES

Figure B.4: (a) Side view of the atoms in the MoS$_2$-Sc heterostructure’s primitive unit cell. Cyan dots represent Mo atoms, yellow S, and gray Sc. (b) The area in the center with the white background represents the unit cell of MoS$_2$ with a scandium top contact that has to be simulated at the DFT level. Shaded regions refer to periodic replicas.

In a simple hexagonal structure with an in-plane lattice parameter $a_0 = 3.309$ Å, that is only 4% larger than in MoS$_2$ ($a_0 = 3.183$ Å). In order to preserve the characteristics of the MoS$_2$ band structure, the size of the common unit cell is fixed to the equilibrium MoS$_2$ lattice constant so that 4% compressive strain is applied to Sc. It has been demonstrated before that 6 atomic layers of a metal top contact reasonably well describes the bulk-like situation and the results do not remarkably change beyond this thickness [141]. Therefore, we also model the heterostructure by placing only 6 scandium layers above MoS$_2$. The unit cell of the compound system is presented in Fig. B.4 (a). The distance between the S and Sc atoms is determined by relaxing the ions other than the top four layers of the metal within this fixed unit cell. A 10 Å vacuum region separates the periodic replicas from each other in the vertical $y$ direction. Van der Waals interactions are included through the DFT-D2 method of Grimme [103].

Using this optimized geometry a rectangular supercell is created that contains many replicas of the complete primitive unit cell along the transport direction $x$ and also a region where the metal atoms are
Figure B.5: Extraction of the Hamiltonian blocks utilized for constructing the Hamiltonian matrix of a device with realistic physical dimensions. Blue rectangles identify the regions from where interactions between cells containing only the metal atoms are taken, red the complete unit cells with the metal and semiconductor parts overlapping, and orange the semiconductor-only part. The green rectangle marks the central part that has to be connected to the left and right extensions.

removed so that only the semiconductor layer is present. This cell is displayed in Fig. B.4 (b). The electronic structure of the system is computed with VASP. The MLWF Hamiltonian matrix elements are generated using Wannier90. The idea behind this procedure is that these matrix elements already contain all the information that is required to construct the Hamiltonian of a system where the metal and semiconductor parts extend to realistic scales.

The method is illustrated in Fig. B.5. A region in the middle of the part where the metal and the semiconductor components overlap is identified. It must be large enough such that it can be divided into three equal volumes whose interactions beyond their nearest neighbors can be neglected. Their volumes are marked with red in Fig. B.5. The Hamiltonian blocks that contain the interactions between the central cell and its left neighbor, itself, and right neighbor are denoted $H_{O,1}$, $H_{O,2}$, and $H_{O,3}$, respectively. The same process is repeated in the semiconductor-only region, where the cells are colored in orange and the Hamiltonian blocks are denoted $H_{S,1}$, $H_{S,2}$, and $H_{S,3}$. It is straightforward to realize that similarly to the example of Appendix A, the $H_{O,i}$ or $H_{S,i}$ blocks $(i = 1, 2, 3)$ placed in a tridiagonal fashion
Figure B.6: Schematics of a scandium top contact deposited on a MoS$_2$ strip with extended lateral dimensions. Typically, the lengths $L_M$, $L_O$, and $L_S$ measure about 10 nm, 50 nm, and 50 nm, respectively.

would give the $H_0$ Hamiltonian of an extended strip with the metal overlapping the semiconductor or with the MoS$_2$ only, respectively. The $H_-$ and $H_+$ matrices that connect the strip with its replicas in the $-z$ and $+z$ directions are created in the same way from blocks that connect a central cell with its neighbors that are shifted along the $z$ axis.

The overlapping part on the left and the semiconductor-only part on the right are connected through the $H_C$ block of the central region marked as green. The off-diagonal block that connects the left extension to the center is the same $H_{O,3}$ matrix that connects the overlapping cells with their right neighbors. Similarly, the connection between the central part and the semiconductor-only extension on the right is $H_{S,1}$.

Additionally, another set of three volumes are identified within the overlapping region, namely the metal-only parts, shown in blue. The corresponding Hamiltonian blocks are denoted $H_{M,1}$, $H_{M,2}$, and $H_{M,3}$. The connection between the rightmost metal-only block and the start of the overlapping part will be denoted as $H_{MO}$. They are all subspaces of the $H_{O,i}$ ($i = 1, 2, 3$) matrices. These components are required to build a metal extension without the underlying semiconductor layer on the left side of the simulation domain, as illustrated in Fig. B.6. The physical justification behind this procedure is that although the semiconductor states are heavily influenced by the presence of a metal contact, states in the metal remain more intact due to the higher charge density there. Thus, the Hamiltonian of a metal extension can be constructed from the matrix elements of the overlapping metal and semiconductor region, while the same could not be done for the semiconductor extension. This is the reason for
Figure B.7: Schematics of the $H_0$ Hamiltonian of the device presented in Fig. B.6. The colors and the notations of the sub-matrices refer to the quantities introduced in Fig. B.5 including a large volume of a semiconductor-only part in the DFT simulation.

Taking all the above into consideration, the Hamiltonian of the system presented in Fig. B.6 is sketched in Fig. B.7 with the help of the quantities introduced in Fig. B.5. The Hamiltonian blocks containing the interactions between the metal-only parts (blue) are repeatedly placed in a tridiagonal fashion first. They are then connected...
through the $H_{\text{MO}}$ block (black) to the region where the metal and
the semiconductor layer overlap (red). The latter part is connected
to the semiconductor-only volume (orange) through the central green
block.

Special attention must be paid to the indexing of the Wannier
functions. In order to make the proposed approach properly work, the
$i$-th MLWF in any of the distinguished cells must describe approxi-
mately the same state as the $i$-th MLWF in its neighbor cells. That
is, the atoms they are centered on should be replicas of each other by
a translation vector with the same length as the cell width and their
symmetry should be the same as well. To satisfy this requirement the
atoms are indexed in ascending order by their $x$ coordinate first and by
$y$ and $z$ in case of ambiguity. In general this is not sufficient yet when
more than one MLWF is centered on the same atom. The different
orbitals could be arbitrarily ordered at various places. However, if
the initial projections are carefully chosen so that they transform the
Bloch space into a state that is already close to the maximally localized
one, they preserve their initial symmetry. If e.g. the three MLWFs
centered on a S atom are of $p_x$, $p_y$, and $p_z$ symmetry, respectively,
then the same order can be expected at every other atom position
when the projections are defined as $S:1=1$.

Due to the minor boundary effects even at the middle of the
supposedly homogeneous regions the matrix connecting a central cell
to its left neighbor is not perfectly identical to the conjugate trans-
pose of the one connecting it to the right neighbor. The device
Hamiltonians are therefore symmetrized as $H_0^{\text{sym}} = (H_0 + H_0^\dagger)/2$, $H_-^{\text{sym}} = (H_- + H_+^\dagger)/2$, and $H_+^{\text{sym}} = (H_+ + H_-^\dagger)/2$.

To validate the approach presented in this work the band structure
of an isolated single-layer MoS$_2$ is compared to the one obtained from
the rightmost blocks of the Hamiltonian created according to Fig. B.7
The former one is calculated in the same rectangular unit cell as the
one colored in orange in Fig. B.5 through the Wannier interpolation
technique, utilizing the MLWF Hamiltonian matrix elements obtained
in the primitive hexagonal unit cell. The results are presented in
Fig. B.8. It can be seen that the band gap of MoS$_2$ is accurately
reproduced, while the energy of satellite valleys deviate from the
original one. This is due to the influence of the metal contact on the
semiconductor states that is still noticeable a few nanometers away
Figure B.8: Band structure of single-layer MoS$_2$ in a rectangular unit cell, as the orange one in Fig. B.5, extracted from the DFT simulation of a periodic system (lines) and from the rightmost blocks of the Sc-MoS$_2$ heterostructure’s Hamiltonian matrix (circles). The path is taken along the $k_z = 0$ line with the Brillouin zone extending from $-\pi$ to $\pi$. Note that the $K$ point in the primitive hexagonal unit cell folds to the $\Gamma$ one here.

from the interface. A longer MoS$_2$ part in the DFT simulation might improve the results, but the Wannier transformation of electron states in larger systems is computationally very demanding.
Upon further investigations it was found that the approximations applied in Chapters 3 and 4 might lead to some undesired effects. The most critical approximation was the neglection of the derivatives of the Hamiltonian matrix elements beyond first nearest-neighbor distances in Eq. (2.64). Therefore, we now alleviate this restriction. The current flowing through a MoS$_2$ or black phosphorus sample with a small bias difference at the source and drain terminals is computed with a varying number of $\nabla_i H_{nl}$ terms retained in the calculation of the scattering self-energy matrices. It has been determined that the inclusion of terms up to third nearest-neighbor distances is essential for an accurate description of the considered systems, while connections beyond this range do not significantly affect the results.

As already discussed in Chapter 3, there was another weak point in the analysis of few-layer MoS$_2$ FETs, namely the modeling of the atomic structure using the experimental bulk lattice parameters.
This simplification results in a lower energy separation between the first two conduction sub-bands compared to the predictions of DFT simulations with ionic relaxation. We originally aimed for an accurate reproduction of the experimental band gap value, which is obtained without ionic relaxation, while in fact this feature has less influence on the device characteristics than the conduction band properties such as effective masses and inter-valley separations. Although there is no direct experimental evidence for it, DFT with GGA or hybrid functionals as well as GW calculations all indicate a higher minimum for the second valley when the ions are relaxed within the unit cell. Thus, we recalculated the MLWF Hamiltonian of the single-layer MoS$_2$ after fully relaxing the lattice and the ions using the PBE (GGA) functional again.

The DFT-optimized lattice parameter of the monolayer MoS$_2$ is $a_0 = 3.183$ Å. The energy separation between the satellite Σ and the $K$ valleys is 0.27 eV. This is a spectacular increase as compared to the 0.04 eV that the unrelaxed DFT simulation predicted. This higher energy separation is expected to reduce the overall scattering rate and increase the carrier mobilities, while the inclusion of more $\nabla_iH_{nl}$ terms in Eq. 2.64 has the opposite effect. Recomputing the $I_d-V_g$ curves of the investigated devices in this more precise approach is out of the scope of this work due to the high computational cost of the self-consistent simulations. We did, however, perform a less demanding electron- and hole mobility calculation of single-layer MoS$_2$ to estimate the magnitude of the changes.

The carrier mobilities as a function of the charge concentration are presented in Fig. C.1. The electron mobility at low carrier concentrations reaches a value as high as 495 cm$^2$/V s, which is closer to the estimations of other theoretical studies than our previous results. It decreases to about 400 cm$^2$/V s at a charge density of $2 \times 10^{13}$ cm$^{-2}$. The overall higher mobility values than what was presented in Chapter 3 suggest that the approximations applied there led to an overestimation of the electron-phonon scattering rates and hence an underestimation of the current. The ON-current of the MoS$_2$ FET is, therefore, likely higher than what we determined in Chapter 3.

Figure C.1 also shows the hole mobility of MoS$_2$. It is significantly lower than the electron one. It varies from 75 cm$^2$/V s at a hole concentration of $10^{11}$ cm$^{-2}$ to 67 cm$^2$/V s at $2 \times 10^{13}$ cm$^{-2}$. This high
Figure C.1: Phonon-limited hole (left) and electron (right) mobilities in single-layer MoS$_2$ as a function of the carrier concentrations. The ions are fully relaxed during the DFT simulations and the derivatives of the Hamiltonian matrix elements are retained up to third nearest-neighbor distances when evaluating the scattering self-energies.

The asymmetry between the electron and hole mobilities is attributed to the small energy difference between the $K$ and $\Gamma$ valleys in the valence band. Although the band maximum is at the $K$ point with a hole effective mass $m^*_{h,K} = 0.6 \ m_0$, the $\Gamma$ valley is situated only 0.015 eV lower with an effective mass $m^*_{h,\Gamma} \approx 2.0 \ m_0$. An increased inter-valley scattering rate and a reduced average hole velocity results from this band configuration.

The single-layer black phosphorus (BP) is further investigated for similar reasons as MoS$_2$. The inclusion of the derivatives of the Hamiltonian matrix elements beyond first nearest-neighbor distances is expected to have a significant impact on the electron-phonon interactions in this material too. Although with BP we had no issue similar to the valley separation uncertainty in MoS$_2$, we can still increase the accuracy of the simulations in other ways. One improvement we make is that the derivatives of the Hamiltonian along the three different
Cartesian directions are extracted separately from DFT calculations with strain in the $x$, $y$, and $z$ direction instead of the hydrostatic strain applied in the previous approach. The other one is the use of the HSE06 hybrid functional instead of the generalized gradient approximation. The former one has been demonstrated to reproduce the experimental lattice parameters and band gap of bulk BP more accurately than the PBE functional [43]. Therefore, it is presumed to be more reliable in determining the properties of single-layer BP too. It also predicts a significantly higher band gap $E_g = 1.5$ eV for the monolayer than the 0.9 eV PBE suggests. We performed a full lattice- and ion relaxation with HSE06 before calculating the MLWF Hamiltonian. Then we determined the carrier mobilities in this more accurate setting.

The results are shown in Fig. C.2. We observe a huge drop in both the electron- and hole mobilities along the $X$ (armchair) and $Y$ (zig-zag) directions as compared to the values obtained by the first nearest-neighbor approximation and with the PBE functional in
The maximum of the hole mobility along $X$ is reduced from 3000 down to 650 cm$^2$/Vs and the electron mobility from 2500 down to 300 cm$^2$/Vs. Despite this reduction, the qualitative behavior of the curves remained similar to our former predictions. Switching the direction of transport resulted in an 18-fold decrease in the hole mobility in the previous case, while it is still around 12-fold now. The former 4-fold reduction of the electron mobilities under the same conditions is closer to 5 now. The electron mobility along $X$ is again lower than the hole one, and when considering transport along $Y$, it is still higher than the value for holes. The increase of the hole mobility along $Y$ with the increasing charge remained as well, due to the non-parabolicity of the valence band along $\Gamma$-$Y$.

These changes are expected to decrease the ON-current of long-channel BP transistors, but are not likely to alter the conclusions of Chapter [4]. As the gate length shrinks below 10 nm, the mobility plays a decreasing role in the performance of the device. Instead, it is the source-to-drain tunneling rate that affects the ON/OFF current ratio of ultra-scales FETs and this effect is related to the effective mass rather than the mobility.
List of publications

Peer-reviewed journal papers


Conferences

• A. Szabo and M. Luisier, “Full-band simulation of p-type ultrascaled silicon nanowire transistors”, *European Solid-State Device Research Conference (ESSDERC)*, p. 77, Bucharest, Romania, September 2013

• A. Szabo, S. J. Koester, and M. Luisier, “Metal-dichalcogenide hetero-TFETs: Are they a viable option for low power electronics?”, *Device Research Conference (DRC)*, p. 19, Santa Barbara, CA, USA, June 2014


• C. Stieger, A. Szabo, P. Khomyakov, and M. Luisier, “Multiscale modeling of electron transport in semiconductor nanosstructures”, Materials Science Engineering (MSE), Darmstadt, Germany, September 2014
Bibliography


[52] Y. Guo, Y. Han, J. Li, A. Xiang, X. Wei, S. Gao, and Q. Chen, “Study on the resistance distribution at the contact between


Curriculum vitae

Áron Szabó was born in Budapest, Hungary, on April 16, 1987. From 2005 to 2011 he studied Engineering Physics at the Budapest University of Technology and Economics. He received his MSc. degree in 2011. In the same year he joined the Integrated Systems Laboratory, ETHZ as a research assistant in the Technology CAD group. His research interests focus on the numerical simulations of quantum transport in nanoscale devices.