Compact Modeling of Ballistic Monolayer Black Phosphorus Metal-Oxide-Semiconductor Field-Effect Transistors

Raphaël Joachim Prentki

Department of Physics, McGill University

Montreal, Quebec, Canada

August 2017

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science.

This thesis is dedicated to my grandfather, Jacques Prentki.

Contents

Abstract

Compact models are tools used by the semiconductor industry for digital prototyping of integrated circuits. They are analytical parametrizations of the current-voltage characteristics of nanoelectronic devices in terms of hundreds or thousands of empirical parameters. In this thesis, we present accurate compact models of ballistic metal–oxide–semiconductor field-effect transistors using less than ten parameters, all of which have a clear physical interpretation. In addition to having great predictive power, the models that we present are important conceptual guides for device research and development. We focus our attention on transistors composed of monolayer black phosphorus, a two-dimensional semiconductor with unique electronic and mechanical properties which make it a promising candidate for novel digital logic applications.

Résumé

Les modèles compacts sont des outils utilisés par l'industrie des semi-conducteurs pour le prototypage numérique de circuits intégrés. Ces modèles décrivent, en utilisant des centaines ou des milliers de paramètres empiriques, les caractéristiques des composants nanoélectroniques. Dans cette thèse, nous présentons des modèles compacts précis de transistors à effet de champ à grille métal-oxyde balistiques utilisant moins de dix paramètres, qui ont tous une interprétation physique claire. Ainsi, en plus d'avoir une grande puissance prédictive, les modèles que nous présentons sont des guides conceptuels importants pour la recherche et le développement dans le domaine de la nanoélectronique. Nous concentrons nos efforts sur des transistors composés de phosphore noire monocouche, un semi-conducteur bidimensionnel aux propriétés électroniques et mécaniques uniques et prometteuses pour des applications technologiques novatrices.

Acknowledgements

First and foremost, I would like to thank Hong Guo for his exceptional altruism in supervising my research project, as well as for his insightful guidance in physics, nanotechnologies, and academic research in general.

I thank Fei Liu for sharing with me his computer program for simulations of monolayer black phosphorus metal-oxide-semiconductor field-effect transistors, and for teaching me how to properly use it.

I would also like to thank the numerous inspiring professors who taught my graduate-level classes and therefore played an foundational role in my understanding of quantum mechanics, condensed matter physics, many-body theory, and nanoelectronics. The names of Guy Moore, Aashish Clerk, William Coish, and Thomas Szkopek spring to mind.

These acknowledgements would not be complete without a mention to my peers, with whom I have had countless debates and edifying discussions. In particular, I wish to thank Daniel Abarbanel, Simon Bilodeau, Felix Fehse, Mohammed Harb, Chen Hu, Paul Jreidini, Mikhail Mamaev, Alexander McDonald, Raj Shampur, Qing Shi, and last but not least, Chenyi Zhou.

I thank the McGill Physics Department for funding part of my stipend through their departmental award. I also thank members of the Physics Department staff for their assistance and kindness.

Finally, I thank Compute Canada and Calcul Québec for providing, through the Guillimin supercomputer, the computational resources needed to generate *in silico* some of the data that I present in this thesis.

Preface and Statement of Originality

The manuscript for this thesis was entirely written by myself, Raphaël Prentki. Additionally, all figures presented in this thesis were designed by myself.

In this thesis, I explore the idea of compact modeling of nanoelectronic devices from robust physical arguments and first-principles simulations, with a focus of ballistic monolayer black phosphorus (ML-BP) metal-oxide-semiconductor field-effect transistors (MOSFETs). Some of the original ideas that I present in this thesis are:

- An approximation formula for the capacitance of a parallel plate capacitor with rectangular plates exact within $\sim 1\%$ for all values of the ratio of plate separation to plate width, as exposed in Sec. 3.1.1.7,
- A threshold voltage extraction method based on the concept of ballistic mobility for ballistic MOSFETs biased at drain voltages smaller than the thermal voltage, as exposed in Sec. 3.2.2.1,
- The demonstration that the transfer and output characteristics of ballistic ML-BP MOSFETs can be described accurately within the capacitor model, as exposed in Sec. 4.2,
- The demonstration that the output characteristics of ballistic ML-BP MOSFETs can be described accurately within the virtual source model, as exposed in Sec. 4.3
- An accurate semiphenomenological parametrization of the velocity of charge carriers in the saturation regime in terms of the gate voltage and the threshold voltage, as exposed in Sec. 4.3.

Lists of Abbreviations, Symbols, and Notations

List of commonly used abbreviations

Eq.	Equation
Fig.	Figure
Tab.	Table
Sec.	Section
Ref.	Reference
2D	Two dimensions, or two-dimensional
2DEG	Two-dimensional electron gas
FET	Field-effect transistor
MOS	Metal-oxide-semiconductor
MOSFET	Metal-oxide-semiconductor field-effect transistor
DFT	Density functional theory
NEGF	Nonequilibrium Green's functions (formalism)
TB	Tight binding
LB	Landauer-Büttiker
VS	Virtual source
ML	Monolayer
BP	Black phosphorus
AD	Armchair direction
ZD	Zigzag direction
TMDC	Transition metal dichalcogenide

List of commonly used symbols

e	Positron charge
ε_0	Vacuum permittivity
h	Planck's constant
$\hbar$	Reduced Planck's constant
k_B	Boltzmann constant
α	Fine structure constant
T	Temperature
$\overline{m^\star}$	Density of states effective mass
$m_t^\star$	Effective mass along the transport direction
E_C	Conduction band minimum, or first conduction subband minimum
g_V	Valley degeneracy
E_F	Fermi level
μ_{ball}	Ballistic mobility
v_T	Nondegenerate unidirectional thermal velocity
I_{DS}	Drain current
V_{DS}	Drain voltage
V_{GS}	Gate voltage
V_T	Threshold voltage
δ	Drain-induced barrier lowering parameter
C_G	Gate capacitance
C_E	Electrostatic capacitance
C_Q	Quantum capacitance
S	Subthreshold swing
α_G	Gate control parameter
α_D	Drain control parameter

List of commonly used mathematical notations

$\log$	Natural logarithm
$\log_a$	Base a logarithm, where $a > 1$
$F(\zeta; k) = F(\phi, k)$	Incomplete elliptic integral of the first kind
$F'(\zeta; k) = F'(\phi, k)$	Complementary incomplete elliptic integral of the first kind
$K(k)$	Complete elliptic integral of the first kind
$K'(k)$	Complementary complete elliptic integral of the first kind
$E(\zeta; k) = E(\phi, k)$	Incomplete elliptic integral of the second kind
$E'(\zeta; k) = E'(\phi, k)$	Complementary incomplete elliptic integral of the second kind
$E(k)$	Complete elliptic integral of the second kind
$E'(k)$	Complementary complete elliptic integral of the second kind

1 Introduction

Transistors and other semiconductor devices have undergone an uninterrupted miniaturization over the past fifty years, and are now reaching the nanoscale. This represents a fundamental barrier in the understanding of electronic devices: at the atomic scale, it is quantum phenomena that dominate the physical properties of these systems. To understand the experimental data and the physics of these nanoelectronic devices, it is important to develop an appropriate theoretical formalism and associated modeling tools which are capable of making quantitative and material specific predictions of device characteristics from first-principles (or "*ab initio*"). One of the most successful formalisms for atomistic modeling of quantum transport is density functional theory (DFT) within the nonequilibrium Green's function (NEGF) formalism. Since its conception by Guo *et al.* [1], the NEGF-DFT technique has emerged as a very powerful and practical method for predicting nonlinear and nonequilibrium quantum transport properties of nanoelectronic devices.

A profusion of semiconductor devices is found in contemporary technologies, ranging from microprocessors to photovoltaic cells. Numerical simulations most often guide the design and control of such devices. *Ab initio* physics-based simulations of nanoelectronic devices take exceedingly large amounts of time to run, and are thus mostly impractical for the semiconductor industry. Instead, this industry uses compact models, which typically parametrize electronic devices by the hundreds or even the thousands of empirical parameters that are fitted from experimental data. However, these parameters become difficult to measure at the nanoscale. Moreover, curve-fitting-based compact models offer little conceptual insight on the physics of these devices, which is critical to guide our thinking in device research and development.

The virtual source (VS) model, which was recently developed by Antoniadis *et al.* [2–4], offers hope in this regard. In this model, nanoelectronic devices are parameterized by around ten parameters, all of which have a clear physical interpretation, such as the charge carrier's effective mass, mean free path, and effective mobility. By measuring these parameters

experimentally, the current-voltage characteristics of metal-oxide-semiconductor nanoscale field-effect transistors (MOSFETs) can be modeled with surprising accuracy within the VS model [2, 3].

My Masters research project consists in computing the parameters of the VS model for various MOSFETs from first principles, as well as investigating first-principles compact modeling more generally. I chose monolayer black phosphorus (ML-BP) MOSFETs as the object of my investigation. Ever since few-layer BP was first obtained by scotch tape-based mechanical exfoliation in 2014 [5], ML-BP has emerged as a promising material for nanoelectronics applications [6, 7]. Indeed, ML-BP possesses critical advantages over other two-dimensional layered materials that have been studied recently. First, unlike graphene, ML-BP has a bandgap, making it suitable for transistors applications. Second, while they do possess an appropriate bandgap, being compounds, monolayer transition metal dichalcogenides are harder to fabricate with high purity than ML-BP; high purity is a requirement to achieve an optimal charge carrier mobility [8]. Overall, my Masters thesis project fits within the research and development of sub-10 nm technology, which as of yet has not been made commercially available.

This thesis is organized as follows. First, in the remainder of Sec. 1, I further my introduction of compact modeling, first-principles simulations, and two-dimensional semiconductors. Second, in Sec. 2, I develop a brief and general tutorial exposition to the operating principles of MOSFETs in terms of energy band diagrams, followed by a more thorough description of charge transport in MOSFETs from the diffusive (long-channel) to the ballistic (short-channel) limits. Additionally, I introduce the concept of ballistic mobility, which describes ballistic MOSFETs in the language of diffusive transport, and I state the regime of validity of this description. Third, in Sec. 3, I expose various calculation methods of four crucial parameters for MOSFET modeling: the gate capacitance, the threshold voltage, the subthreshold swing, and the drain-induced barrier lowering. These calculation methods are either based on purely analytical arguments, or on extraction from current-voltage characteristics. Fourth, in Sec. 4, I compare compact models based upon the ideas developed in the two previous sections to current-voltage characteristics of ballistic ML-BP MOSFETs

computed from first-principles simulations. Finally, in Sec. 5, I conclude by summarizing this thesis and outlining a future project.

1.1 Compact modeling

A compact model is an analytical parametrization of the current through an electronic device as a function of its input voltages. Compact models are used in circuit simulators for the purpose of predicting the integrity and behaviour of large-scale circuits, in which a great number of devices are interconnected. Compact models must therefore have the two following characteristics:

- They must be sufficiently simple and stable for a large-scale circuit simulator to run quickly and robustly.
- They must be sufficiently accurate to avoid any appreciable compounding of errors on such a simulator.

The simplest example of a compact model, familiar even to most high school students, is Ohm's law, which states that the current I through a resistor is proportional to the voltage V applied across its electrodes:

$$I = \frac{1}{R}V \tag{1}$$

The coefficient of proportionality $\frac{1}{R}$ is known as the conductance of the resistor. As expressed in Eq. 1, the compact model parameter $\frac{1}{R}$ can be thought of as being entirely phenomenological. It can be obtained from linear fitting on the current-voltage characteristic of a resistor, without any underlying understanding of the scattering processes ultimately leading to the accuracy of Eq. 1. Alternatively, $\frac{1}{R}$ may be defined and characterized more formally within a physical theory. For example, semiclassically, the Drude model [9, 10] relates $\frac{1}{R}$ to material-specific parameters and physical quantities pertaining to the scattering of electrons. Eq. 1 may also be derived from a microscopic approach based on dressed Green's functions and vertex corrections [11, 12].

Thus, we see that compact models may be either constructed from a theory based on firm physics-based arguments, or from a purely *ad hoc* approach where the physical phenomena dictating charge transport in the device are not considered. The former is often referred as a bottom-up approach, while the latter is referred as a top-down approach. Top-down approaches to compact modeling focus on intended applications rather than charge transport properties of modeled device. As such, models based on this approach are most often difficult to generalize. For example, a top-down model developed for a Si-channel SiO_2-oxide MOSFET may not be easily adapted to model a MOSFET of identical geometry but built with different materials. Bottom-up compact models, in which all parameters have a precise physical interpretation, do not suffer from this flaw. Furthermore, a particularly desirable property of a compact model is scalability, namely the ability to properly model devices of various dimensions. Top-down models may or may not be scalable. However, bottom-up compact models all have this property, at least to a certain extent. This makes bottom-up compact models important guides for device research and development in a field governed by Moore's law.

Compact modeling has become increasingly important for the semiconductor industry in the past decades [13]. Indeed, transistors are becoming smaller in size, and more numerous on integrated circuits. As such, manufacturing circuit prototypes using state-of-the-art techniques such as photolithography has become prohibitively expensive. The semiconductor industry has thus turned to compact modeling for the purpose of research and development of complicated integrated circuits. Circuits are simulated and adjusted until the desired operation is reached. Only then is the circuit prototype first manufactured. We note that the most commonly used circuit simulation software is SPICE (Simulation Program with Integrated Circuit Emphasis) [14].

The SPICE Level 1 model [15] was amongst the first compact models for very long channel MOSFETs, and was developed in the early 1970s. This model was derived from a careful examination of drift-diffusion transport and electrostatic effects in MOSFETs [16]. In the SPICE Level 1 model, the drain current of a MOSFET in expressed as a function of

the drain voltage V_{DS} and gate voltage V_{GS} as:

$$I_{DS}\left(V_{DS}, V_{GS}\right) = \begin{cases} 0 & \text{for } V_{GS} \leq V_T \\ \beta\left(V_{GS} - V_T - \frac{V_{DS}}{2}\right) V_{DS} & \text{for } V_{GS} > V_T \text{ and } V_{GS} - V_T \geq V_{DS} \\ \frac{\beta}{2}\left(V_{GS} - V_T\right)^2 \left(1 + \lambda V_{GS}\right) & \text{for } V_{GS} > V_T \text{ and } V_{GS} - V_T < V_{DS} \end{cases} \tag{2}$$

where β, V_T, and λ are the parameters of the model. The SPICE Level 1 model is of great simplicity and based upon robust physics, but suffers from two important flaws:

- The drain current is approximated to be 0 for gate voltages V_{GS} smaller than the threshold voltage V_T. This approximation is incorrect, and corresponds to ignoring an important fraction of leakage currents and power dissipations.
- For $V_{GS} > V_T$, the drain current is approximated to be piecewise linear function of V_{DS}. In reality, the drain current of a MOSFET is a smooth function of V_{DS}. Furthermore, as a result of this approximation, the drain current in the SPICE Level 1 model is systematically overestimated.

We note, however, that while these shortcomings alone would make the SPICE Level 1 model unusable for the development of present-day integrated circuits, they were not considered to be severe in the 1970s.

Over the years, more complicated compact models were developed [17, 18]. Currently, typical MOSFET compact models used by the semiconductor industry include hundreds or thousands of parameters, a significant fraction of which are purely phenomenological in nature, and obtained from fits with experimental data. Synopsys Sentaurus ® is an example of such a modern-day compact model, and requires approximately 750 parameters to simulate a MOSFET [19]. We wish to give a flavour of the type of parameters used in this compact modeling software. The University of Bologna bulk mobility model [20, 21] is used in Synopsys Sentaurus ® to describe the temperature (T) dependence of the contribution of dopant scattering to the carrier mobility. In this model, the dopant-limited mobility is parametrized as:

$$\mu_{\text{dop}}\left(T\right) = \mu_0\left(T\right) + \frac{\mu_L\left(T\right) - \mu_0\left(T\right)}{1 + \left(\frac{N_D}{C_{r1}}\right)^{\alpha} + \left(\frac{N_A}{C_{r2}}\right)^{\beta}} - \frac{\mu_1\left(N_D, N_A, T\right)}{1 + \left(\frac{N_D}{C_{s1}} + \frac{N_A}{C_{s2}}\right)^{-2}} \tag{3}$$

where N_D and N_A are respectively the donor and acceptor dopant concentrations, and where:

$$\begin{aligned} \mu_L(T) &= \mu_{\max}\left(\frac{T}{300\ \mathrm{K}}\right)^{-\gamma+c\left(\frac{T}{300\ \mathrm{K}}\right)} \\ \mu_0(T) &= \frac{\mu_{0d}N_D + \mu_{0a}N_A}{N_D + N_A} \\ \mu_1(T) &= \frac{\mu_{1d}N_D + \mu_{1a}N_A}{N_D + N_A} \end{aligned} \tag{4}$$

The bulk dopant-limited mobility is thus parametrized by 13 (possibly T-dependent) parameters that were fitted from experimental measurements of Hall mobilities [21]. These parameters depend not only on the chemical nature of the host material (typically silicon), but also the nature of the dopant atoms themselves. Hall mobility measurements and fitting procedures must therefore be repeated for any desired combination of host material and dopant atoms.

Moore's law is a prime example of an epiphenomenon, namely a mathematically described trend emerging from a swarm of events characterized by irregularity and sporadicity. Those events are the small and great nanoelectronics breakthroughs driven by the thousands of device engineers and physicists around the world. As transistors are made smaller, the nature of the physical phenomena which most accurately describes charge transport changes. Additionally, the industry requirements for compact models have become increasingly strict [13]. This makes compact modeling an exciting and challenging research field for device physicists and electrical engineers.

1.2 First-principles simulations

As we have seen, compact model parameters are typically obtained through fitting procedures from experimental data taken on the material or electronic device to be modeled. One of the goals of this thesis is to demonstrate the possibility of compact modeling from theoretical arguments and first-principles simulations. Condensed matter systems are all composed of positively charged nuclei and negatively charged electrons interacting electromagnetically.

In the nonrelativistic limit, their properties are best described by their quantum states, which evolve according to the Schrödinger equation. A first-principles calculation of a condensed matter system consists in describing its properties at the atomic level by its quantum state, and calculating its dynamics by solving the Schrödinger equation. These calculations are seldom tractable analytically, and thus most often require extensive numerical calculations. In this latter case, we refer to this process as a first-principles simulation. These simulations offer the greatest amount of predictive power available to us within our present theories of physics about condensed matter systems. FETs are no exception to this rule: their behaviour is best predicted *in silico*. To properly simulate a FET, at least three ingredients are required.

First, a formalism to describe the interactions between charges is required. FETs contains free charges - such as electrons and holes - as well as bound charges - such as dopant atoms. The electromagnetic interaction between these charges is most accurately described by quantum electrodynamics (QED). However, due to its simplicity and validity down to the atomic scale, classical electrodynamics is typically used to describe electromagnetic interactions in nanoscale devices. Indeed, the lowest-order QED correction to the Coulomb potential of the electric potential $V(r)$ of an electron, known as the Uehling potential [22], is exponentially suppressed over a lengthscale of the order of the Compton wavelength of the electron $\lambda \sim 10^{-12}$ m:

$$V(r) = -\frac{e}{4\pi\varepsilon_0}\frac{1}{r}\left[1 + \frac{\alpha}{4\pi^2}\frac{e^{-\frac{4\pi r}{\lambda}}}{\left(\frac{2r}{\lambda}\right)^{\frac{3}{2}}} + \cdots\right] \tag{5}$$

where r denotes the distance from the electron. Therefore, in the context of FET simulations, Poisson's equation for the electric potential ϕ is the descriptor of electromagnetic interactions between charges:

$$\nabla(\varepsilon\nabla\phi) = -\rho \tag{6}$$

where ρ and ε respectively denote the charge density and the permittivity. Note that in the most general case, the permittivity $\varepsilon = \varepsilon(\vec{q}, \omega)$ is a function of electronic wavevector $\vec{q}$ and electronic energy $E = \hbar\omega$. Under such circumstances, Poisson's equation is most readily

solved in Fourier space, in which it can be expressed as:

$$q^2 \varepsilon\left(\vec{q}, \omega\right) \phi\left(\vec{q}, \omega\right) = \rho\left(\vec{q}, \omega\right) \tag{7}$$

The real-space potential can then be obtained by means of inverse Fourier transform. The $\vec{q}$ and ω dependences of the permittivity are of prime importance in condensed matter systems; notably, the former describes the screening of electric fields by mobile charges in metals.

Second, a formalism to describe atomistic properties of charge carriers is required. Such a formalism describes the interaction between mobile charges and the crystalline lattice, namely the bandstructure of the material of which the simulated MOSFET is composed. For example, in a semiconductor, the bandgap and carrier effective masses are described by the chosen atomistic formalism. Quantum Monte Carlo [23] and density functional theory (DFT) [24–26] are often used for this purpose. Another such formalism is the tight-binding model, where electronic states are approximated as superpositions of finitely many orbitals centered around each atomic site, which we label by i. In the tight-binding Hamiltonian H, mobile charges are allowed to jump between neighbouring atomic sites i and j; the probability of such events is quantified by the hopping parameters t_{ij}. Note that these hopping parameters typically need to be computed from more elaborate atomistic formalisms, such as density functional theory. The tight-binding Hamiltonian is given by:

$$H = \sum_{\langle i,j \rangle} t_{ij} a_i^\dagger a_j - q \sum_i V_i a_i^\dagger a_i \tag{8}$$

where $\sum_{\langle i,j \rangle}$ denotes a sum over pairs of neighbouring sites, $a_i^\dagger$ and a_i respectively the creation and annihilation operators on site i, and V_i the on-site energy on site i.

Third, a formalism to describe the transport properties of charge carriers under nonequilibrium conditions is required. Biasing the leads of a transistor corresponds to introducing a spatial dependence to the Fermi level of carriers across the device. Different Fermi levels describe different occupation functions, and as a result, charges flow along Fermi level gradients. In particular, electrons flow from regions of high Fermi level to regions of low Fermi level. The Boltzmann transport equations [27] describe this flow of charges semiclassically. On the other hand, the nonequilibrium Green's functions formalism (NEGF) [1,28] describes

the same phenomenon on purely quantum mechanical grounds, and thus incorporates phenomena such as quantum interference and quantum tunnelling. The central mathematical object in NEGF is the Green's function, defined for a device with two leads (which we will refer to as the source and the drain) as a function of carrier energy E as:

$$G(E) = \lim_{\substack{\eta \to 0 \\ \eta > 0}} \left[(E + i\eta) I - H - \Sigma_S(E) - \Sigma_D(E) \right]^{-1} \tag{9}$$

where I denotes the identity matrix, H the Hamiltonian of the device channel (which could be, for example, a tight-binding Hamiltonian as defined in Eq. 8), and Σ_S (Σ_D) the self-energy operator quantifying the interactions between the channel and the source (drain) lead [28]. One then defines the transmission function as:

$$\overline{T}(E) = \mathrm{Tr}\left[\Gamma_S(E) G(E) \Gamma_D(E) G^{\dagger}(E) \right] \tag{10}$$

where $\Gamma_{S,D}(E) = i\left[\Sigma_{S,D}(E) - \Sigma_{S,D}(E)\right]$. The transmission function can be thought as the transmission probability through the channel of a carrier of energy E multiplied by the number [29] of conduction modes at energy E. The current through the device (from the source to the drain) is then obtained from the Landauer-Büttiker formula [12]:

$$I = -\frac{2e}{h} \int_{-\infty}^{+\infty} \overline{T}(E) \left[f_S(E) - f_D(E) \right] dE \tag{11}$$

where $f_S(E)$ and $f_D(E)$ are respectively the source and drain Fermi distributions, and where we have assumed spin degeneracy as well as coherent transport. Note that as expressed in Eq. 10, the transmission function $\overline{T}(E)$ can describe transport under ballistic conditions, as well as transport limited by some forms of impurity scattering and surface scattering (if a self-energy term associated with scattering is included in the definition of the Green's function in Eq. 9). However, Eq. 10 and Eq. 11 cannot describe scattering in general; notably, they are unsuitable to describe systems with phonon scattering and electron-electron scattering. Fortunately, they can be generalized to describe more general charge transport mechanisms [30].

1.3 Two-dimensional semiconductors

Two-dimensional materials exhibit a number of surprising properties. Perhaps most surprising to the theoretical physicist is their very existence. Indeed, according to the Mermin-Wagner theorem [31], there is no stable long-range ordered continuous symmetry state with finite-range interactions in 2D and at non-zero temperatures. The Mermin-Wagner theorem thus forbids the existence of arbitrarily large 2D crystals. Nevertheless, finitely-sized 2D crystals are possible, as long as their dimensions are sufficiently small. It is unlikely that the limitation set by the Mermin-Wagner theorem will ever hinder real-world applications of two-dimensional materials. Indeed, the magnitude of the fluctuations associated with the instability of two-dimensional materials diverges logarithmically with their sizes; the slow nature of this divergence makes it seldom prohibitive. Perhaps even more importantly, 2D crystals are typically supported by substrates that, by their three-dimensional nature, quench the out-of-plane phonons corresponding to the aforementioned fluctuations. As a proof of concept, a 100 m long sheet of graphene was fabricated in 2012 by Kobayashi *et al.* [32].

Two-dimensional materials have been an active research topic since the isolation and characterization of graphene in 2004 [33]. Since then, several two-dimensional materials have been considered for digital logic applications. We consider three categories of such materials:

- Monolayer semimetals: graphene, silicene, and germanene,
- Monolayer transition metal dichalcogenides, such as molybdenum disulfide (MoS_2) and tungsten diselenide (WSe_2),
- Monolayer black phosphorus.

In the case of MOSFET applications, the dimensionality of those materials offers two principal advantages. First, due to the subnanoscopic thickness of two-dimensional semiconductors, the position of the inversion layer in a MOSFET channel composed of a two-dimensional semiconductor will inevitably be extremely close to the gate oxide layer. The capacitive coupling between the gate and the channel will thus be enhanced, leading to a smaller subthreshold swing, and therefore smaller leakage currents and power dissipations.

Second, due to their out-of-plane structural flexibility, two-dimensional semiconductors offer new prospects for bendable electronics.

Composed of carbon atom on a monolayer honeycomb lattice, graphene is a semiconductor with no bandgap [34], namely a semimetal. The corresponding "low effective mass" of electrons and holes in graphene leads to extremely high charge carrier mobilities [35] of the order of 15 000 $\mathrm{cm}^2 \cdot \mathrm{V}^{-1} \cdot \mathrm{s}^{-1}$ [36]. For this reason, graphene has been considered as a promising surrogate to silicon in transistors [37]. However, the lack of bandgap of graphene leads to transistors with rather poor ON-OFF ratios [38]. This significantly decreases the power efficiency of graphene-based transistors, thus hindering the adoption of graphene towards digital logic applications. A solution to this problem would be create a bandgap in graphene [39], but such procedures typically decrease the charge carrier mobility. Other group-IV semiconductor monolayers, silicene [40] and germanene [41], have similar electronic properties. Importantly, we note that the bandgap of silicene is tunable through appropriate doping [42].

Transition metal dichalcogenide (TMDC) monolayers are atomically thin materials with chemical formula MX_2, where M is a transition metal, and X is a chalcogen. TMDC monolayer are direct bandgap semiconductors [43], making them appropriate candidates for transistor applications [44]. Notably, a MoS_2-based MOSFET with a 1 nm channel length was fabricated in 2016 [45]. Despite these exciting prospects, TMDCs suffer from low charge carrier mobilities, typically ranging from $\sim 10\ \mathrm{cm}^2 \cdot \mathrm{V}^{-1}\mathrm{s}^{-1}$ to $\sim 200\ \mathrm{cm}^2 \cdot \mathrm{V}^{-1}\mathrm{s}^{-1}$ [46–48].

Monolayer black phosphorus (ML-BP), or phosphorene, is a two-dimensional allotrope of phosphorus. Few-layers BP, which is composed of a small number of phosphorene layers bound by Van der Waals forces, was first isolated in 2014 by means of mechanical exfoliation from bulk black phosphorus [5,49]. Due to its large bandgap of ~ 1.8 eV [50,51], ML-BP is a strong competitor to graphene for FET applications. Charge carrier mobility in few-layers BP FETs were measured to be of the order of $10^2\ \mathrm{cm}^2 \cdot \mathrm{V}^{-1}\mathrm{s}^{-1}$ to $10^3\ \mathrm{cm}^2 \cdot \mathrm{V}^{-1}\mathrm{s}^{-1}$ [52–55]. Further theoretical investigations produced similar results, together with similarly high charge carrier mobility in ML-BP [8, 56]. ML-BP is thus seen to be a two-dimensional material solving

the shortcomings of graphene and TMDC monolayers for FET applications: it possesses a bandgap, and has large charge carrier mobility [57]. Furthermore, ML-BP is anisotropic; the two principal axes of the ML-BP crystal structure are known as the armchair direction (AD) and the zigzag direction (ZD). As such, the charge carriers effective masses (relative to the free electron mass) are direction-dependent, varying from ~ 0.17 (AD) to ~ 0.87 (ZD) in the case of electrons, and from ~ 0.19 (AD) to ~ 1.17 (ZD) in the case of holes [58]. This effective mass anisotropy has been shown to provide means towards low OFF-state currents and low switching times for ML-BP MOSFETs [59]. Few-layers BP and ML-BP were also shown to have good mechanical flexibility [60, 61], making it a good candidate for bendable electronics. We thus believe ML-BP to be a very promising two-dimensional material for transistor and digital logic applications [6, 7, 62].

Parenthetically, we note that the saga of few-layers BP is one where the relevance of first-principles simulations is particularly outstanding. Indeed, in practice, few-layers BP is difficult to isolate and manufacture into nanoelectronic devices, as it is particularly prone to photooxidation [63]. Few properties of this novel material could be probed experimentally after it was first isolated, especially in the monolayer case. The quick response of computational physicists [56] bridged this gap in our knowledge; to this day, three years later, much of what we know about few-layer BP still comes from first-principles simulations.

1.4 Goals

In this thesis, we aim to build a bottom-up compact model for nanoscale ML-BP MOSFETs. The "empirical" data to which we will compare our model was obtained from first-principles simulations based on the TB-NEGF formalism [6]. The TB Hamiltonian used in these simulations considers up to third-nearest neighbour hopping; the hopping parameters and on-site energies were computed using DFT and the GW approximation [58, 64].

We focus our attention on four particular ML-BP MOSFETs: one with a 10 nm channel oriented along AD, one with a 10 nm channel oriented along ZD, one with a 5 nm channel

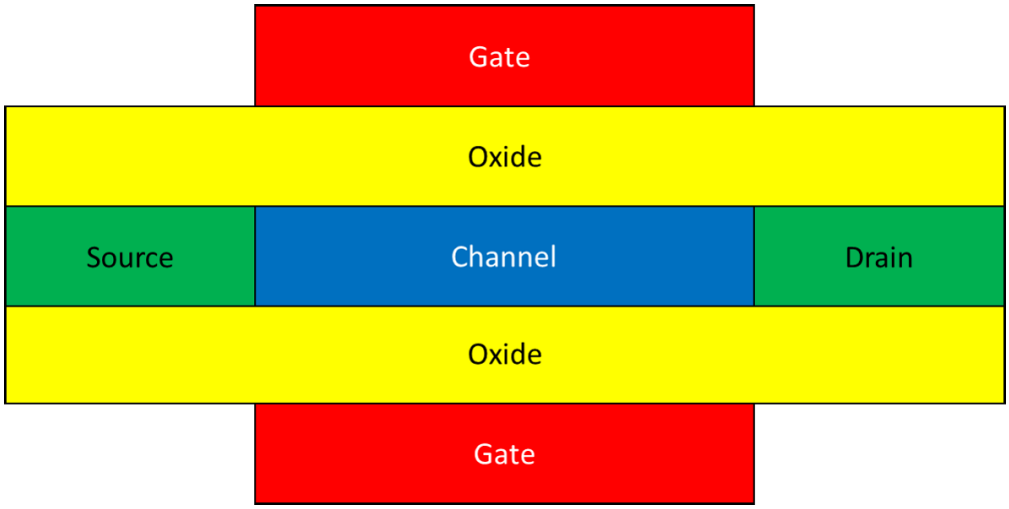


Figure 1 – Device structure of the MOSFET under consideration throughout this thesis.

oriented along AD, and finally, one with a 5 nm channel oriented along ZD. The source and drain of these devices is taken to be ML-BP oriented along the same direction as the channel. The channel is undoped, while the source and drain are n-doped at a concentration of 7.0×10^{13} cm^{-2}. Furthermore, the devices are taken to be double-gated, with gate lengths identical to the channel length. The oxide layers are 2 nm thick and composed of HfO_2, a dielectric material with relative permittivity $\varepsilon_r \approx 25$. Finally, the device is taken to be periodic along the axis perpendicular to the transport direction in the channel plane. The geometry of these devices is illustrated in Fig. 1.

2 From the Diffusive to the Ballistic Limit

This section starts with a brief overview of the operating principles of MOSFETs, tied with an exposition of some of the standard terminology used in the study of MOSFETs and transistors in general. We then introduce the fundamentals of diffusive transport in MOSFETs, a description valid in the long-channel limit. Two formalisms for ballistic transport are then exposed: one based on the Landauer-Büttiker formalism, and one based on self-consistent MOS electrostatics. We conclude this section by introducing the concept of ballistic mobility, which relates the ballistic transport and diffusive transport formalisms.

2.1 Operating principles of MOSFETs

A transistor is an electronic device composed of a semiconducting channel and three electrodes for connection to an external circuit. Charge transport in the channel occurs between two of these electrodes, usually known as source and drain, or emitter and collector. The current through a transistor increases with the voltage applied between these two electrodes, which is known as the drain voltage (and denoted by V_{DS}) in the context of MOSFETs. The conduction through the channel is modulated by the voltage applied on the third electrode, known as the gate, which is capacitively coupled to the channel. The voltage applied between the gate and the source is known as the gate voltage, and denoted by V_{GS}. In what is known as an enhancement mode n-channel transistor, the conductivity increase as V_{GS} increases. Therefore, in such a transistor, the drain-to-source current I_{DS} is only "high" if both the drain and gate voltages are high. Informally, in terms of digital logic, an enhancement mode transistor can therefore be thought of as an AND gate in which V_{DS} and V_{GS} are the gate input and I_{DS} is the gate output.

As a side note, n-channel transistors have for principal charge carriers electrons, while p-channel transistors have holes as principal charge carriers; conduction through the transistor channel increases with V_{GS} in the n-channel case, and decreases with V_{GS} in the p-channel

case. Furthermore, an enhancement mode transistor is one where the conduction through the channel is minute at $V_{GS} = 0$, while a depletion mode transistor is one where this conduction is significant at $V_{GS} = 0$. For the remainder of this thesis, we will solely consider enhancement mode n-channel transistors.

2.1.1 Field effect

The field effect is the physical mechanism by which the conduction through the channel of a transistor is modulated by the applied gate voltage. Applying a gate voltage to a transistor induces an electric field orthogonal to the channel plane. Given the small free charge carriers densities in semiconductors, the screening of this field within the channel is relatively small, resulting in the penetration of this field into the channel. Electric fields alter the energy states available for charge carriers and therefore their occupancies. This shows how an applied gate voltage affects the conductivity of a transistor channel.

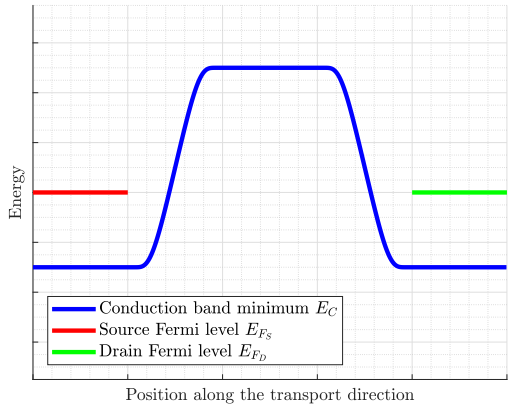
2.1.2 Band diagrams

Band diagrams are useful tools to study charge transport in semiconductor devices. By definition, they are plots of the Fermi level and energy band edges (such as the conduction band minimum and valence band maximum) as a function of some spatial dimension through the structure of interest. It is important to note that the concept of a band diagram is only appropriate in the regime of validity of band structure theory. Thus, for devices with features of very small sizes, in which Bloch's theorem is not applicable, band diagrams are not suitable tools.

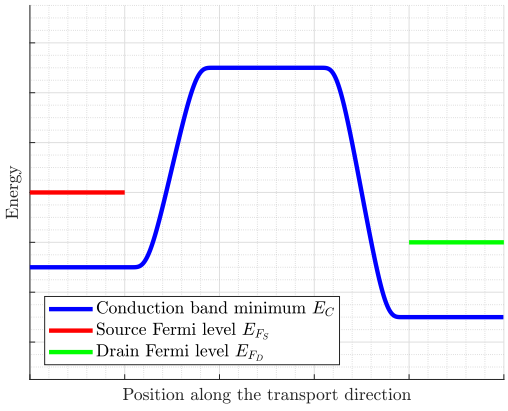
Herbert Kroemer's 2000 Nobel Lecture [65] best epitomizes the importance of band diagrams in semiconductor physics.

Kroemer's Lemma: If, in discussing a semiconductor problem, you cannot draw an energy band diagram, this shows that you don't know what you are talking about.

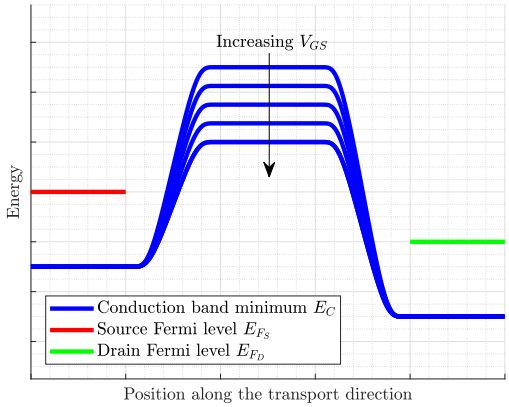
Kroemer's Corollary: If you can draw one, but don't, then your audience won't know what you are talking about.



(a) $V_{DS} = V_{GS} = 0$



(b) $V_{DS} > 0$ and $V_{GS} = 0$



(c) $V_{DS} > 0$ and $V_{GS} > 0$

Figure 2 – Pictorial band diagrams of a MOSFET with degeneratively doped source and drain and intrinsic channel, under equilibrium conditions in (a) and out of equilibrium in (b) and (c).

Following Kroemer's sagacious advices, we show typical band diagrams of a MOSFET in Fig. 2. We assume that the channel of the MOSFET is intrinsic, while the source and drain are degeneratively doped. At equilibrium (namely for $V_{DS} = V_{GS} = 0$), the Fermi level is constant throughout the entire device. The doping profile of this MOSFET imposes the conduction band minimum E_C to be below the Fermi level in the source and drain, and

above the Fermi level in the channel; charge rearrangement effects make the transition of E_C from its source/drain value to its channel value smooth. Thus, the equilibrium band diagram resembles that shown in Fig. 2a. By applying a drain voltage $V_{DS} > 0$, the drain Fermi level is brought down relative to the source Fermi level, resulting in a band diagram akin to that of 2b. Finally, by further applying a gate voltage $V_{GS} > 0$, the value of E_C in the channel is decreased, by virtue of the field effect, as shown in Fig. 2c.

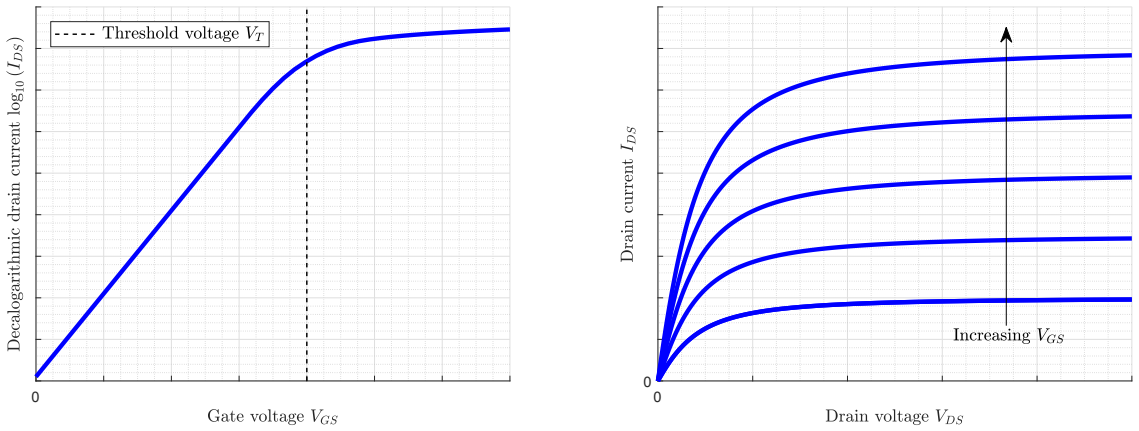
The principal charge transport mechanism in MOSFETs is thermionic emission: electrons in the high-energy thermal tail of the source and drain Fermi distributions may be injected over the potential barrier formed by E_C in the channel, resulting in charge transport through the device. If $V_{DS} = 0$ the injection probabilities of electrons in the drain and source over the barrier are identical: the net drain current is $I_{DS} = 0$. These probabilities are no longer identical for $V_{DS} \neq 0$, resulting in a drain current $I_{DS} \neq 0$. Furthermore, given the exponential nature of the thermal tail of Fermi distributions, the injection probability of electrons over the barrier is significantly increased by lowering the barrier height. As a result, given $V_{DS} > 0$, I_{DS} is significantly increased by increasing V_{GS}. This picture of charge transport in MOSFETs is why they are sometimes referred to as barrier controlled devices.

Finally, we mention two of the many caveats to the discussion above. First, thermionic emission is not the sole charge transport mechanism in MOSFET. Indeed, low-energy electrons can undergo quantum mechanical tunnelling from the source to the drain and *vice versa*. This effect becomes more important as the channel length is decreased. However, even for sub-10 nm MOSFETs, the contribution of the tunnelling current to the total current is very modest, especially for large values of V_{GS} and V_{DS}. It is nevertheless not completely negligible, as the induced leakage currents can increase power consumption. As a side note, tunnelling is not readily described by the Boltzmann transport equations, making NEGF a more appropriate formalism for studying charge transport in nanoscale MOSFETs. Second, we have so far assumed that V_{GS} was the sole voltage controlling the barrier height. However, this is not the case, especially in short-channel MOSFETs. Indeed, when the capacitive coupling between the drain and channel is not negligible, as V_{DS} is increased, the barrier height is lowered. This effect is known as drain-induced barrier lowering (DIBL), and will

be explored in greater depth in Sec. 3.4.

2.1.3 Current-voltage characteristics

As its name suggests, a current-voltage characteristic of a device is a graph of the current through the device as a function of one or several of its input voltages. Beyond the physical mechanism that dictate what this current might be, it is the current-voltage characteristics that ultimately define the nature of a device, and how it can be used for real-world applications. For transistors, we distinguish two types of current-voltage characteristics: the transfer characteristics, and the output characteristics. Transfer characteristics are graphs of I_{DS} as a function of V_{GS} and fixed V_{DS}. On the other hand, output characteristics are graphs of I_{DS} as a function of V_{DS} and fixed V_{GS}. We illustrate these two types of current-voltage characteristics in Fig. 3.



(a) Transfer characteristic at a fixed $V_{DS} > 0$ (b) Output characteristics at fixed $V_{GS} > 0$

Figure 3 – Pictorial representations of the current-voltage characteristics of MOSFETs akin to those assumed in the making of the band diagrams of Fig. 2.

2.1.3.1 Transfer characteristics

Typical transfer characteristics are illustrated in Fig. 3a. As V_{GS} is increased, due to the facilitation of thermionic emission over the barrier, I_{DS} increases. For small values of V_{GS},

the dependence of the barrier height on V_{GS} is linear [66], and hence I_{DS} is an exponential function of V_{GS}. For large values of V_{GS}, as the amount of free charges in the channel is made larger, the electric field induced by the gate is screened. Therefore, the dependence of the barrier height on V_{GS} is logarithmic, and thus, I_{DS} is an algebraic function of V_{GS}.

The value of V_{GS} at the crossover between the exponential and algebraic regimes is known as the threshold voltage, and is denoted by V_T. We thus refer to the regime of exponential dependence of I_{DS} upon $V_{GS} < V_T$ as the subthreshold regime, while we refer to the regime of algebraic dependence of I_{DS} upon $V_{GS} > V_T$ as the superthreshold regime, or the inversion regime. Typically, in the inversion regime, $I_{DS} \propto (V_{GS} - V_T)^m$, where $1 \leq m \leq 2$.

When DIBL is present, transfer characteristics are shifted horizontally to the left as V_{DS} is increased. DIBL can thus be thought of as a modification of the threshold voltage: $V_T \rightarrow V_{T_0} - \delta V_{DS}$, where V_{T_0} is the threshold voltage at $V_{DS} \approx 0$, and δ is the DIBL parameter.

2.1.3.2 Output characteristics

Typical output characteristics are illustrated in Fig. 3b. For small values of $V_{DS} > 0$, as V_{DS} increases, the injection probability of electrons in the drain over the barrier decreases. The overall dependence of I_{DS} on V_{DS} is linear, and for this reason, this regime is called the linear regime, or triode regime. Now, as V_{DS} is made larger and larger, eventually, the injection probability of electrons from the drain is completely suppressed. The drain current then saturates to a constant value. This regime is thus referred to as the saturation regime, or pentode regime. A MOSFET can therefore be thought of as a resistor in the triode regime, and as a constant current source in the pentode regime.

When DIBL is present, I_{DS} is no longer constant in the pentode regime. Instead, the pentode regime current exhibits a positive and roughly constant slope.

2.1.4 Modeling the current-voltage characteristics of MOSFETs

The starting point for any analytical calculation of the current-voltage characteristics of a MOSFET is to parametrize the drain current as the product of charge and carrier velocity:

$$I_{DS} = WQ\left(x\right)v\left(x\right) \tag{12}$$

It is important to properly define the terms in the formula above. First, W represents the width of the channel. The number of conduction modes in a MOSFET channel is proportional to its width [28], which in turns implies that $I_{DS} \propto W$. This relationship only fails for channel with very small widths, in which quantum confinement effects become important. This limit is, however, very rarely reached, even in the smallest modern-day MOSFETs. Second, $Q\left(x\right)$ represents the areal charge density at position x along the transport direction in the channel. Third, $v\left(x\right)$ represents the mean velocity of charge carriers at position x.

The current, as expressed in Eq. 12, does not depend on the parameter x. This fact is a consequence of current continuity [12], an assumption which is exact even when charge carrier recombination or generation effects take place. Depending on the problem at hand, calculations of the drain current might be easier to carry out at a specific position x.

2.2 Diffusive transport

The first complete, physical, analytical models for the current-voltage characteristics of MOSFETs were proposed by Hofstein and Heiman in 1963 [67], and by Sah in 1964 [16]. Both models rely on the drift-diffusion theory for charge transport. Within this fully classical theory, at or near equilibrium, electrons are frequently scattered and thus move erratically: they undergo diffusion. Out of equilibrium, on the other hand, electrons acquire a net motion in the direction opposite to that applied electric field, on top of their Brownian motion. This net motion along the field line is known as the drift motion. The average speed of electrons

can then be modeled as:

$$v(x) = \mu E(x) \tag{13}$$

where μ is the electron mobility, and $E(x)$ the norm of the electric field at position x along the channel.

Using Eq. 12, the drain current can thus be expressed above threshold as:

$$I_{DS} = WC_G(V_{GS} - V_T)\mu E_S \tag{14}$$

where we have chosen x to be the position of the source-channel interface, so that $E(x) = E_S$, and have parametrized the channel charge to be $C_G(V_{GS} - V_T)$. C_G denotes the total gate capacitance.

In the triode regime, the gradual channel approximation holds, so that given an applied drain voltage V_{DS}, the electric field is uniform throughout the channel: $E_S = \frac{V_{DS}}{L}$. Hence, the triode regime current is:

$$I_{DS_{\text{triode}}} = \frac{W}{L}\mu C_G V_{DS}(V_{GS} - V_T) \tag{15}$$

In the pentode regime, a careful analysis of MOS electrostatics [16, 67] indicates that the electric field at the source is $E_S = \frac{1}{2L}(V_{GS} - V_T)$. The pentode regime current is thus given by:

$$I_{DS_{\text{pentode}}} = \frac{W}{2L}\mu C_G(V_{GS} - V_T)^2 \tag{16}$$

The mean speed of charge carriers at the source in the pentode regime is thus seen to be:

$$v_{S_{\text{pentode}}} = \frac{\mu}{2L}(V_{GS} - V_T) \tag{17}$$

Eq. 17, however, fails to consider the increased levels of interaction with the lattice that arise once the charge carrier kinetic energy is large enough. Most prominent is the scattering of electrons by optical phonons once the field $E_S = \frac{1}{2L}(V_{GS} - V_T)$ is large enough [68]. Therefore, in the limit of small channel length L or particularly large gate voltage V_{GS},

the pentode regime carrier speed must be replaced by a material-specific constant known as the saturation velocity v_{sat} [69]. Typically, the electron velocity saturates at fields of $E \sim 5 \times 10^6 \text{ V} \cdot \text{m}^{-1}$, which given a typical applied drain to source of $V_{DS} \sim 1$ V, occurs for a channel length of $L \sim \frac{V_{DS}}{E} \sim 200$ nm. Velocity saturation is thus an important effect to consider for current-day MOSFETs modeling. Heuristically, the saturation velocity can be estimated as the speed of an electron with kinetic energy equal to the typical optical phonon energy E_{op}:

$$v_{\text{sat}} \sim \sqrt{\frac{2E_{\text{op}}}{m_t^\star}} \sim 10^5 \text{ m} \cdot \text{s}^{-1} \tag{18}$$

where $m_t^\star$ denotes the carrier effective mass along the transport direction.

The pentode regime drain current is then given by:

$$I_{DS_{\text{pentode}}} = WC_G \left(V_{GS} - V_T\right) v_{\text{sat}} \tag{19}$$

From Eq. 16 and Eq. 19, we note that $I_{DS_{\text{pentode}}} \propto \left(V_{GS} - V_T\right)^m$ with $1 \leq m \leq 2$, and that the exact value of the exponent m depends upon the importance of velocity saturation, and thus on the channel length L.

Finally, we note that the triode and pentode regime currents, as expressed in Eq. 15, Eq. 16, and Eq. 19, are divergent for $L \to 0$. This calls for revised models for short-channel devices, which we explore in Sec. 2.3.

2.3 Ballistic transport

As MOSFETs approach the nanoscale, the mean free path of charge carriers between scattering events approaches channel lengths. Electrons traverse entire devices without undergoing any scattering: they are undergoing ballistic transport. In this regime, the diffusive transport model of Sec. 2.2 is not an appropriate modeling framework anymore. Most notably, the concept of mobility breaks down in the ballistic limit. Indeed, as explained in Sec. 2.4, mobility is intimately linked to the timescales associated with scattering events. When no scattering events take place, such as for ballistic transport, this concept loses its

meaning. Eq. 15 and Eq. 16 for the drift-diffusion triode and pentode regimes drain current are thus not valid. Furthermore, velocity saturation effects do not occur in ballistic transistors [70]. Indeed, without optical phonon scattering, there is no velocity saturation. Eq. 19 for the drift-diffusion pentode regime current is thus also not valid.

One of the first complete and consistent ballistic transport model was introduced by Natori in 1994 [71]. The two ballistic transport frameworks that we expose in this section are based to a great extent on Natori's theory.

The development of these models relies heavily on complete Fermi-Dirac integrals, which we introduce in Appendix A.

2.3.1 Landauer-Büttiker approach to ballistic transport

First, we present a ballistic transport model based on the Landauer-Büttiker formalism, within which the current is expressed as [28]:

$$I_{DS} = \frac{e}{h} \int_{-\infty}^{+\infty} T(E)\, M(E) \left[f_S(E) - f_D(E) \right] dE \tag{20}$$

Several terms in Eq. 20 need to be defined. First, $f_{S,D}(E) = \left[1 + \exp\left(\frac{E - E_{F_{S,D}}}{k_B T}\right)\right]^{-1}$ represent the source and drain Fermi distributions, with $E_{F_{S,D}}$ the source and drain Fermi levels, and T the temperature of the device. Second, $T(E)$ represents the probability ($0 \leq T(E) \leq 1$) that an electron of energy E will transmit from the source to the drain, or *vice versa*. We note that the assumption that the source-to-drain and drain-to-source transmission probabilities are equal is in the most general case not valid, but will be appropriate for the purpose of our analysis. Third, $M(E)$ represents the number of transverse modes of energy less than E. We include spin degeneracy in $M(E)$, which explains the discrepancy between Eq. 11 and Eq. 20.

For ballistic transport, tautologically, $T(E) = 1$. Furthermore, for a semiconductor channel with parabolic conduction band, $M(E)$ is obtained by integrating [72] the 1D density

of states [73] along the width of the device channel:

$$\begin{aligned} M(E) &= W \int_{E_C}^{E} \frac{g_V}{\pi\hbar} \sqrt{\frac{2\frac{(\overline{m}^\star)^2}{m_t^\star}}{(E-E_C)}} dE \\ &= W \frac{2g_V}{\pi\hbar} \sqrt{2\frac{(\overline{m}^\star)^2}{m_t^\star}(E-E_C)} \end{aligned} \tag{21}$$

where $\overline{m}^\star$, $m_t^\star$, E_C, and g_V are respectively the density of states effective mass, effective mass along the transport direction, conduction band minimum, and valley degeneracy of the channel electron gas.

By Eq. 20, Eq. 21, and Eq. 128, the drain current can then be expressed in terms of complete Fermi-Dirac integrals as:

$$I_{DS} = We\frac{N_2}{2} v_T \left[\mathscr{F}_{\frac{1}{2}}(\eta_{F_S}) - \mathscr{F}_{\frac{1}{2}}(\eta_{F_D})\right] \tag{22}$$

where $N_2 = \frac{g_V \overline{m}^\star k_B T}{\pi\hbar^2}$, $v_T = \sqrt{\frac{2k_BT}{\pi m_t^\star}}$, $\eta_{F_S} = \frac{E_{F_S}-E_C}{k_BT}$, and $\eta_{F_D} = \eta_{F_S} - \frac{eV_{DS}}{k_BT}$.

Now, we wish to express Eq. 22 in terms of the charge density at the top of the barrier Q_0. In the ballistic limit, the top of the barrier is populated by two categories of electrons: right-moving electrons in equilibrium with the source, and left-moving electrons in equilibrium with the drain. In both cases, the density of states of these electrons is given by [73] $\rho_2(E) = \frac{\overline{m}^\star}{2\pi\hbar^2}\theta(E-E_C)$. Using Eq. 128, the charge density at the top of the barrier is then given by:

$$\begin{aligned} Q_0 &= -e\int_{E_C}^{\infty} \frac{\overline{m}^\star}{2\pi\hbar^2}\left[f_S(E) + f_D(E)\right] dE \\ &= -e\frac{N_2}{2}\left[\mathscr{F}_0(\eta_{F_S}) + \mathscr{F}_0(\eta_{F_D})\right] \end{aligned} \tag{23}$$

Combining Eq. 22 and Eq. 23, we then obtain:

$$I_{DS} = W|Q_0| \left[v_T \frac{\mathscr{F}_{\frac{1}{2}}(\eta_{F_S})}{\mathscr{F}_0(\eta_{F_S})}\right] \left[\frac{1 - \frac{\mathscr{F}_{\frac{1}{2}}(\eta_{F_D})}{\mathscr{F}_{\frac{1}{2}}(\eta_{F_S})}}{1 + \frac{\mathscr{F}_0(\eta_{F_D})}{\mathscr{F}_0(\eta_{F_S})}}\right] \tag{24}$$

The first term between square brackets in 24 represents the saturation-regime injection velocity of electrons at the top of the barrier. In the nondegenerate limit, it equals the unidirectional thermal velocity v_T, while it equals the Fermi velocity in the degenerate limit. The second term between square brackets quantifies the saturation of the current as a function of the drain voltage V_{DS}: for $V_{DS} = 0$, this factor equals 0, while it approaches 1 as $V_{DS} \to \infty$.

Note that the drain current, as expressed in Eq. 24, depends on η_{F_S}, a parameter which *a priori* is unknown, and can only be calculated after a full bandstructure analysis. Thus, for each value of the drain voltage V_{DS} and the gate voltage V_{GS}, to obtain η_{F_S}, one must first compute the charge at the top of the barrier Q_0 (for example, above threshold, this charge can be parametrized as $|Q_0| = C_G\,(V_{GS} - V_T)$), and use Eq. 23 to extract η_{F_S}. Eq. 24 can then be used directly to compute the drain current. In the nondegenerate limit, the procedure described in this paragraph is not necessary. Indeed, using the asymptotic form of Fermi-Dirac integrals exposed in Eq. 134, it can be shown that the right-hand side Eq. 24 is independent of E_{F_S}.

To obtain the scaling of I_{DS} on $(V_{GS} - V_T)$ in the inversion regime, we first note that $Q_0 \propto (V_{GS} - V_T)$. In the degenerate limit $\eta_{F_S} \gg 0$, by Eq. 23 and Eq. 135, $Q_0 \propto \eta_{F_S}$, so that $(V_{GS} - V_T) \propto \eta_{F_S}$. Furthermore, by Eq. 24 and Eq. 135, given that $\eta_{F_S} \geq \eta_{F_D}$, at the most, $I_{DS} \propto \eta_{F_S}^2 \propto (V_{GS} - V_T)^2$. On the other hand, in the nondegenerate limit $\eta_{F_S} \ll 0$, the drain current I_{DS} is independent of η_{F_S}, and hence $I_{DS} \propto (V_{GS} - V_T)$. We deduce that in ballistic MOSFETs, $I_{DS} \propto (V_{GS} - V_T)^m$, with $1 \leq m \leq 2$, and that the exact value of the exponent m depends upon the degree of degeneracy of the inversion channel 2DEG.

The model developed in this section can be extended to account for scattering [74], resulting in a description of transport continuous from the drift-diffusion limit to the ballistic limit. In this model, the expression for the drain current is similar to that of our Landauer-Büttiker approach-based model, with an additional parameter r (with $0 \leq r \leq 1$) representing the backscattering coefficient of electrons from the drain back to the drain. It

is given by:

$$I_{DS} = W\left|Q_0\right|\left[\frac{1-r}{1+r}\right]\left[v_T\frac{\mathscr{F}_{\frac{1}{2}}\left(\eta_{F_S}\right)}{\mathscr{F}_0\left(\eta_{F_S}\right)}\right]\left[\frac{1-\frac{\mathscr{F}_{\frac{1}{2}}\left(\eta_{F_D}\right)}{\mathscr{F}_{\frac{1}{2}}\left(\eta_{F_S}\right)}}{1+\left(\frac{1-r}{1+r}\right)\left(\frac{\mathscr{F}_0\left(\eta_{F_D}\right)}{\mathscr{F}_0\left(\eta_{F_S}\right)}\right)}\right] \tag{25}$$

Finally, while our analysis assumed that the main charge carriers were electrons, we note that it can be adapted to model hole transport [75].

2.3.2 Capacitor model

The model of Sec. 2.3.1 has the advantage of being completely analytical, in the sense that the current can be evaluated directly from Eq. 24, without the need of any iterative numerical procedure. However, it fails to consider certain effects crucial in the modeling of short-channel ballistic MOSFETs, such as DIBL, the quantum capacitance effect (see Sec. 3.1.3), and the floating source effect (see Sec. 2.3.2.1). The first two effects can be included *ad hoc* within Eq. 24 by defining the charge at the top of the barrier $|Q_0|$ to be $C_G\left(V_{GS}-V_T\right)$, with C_G including a quantum capacitance correction term, and V_T a DIBL correction term. However, this procedure can only be carrier out above threshold. Moreover, there is no easy way to adapt Eq. 24 to account for the floating source effect. For these reasons, we will introduce a model for ballistic MOSFETs, called the capacitor model, which includes a simple iterative calculation accounting self-consistently for MOS electrostatics, with great predictive power [76].

The first step in developing the capacitor model is to appreciate the fact that the top of the barrier is a distinguished position along the channel to carry out calculations of the drain current. Indeed, at the top of the barrier, left-moving states are in equilibrium with the source, while right-moving states are in equilibrium with the drain (as illustrated in Fig. 4). The top of the barrier is the sole point in along the channel with this property. The charge at the top of the barrier Q_0 depends upon the source and drain Fermi levels E_{F_S} and E_{F_D}, as well as the top of the barrier energy, which we will call the self-consistent potential U_{SCF}.

The heart of the capacitor model is to consider the capacitive coupling between the top of the barrier and the source, drain, and gate electrodes respectively, as illustrated in Fig. 5.

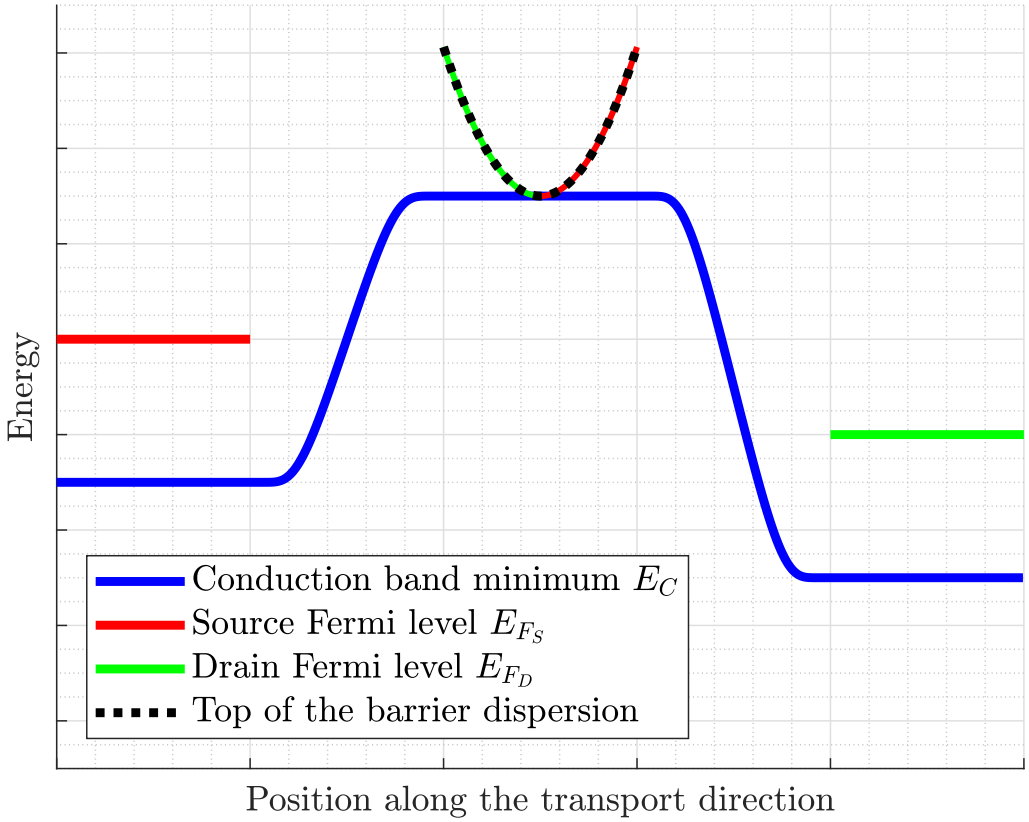


Figure 4 – Illustrative band diagram for MOSFETs studied in the capacitor model. At the top of the barrier, the right-moving states are in equilibrium with the source (represented by a red overline on the dotted dispersion), while the left-moving states are in equilibrium with the drain (represented by a green overline on the dotted dispersion).

The equilibrium charge density at the top of the barrier is given, according to Eq. 128, by:

$$Q_{0_{\mathrm{eq}}} = -eN_2\mathscr{F}_0\left(\frac{E_F - U_{SCF_{\mathrm{eq}}}}{k_BT}\right) \tag{26}$$

where $U_{SCF_{\mathrm{eq}}}$ is the equilibrium self-consistent potential. Out of equilibrium, $E_{F_S} \neq E_{F_D}$, so that the top of the barrier is populated by left-moving electrons in equilibrium with the source and right-moving electrons in equilibrium with the drain. The top of the barrier

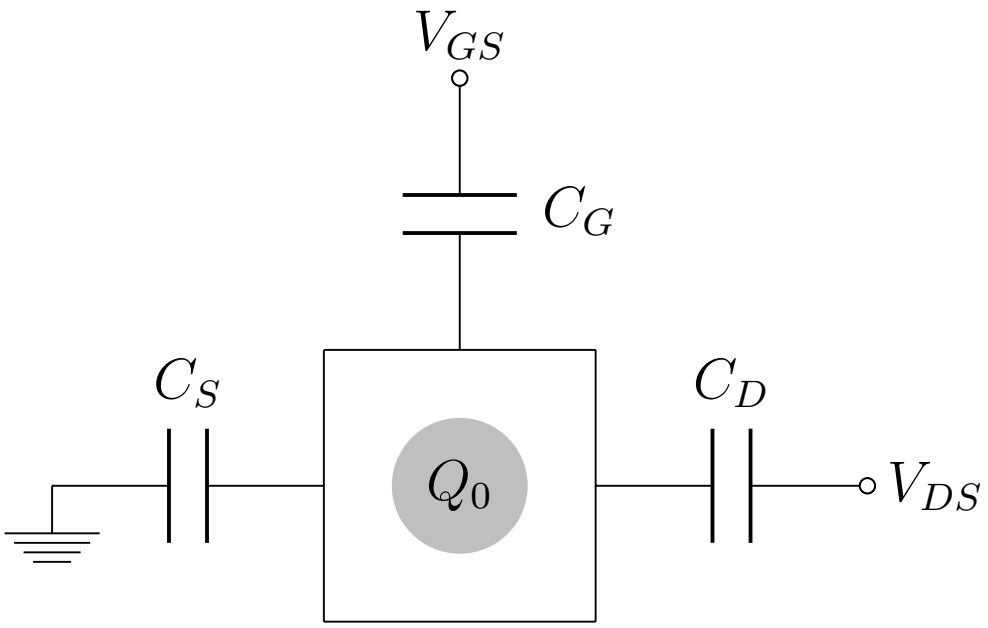


Figure 5 – Circuit diagram for the ballistic transport capacitor model. The top of the barrier, on which the areal charge density is given by Q_0, is capacitively coupled to the source, drain, and gate with capacitances C_S, C_D, and C_G respectively. The drain and gate biases are respectively V_{DS} and V_{GS}. Note that C_G is a purely electrostatic capacitance, and thus does not include quantum capacitance corrections (see Sec. 3.1.3).

charge density is then given by $Q_0 = Q_{0_L} + Q_{0_R}$ where Q_{0_L} and Q_{0_R} are the charge densities associated with the two aforementioned populations of electrons:

$$\begin{aligned} Q_{0_L} &= -\frac{eN_2}{2}\mathscr{F}_0\left(\frac{E_{F_S} - U_{SCF}}{k_B T}\right) \\ Q_{0_R} &= -\frac{eN_2}{2}\mathscr{F}_0\left(\frac{E_{F_D} - U_{SCF}}{k_B T}\right) \end{aligned} \tag{27}$$

Finally, we define the induced charge density at the top of the barrier to be:

$$\begin{aligned} &\Delta Q_0 = Q_{0_L} + Q_{0_R} - Q_{0_{\text{eq}}} \\ \implies &\Delta Q_0 = -\frac{eN_2}{2}\mathscr{F}_0\left(\frac{E_{F_S} - U_{SCF}}{k_B T}\right) - \frac{eN_2}{2}\mathscr{F}_0\left(\frac{E_{F_D} - U_{SCF}}{k_B T}\right) \\ &\qquad + eN_2\mathscr{F}_0\left(\frac{E_F - U_{SCF_{\text{eq}}}}{k_B T}\right) \end{aligned} \tag{28}$$

It is thus seen that the induced top of the barrier charge ΔQ_0 can be found if the self-consistent potential U_{SCF} is known. On the other hand, U_{SCF} is found by solving Poisson's

equation. Being linear, Poisson's equation can be solved by superposition. First, we ignore the contributions to U_{SCF} due to the charge induced at the top of the barrier ΔQ_0. Defining the drain and gate control parameters to be respectively:

$$\begin{aligned} \alpha_D &= \frac{C_D}{C_S + C_D + C_G} \\ \alpha_G &= \frac{C_G}{C_S + C_D + C_G} \end{aligned} \tag{29}$$

we find the contribution to U_{SCF} due to the applied voltages to be:

$$U_{SCF_1} = -e\left(\alpha_D V_{DS} + \alpha_G V_{GS}\right) \tag{30}$$

Note that we have assumed the source electrode to be grounded. The second contribution to U_{SCF} due to the induced charge ΔQ_0 is simply the capacitive charging energy associated with ΔQ_0:

$$U_{SCF_2} = -\frac{1}{C_S + C_D + C_G}\Delta Q_0 \tag{31}$$

By superposition, the net self-consistent potential is then given by:

$$\begin{aligned} & U_{SCF} = U_{SCF_1} + U_{SCF_2} \\ \implies & U_{SCF} = -e\left(\alpha_D V_{DS} + \alpha_G V_{GS}\right) - \frac{1}{C_S + C_D + C_G}\Delta Q_0 \end{aligned} \tag{32}$$

It is thus seen that the self-consistent potential U_{SCF} can be found if ΔQ_0 is known. The current can then be found using Eq. 22:

$$I_{DS} = We\frac{N_2}{2}v_T\left[\mathscr{F}_{\frac{1}{2}}\left(\frac{E_{F_S} - U_{SCF}}{k_B T}\right) - \mathscr{F}_{\frac{1}{2}}\left(\frac{E_{F_D} - U_{SCF}}{k_B T}\right)\right] \tag{33}$$

Overall, we obtain an iterative algorithm, whose steps are as follows:

1. Guess the value of the self-consistent potential U_{SCF}.
2. Compute the induced charge density ΔQ_0 using U_{SCF} and Eq. 32.
3. Compute the self-consistent potential U_{SCF} using ΔQ_0 and Eq. 28.
4. Repeat Step 2 and Step 3 until convergence is attained.
5. Compute the current using Eq. 33.

2.3.2.1 Floating source effect

Fig. 6 shows that as the gate voltage V_{GS} is increased, the conduction band minimum in the source E_{C_S} decreases. This phenomenon is characteristic of ballistic MOSFETs and can be explained as follows. When both the drain voltage V_{DS} and the gate voltage V_{GS} are non-zero, the source electrode is populated by three groups of charge carriers:

- Right-moving carriers with energy higher than the barrier height U_{SCF} that emanated from the source and are heading towards the drain. This group is in equilibrium with the source.
- Left-moving carriers with energy lower than the barrier height U_{SCF} that emanated from the source and were reflected by the barrier. This group is in equilibrium with the source.
- Left-moving carriers with energy higher than the barrier height U_{SCF} that emanated from the drain and are heading towards the source. This group is in equilibrium with the drain.

Thus, given a doping density in the source of N_{D_S}, charge neutrality guarantees that:

$$\begin{aligned} N_{D_S} &= \int_{-\infty}^{U_{SCF}} \rho_2 \left(E - E_{C_S}\right) f\left(E - E_{F_S}\right) dE \\ &\quad + \frac{1}{2} \int_{U_{SCF}}^{\infty} \rho_2 \left(E - E_{C_S}\right) \left[f\left(E - E_{F_S}\right) + f\left(E - E_{F_S} - eV_{DS}\right)\right] dE \\ N_{D_S} &= N_2 \log\left[\frac{1 + \exp\left(\frac{E_{F_S} - E_{C_S}}{k_B T}\right)}{1 + \exp\left(\frac{E_{F_S} - U_{SCF}}{k_B T}\right)}\right] + \frac{N_2}{2} \log\left[1 + \exp\left(\frac{E_{F_S} - U_{SCF}}{k_B T}\right)\right] \\ &\quad + \frac{N_2}{2} \log\left[1 + \exp\left(\frac{E_{F_S} + eV_{DS} - U_{SCF}}{k_B T}\right)\right] \end{aligned} \tag{34}$$

It is thus seen that as V_{GS} increases, which correspond to decreasing U_{SCF}, the source conduction band minimum E_{C_S} must decrease to ensure that the right-hand side of Eq. 34 remains constant. Equivalently, E_{C_S} can be kept fixed, while E_{F_S} increases.

To account for the floating source effect, the algorithm presented in Sec. 2.3.2 must be

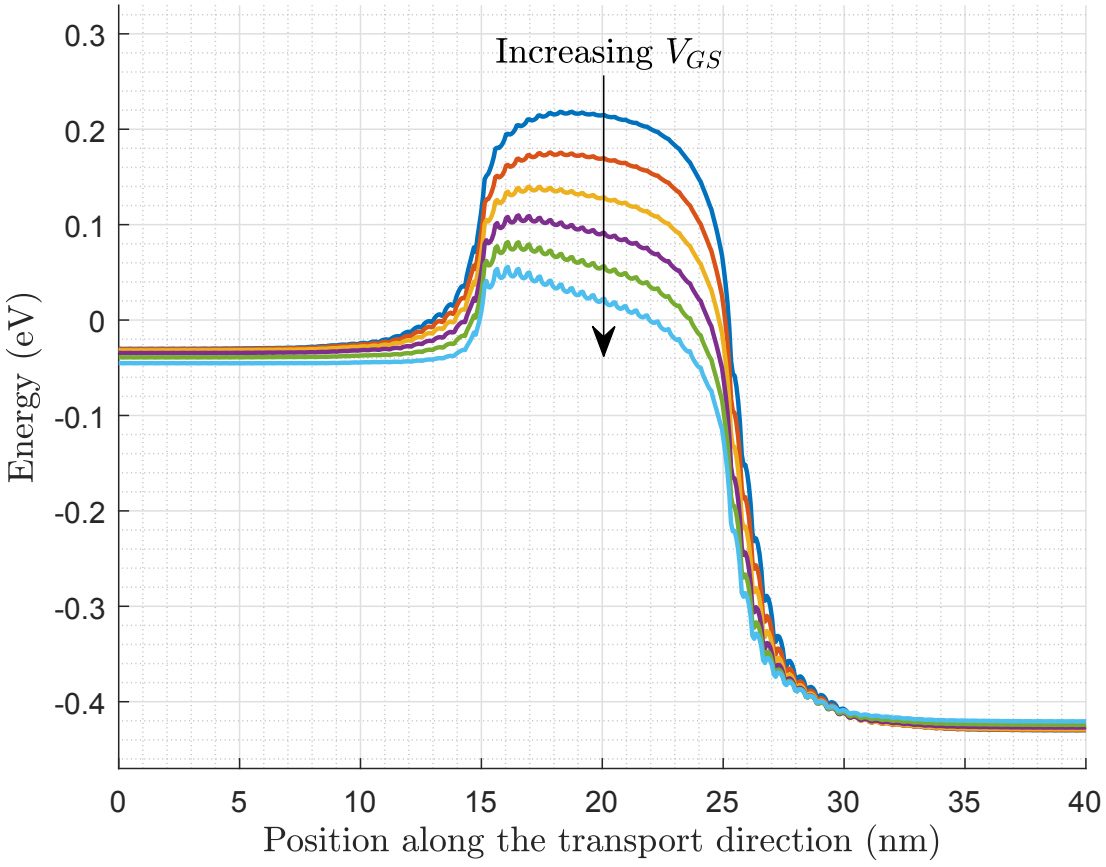


Figure 6 – Band diagrams of a 10 nm channel ML-BP MOSFET oriented along AD, calculated using self-consistent TB-NEGF simulations. The drain voltage is fixed at $V_{DS} = 0.4$ V, and the gate voltage is swept from $V_{GS} = 0.725$ V to $V_{GS} = 0.975$ V in steps of $\Delta V_{GS} = 0.05$ V. We note that the conduction band minimum in the source decreases ~ 10 mV as V_{GS} increases from 0.725 V to 0.975 V. This phenomenon is known as the floating source effect. Note that the undulations displayed by these band diagrams are a sign of the decline of the validity of Bloch's theorem: peaks and troughs correspond to individual atomic sites.

repeated for various values of E_{F_S} until one that ensures charge neutrality in the source is found.

2.4 Mobility

By the Lorentz force law, in a vacuum, an electric field $\vec{E}$ has the effect of accelerating a charge q of mass m with acceleration $\vec{a} = \frac{q\vec{E}}{m}$. However, in a typical bulk solid, moving charges (electrons or holes) are constantly being scattered (by crystal defects, phonons, impurities, *etc.*), and are thus not indefinitely accelerated by an external electric field. Instead, the motion of such charges can be described by their average velocity, called the drift velocity, which points in the same (opposite) direction of $\vec{E}$ if $q > 0$ (if $q < 0$), and whose magnitude is given by Ohm's law:

$$v_D = \mu E \tag{35}$$

The constant of proportionality μ between the drift velocity v_D and the magnitude of the external field E is known as the mobility, and is a material property depending on a number of factors such as the electronic and phononic bandstructures, the impurity and crystal defect densities, and temperature.

2.4.1 Relationship between mobility and scattering time

Let us consider electrons/holes in a crystal with effective masses $m^\star$ and charge q. We make the two following assumptions about the scattering of these charges:

- Charges move freely between two scattering events. We define the mean time between two scattering events as τ, and will henceforth refer to it as the relaxation time or the scattering time.
- Scattering events are memoryless, in the sense that the distribution of directions of the motion of a charge after a scattering event is isotropic.

Now, let $n(t)$ denote the number of charges that undergo a scattering event at time t. Then, at time $t + dt$, the size of this population of charges has decreased by an amount $dn = -\frac{1}{\tau} n(t)\, dt$. It follows that the amount of charges that have not undergone any scattering

event since an arbitrary time t_0 follows an exponential distribution:

$$n(t) = n(t_0) \exp\left(-\frac{t - t_0}{\tau}\right) \tag{36}$$

Under the semiclassical approximation [77], under the application of an external electric field $\vec{E}$, charges' velocities $\vec{v}$ satisfy:

$$\begin{aligned} m^\star \frac{d}{dt}\vec{v} &= q\vec{E} \\ \implies \vec{v}(t) &= \vec{v}(t_0) + \frac{q}{m^\star}(t - t_0)\vec{E} \end{aligned} \tag{37}$$

and by isotropy of scattering events, $\vec{v}(t_0) = \vec{0}$.

Averaging out, the drift velocity of charges in response to the external field is:

$$\begin{aligned} v_D &= \frac{1}{n(t_0)} \int_{n(t_0)}^{0} \frac{q}{m^\star}(t - t_0)\vec{E} dn \\ &= \frac{q}{m^\star}\vec{E} \int_{t_0}^{\infty} \frac{1}{\tau}(t - t_0) \exp\left(-\frac{t - t_0}{\tau}\right) dt \\ \implies v_D &= \frac{q\tau}{m^\star}\vec{E} \end{aligned} \tag{38}$$

Thereupon, using Eq. 35 and Eq. 38, we find that the mobility μ and scattering time τ are related through:

$$\mu = \frac{q\tau}{m^\star} \tag{39}$$

2.4.2 Matthiessen's rule

Consider $n \in \mathbb{N}$ scattering mechanisms with associated scattering times $\tau_1,\ \tau_2,\ \cdots,\ \tau_n$. Assuming that all of these mechanisms are independent from each other, using Eq. 36, the probability that a charge carrier will be scattered at some time $t > t_0$ is:

$$\begin{aligned} P &\propto \exp\left(-\frac{t - t_0}{\tau_1}\right) \exp\left(-\frac{t - t_0}{\tau_2}\right) \cdots \exp\left(-\frac{t - t_0}{\tau_n}\right) \\ &\propto \exp\left[-(t - t_0)\left(\frac{1}{\tau_1} + \frac{1}{\tau_2} + \cdots + \frac{1}{\tau_n}\right)\right] \end{aligned} \tag{40}$$

Thus, as a whole, the scattering mechanisms can be described by a single scattering time, given by:

$$\tau = \left(\frac{1}{\tau_1} + \frac{1}{\tau_2} + \cdots + \frac{1}{\tau_n}\right)^{-1} \tag{41}$$

Using Eq. 39 and Eq. 41, we then obtain the following important result:

Matthiessen's rule: If charge carriers in a material are under the influence of $n \in \mathbb{N}$ independent scattering mechanisms with mobilities $\mu_1,\ \mu_2,\ \cdots,\ \mu_n$, then their drift velocities can be described by the single mobility:

$$\mu = \left(\frac{1}{\mu_1} + \frac{1}{\mu_2} + \cdots + \frac{1}{\mu_n}\right)^{-1} \tag{42}$$

2.4.3 Ballistic mobility

As we have seen, ballistic transport consists in the transport of charge carriers with little or no scattering. This idealized regime of conduction is attained in nanoelectronic devices with dimensions smaller than the mean free path of charge carriers (namely the distance travelled by charge carriers during a scattering time). The concept of mobility thus breaks down in the ballistic regime: charge carriers are accelerated by an external electric field without being scattered, and their motions cannot be described by a drift velocity. Nevertheless, in nanoelectronic devices with small dimensions, transport of charge carriers can still effectively be described by a mobility known as the ballistic mobility, which was first introduced by Shur [78, 79].

2.4.3.1 Derivation of the ballistic mobility in the nondegenerate limit

Consider a MOSFET with channel length L and width W having for main charge carriers spin-degenerate electrons of density-of-states effective mass $\overline{m}^\star$, effective mass along the transport direction $m_t^\star$, and valley degeneracy g_V. Furthermore, let T be the temperature of the system. Let E_C denote the conduction band minimum in the channel. Let E_{F_S} and

E_{F_D} respectively denote the Fermi levels in the source and drain of the MOSFET. Assume that the charge carriers obey nondegenerate statistics, so that their areal densities in the source and drain are:

$$n_{S,D} = \frac{g_V \overline{m}^\star k_B T}{\pi \hbar^2} \exp\left(\frac{E_{F_{S,D}} - E_C}{k_B T}\right) \tag{43}$$

In the nondegenerate limit, assuming that there are no applied biases on the terminals of the MOSFET, the velocity distribution for charge carriers is a Maxwell-Boltzmann distribution:

$$P\left(\vec{v}\right) d^2\vec{v} = \frac{\alpha}{\pi} e^{-\frac{m_x^\star}{2k_B T}v_x^2 - \frac{m_y^\star}{2k_B T}v_y^2} d^2\vec{v} \tag{44}$$

where $m_x^\star$ and $m_y^\star$ are respectively the effective masses along the Cartesian principal axes x and y of the effective mass tensor of the channel two-dimensional electron gas (2DEG). By the equipartition theorem, $\alpha = \frac{\overline{m}^\star}{2k_B T}$. The unidirectional thermal speed of charge carriers along the transport direction t is thus given by:

$$\begin{aligned} v_T = \langle |v_t| \rangle &= \frac{\alpha}{\pi} \iint\limits_{\mathbb{R}^2} |v_t| \, e^{-\frac{m_x^\star}{2k_B T}v_x^2 - \frac{m_y^\star}{2k_B T}v_y^2} dv_x dv_y \\ \implies v_T &= \sqrt{\frac{2k_B T}{\pi m_t^\star}} \end{aligned} \tag{45}$$

where m_t is the effective mass along t. The unidirectional thermal speed corresponds to the speed of charge carriers in moving from the source to the drain, and *vice versa*.

The fundamental assumption underlying ballistic transport is that charge carriers remain in thermal equilibrium with the electrode from which they emanate as they travel through the channel of the MOSFET. The current through the MOSFET is thus given by:

$$\begin{aligned} I &= -eW\left(n_S v_T - n_D v_T\right) \\ I &= WQv_T \frac{1 - (n_D/n_S)}{1 + (n_D/n_S)} \end{aligned} \tag{46}$$

where we have used the fact that the charge in the inversion channel is $Q = -e\left(n_S + n_D\right)$.

Now, let $x = \frac{n_D}{n_S} = \exp\left(\frac{E_{F_D} - E_{F_S}}{k_B T}\right)$ by Eq. 43. We note that the applied drain to source voltage is by definition $V_{DS} = \frac{1}{e}\left(E_{F_S} - E_{F_D}\right)$. Assuming that $V_{DS} \ll \frac{k_B T}{e}$, by Taylor expansion:

$$x \approx 1 - \frac{eV_{DS}}{k_B T} \tag{47}$$

$$\frac{1-x}{1+x} \approx \frac{eV_{DS}}{2k_BT} \tag{48}$$

Combining Eq. 46, Eq. 47, and Eq. 48, we find that the ballistic current for small V_{DS} is:

$$I_{\text{ball}} = WQ\frac{eV_{DS}v_T}{2k_BT} \tag{49}$$

On the other hand, the corresponding drift-diffusion current for small V_{DS} can be described by a mobility μ:

$$I_{\text{diff}} = WQ\frac{\mu V_{DS}}{L} \tag{50}$$

Thus, an effective mobility, namely the ballistic mobility, can be found by setting $I_{\text{ball}} = I_{\text{diff}}$. From Eq. 49 and Eq. 50:

$$\mu_{\text{ball}} = \frac{v_T L}{2k_BT/e} = \frac{eL}{\pi m_t^\star v_T} \tag{51}$$

Physically, μ_{ball} describes scattering at the source/channel and channel/drain interfaces. Now, suppose that the channel of the MOSFET under consideration is made of a material with bulk mobility μ_{bulk}. Then, according to Matthiessen's rule (Eq. 42), the net mobility of the channel is given by:

$$\mu_{\text{tot}} = \left(\frac{1}{\mu_{\text{bulk}}} + \frac{1}{\mu_{\text{ball}}}\right)^{-1} \tag{52}$$

We can conclude that for $V_{DS} \ll \frac{k_BT}{e}$, the mobility of the channel has a supremum, the ballistic mobility μ_{ball}, set by the channel length L, the carrier effective mass $m_t^\star$, and the temperature T. This is illustrated in Fig. 7.

2.4.3.2 Derivation of the ballistic mobility for any level of degeneracy

In the drift-diffusion regime, the drain current under low drain bias is expressed as (Eq. 15):

$$I_{DS} = WQ\mu\frac{V_{DS}}{L} \tag{53}$$

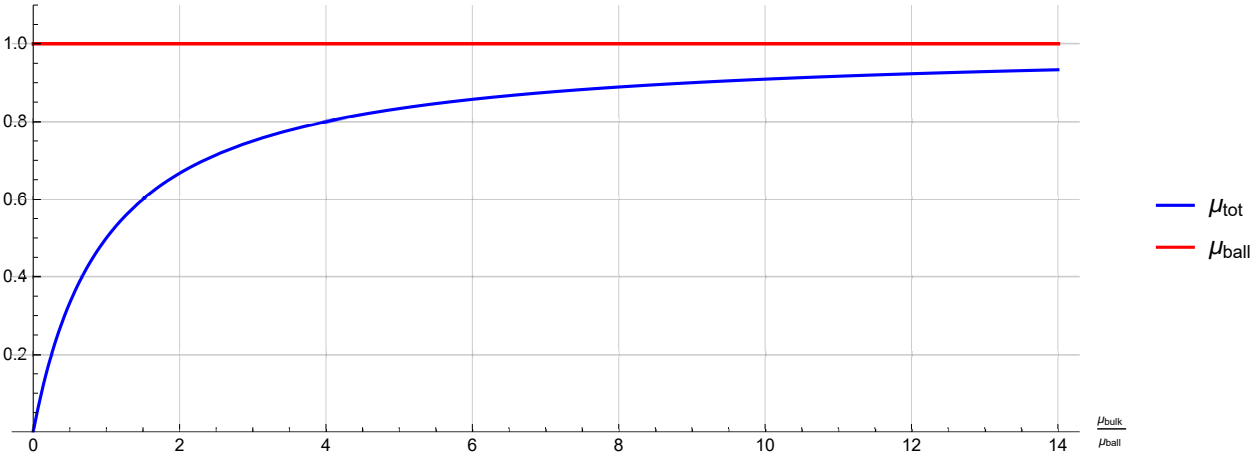


Figure 7 – μ_{tot} as a function of μ_{bulk}, where we have set $\mu_{\text{ball}} = 1$. The total mobility μ_{tot} is seen to be bounded above by μ_{ball}. Note that this conclusion is only valid for $V_{DS} \ll \frac{k_B T}{e}$.

On the other hand, in the ballistic limit, the drain current is (Eq. 24):

$$I_{DS} = WQv_T \frac{\mathscr{F}_{\frac{1}{2}}(\eta_F) - \mathscr{F}_{\frac{1}{2}}(\eta_F - U_D)}{\mathscr{F}_0(\eta_F) + \mathscr{F}_0(\eta_F - U_D)} \tag{54}$$

where $\eta_F = \frac{E_{F_S} - E_C}{k_B T}$ and $U_D = -\frac{eV_{DS}}{k_B T}$. For low drain biases $V_{DS} \ll \frac{k_B T}{e}$, one has $U_D \approx 0$, so that the ballistic drain current becomes:

$$\begin{aligned} I_{DS} &= WQv_T \frac{1}{2\mathscr{F}_0(\eta_F)} \frac{\partial \mathscr{F}_{\frac{1}{2}}(\eta_F)}{\partial \eta_F} U_D \\ &= WQv_T \frac{1}{2\mathscr{F}_0(\eta_F)} \mathscr{F}_{-\frac{1}{2}}(\eta_F) \frac{eV_{DS}}{k_B T} \end{aligned} \tag{55}$$

As before, we simply equate the drift-diffusion current to the ballistic current (Eq. 54 and Eq. 55) to extract the ballistic mobility:

$$\mu_{\text{ball}} = \frac{eL}{\pi m_t^\star v_T} \frac{\mathscr{F}_{-\frac{1}{2}}(\eta_F)}{\mathscr{F}_0(\eta_F)} \tag{56}$$

Note that Eq. 56 reduces to Eq. 51 in the nondegenerate limit, by Eq. 134.

We conclude that under ballistic conditions, the current through a device can still be described in terms of the drift-diffusion formalism, but this procedure only works for drain biases smaller than the thermal voltage. At room temperature, this voltage equal $\frac{k_B T}{e} \approx 25.8$ mV. Since typical applied drain biases in MOSFETs are in the range $\sim 0\text{ V} - 2\text{ V}$, the

concept of ballistic mobility has a highly limited applicability in the modeling of MOSFETs. However, it can be used to extract low bias quantities, such as the channel conductivity $\left.\frac{\partial I_{DS}}{\partial V_{DS}}\right|_{V_{DS}=0}$ [80]. In Sec. 3.2.2, we use the concept of ballistic mobility to extract the threshold voltage from the transfer characteristics of MOSFETs biased at $V_{DS} \ll \frac{k_B T}{e}$.

3 Important Parameters for MOSFET Modeling

In this section, we develop methods to compute four important parameters in the modeling of MOSFETs: the gate capacitance, the threshold voltage, the subthreshold swing, and the DIBL. The gate capacitance C_G is calculated analytically by considering fringing field effects, wavefunction penetration effects, and quantum capacitance effects. The threshold voltage V_T is extracted from transfer characteristics; 7 different methods are compared, two of which are promoted as being more appropriate for modeling purposes. The subthreshold swing S and DIBL δ are also extracted from transfer characteristics. Finally, we examine how the gate and drain control parameters α_G and α_D, that were introduced in Sec. 2.3.2, vary as a function of V_{DS}.

3.1 Gate capacitance

The (areal) gate capacitance C_G is a crucial physical parameter in the modeling of the current-voltage characteristics of MOSFETs. Indeed, for gate voltages V_{GS} larger than the threshold voltage V_T, the absolute value of the (areal) charge density Q at the top of the potential barrier of a MOSFET is given by:

$$Q = C_G \left(V_{GS} - V_T \right) \tag{57}$$

While Eq. 57 is not exact, it is accurate enough for the purpose of our compact modeling. A naive expression for C_G is simply the (areal) capacitance of a parallel plate capacitor with infinite plane plates:

$$C_G = \frac{\varepsilon}{t} \tag{58}$$

where ε and t respectively denote the permittivity and thickness of the oxide layer between the metal gate and the semiconductor channel. Nevertheless, this expression fails to consider a number of physical effects which affect the capacitive coupling between the gate and the channel. In the following pages, we introduce some of these effects, and attempt to quantify their contributions to C_G.

3.1.1 Fringing fields

Deriving Eq. 58 is an exercise typically done in undergraduate classes on electrostatics. From Gauss' law, the electric field produced by an infinite plane with uniform areal charge density σ immersed in a medium of permittivity ε is equal to $\frac{\sigma}{2\varepsilon}$ in norm, and points away from (towards) the plane when $\sigma > 0$ (when $\sigma < 0$). Consequently, when two such planes with opposite areal charge densities $\pm\sigma$ are placed parallel to each other and separated by a distance t, the field between them has norm $\frac{\sigma}{\varepsilon}$ (and points from the plate of positive areal charge density towards the plate of negative areal charge density), and is $\vec{0}$ everywhere else. Integrating this field between the two planes, we find that the potential difference between them is $\frac{\sigma t}{\varepsilon}$. Therefore, by definition, the capacitance of the capacitor formed by the two planes is $\frac{\varepsilon}{t}$, as stated in Eq. 58.

The derivation above assumes that the field lines are entirely located between the two planes, and are strictly orthogonal to them. However, when the dimensions of the planes are finite, fringing fields appear. Fringing fields are fields that curl around the edges of the plates of a finitely sized capacitor. Their existence is guaranteed by Gauss' law. Indeed, consider the closed loop γ illustrated in Fig. 8. If the electric field $\vec{E}$ remained orthogonal to the planes and terminated abruptly near their edges, then the line integral $\oint_\gamma \vec{E} \cdot d\vec{l}$ could not be equal to 0, thus violating Gauss' law. The effect of fringing fields is to increase the value of the capacitance compared to that calculated from Eq. 58.

3.1.1.1 Statement of the problem and outline of its solution

Consider two metallic plates, P_1 and P_2, of infinitesimal thickness immersed in an environment of permittivity ε. These plates are both of infinite extent along the z dimension, and finitely sized along the x dimension. We define the width w of the plates to be their extent along the x dimension. Finally, the two plates are placed parallel to each other, separated

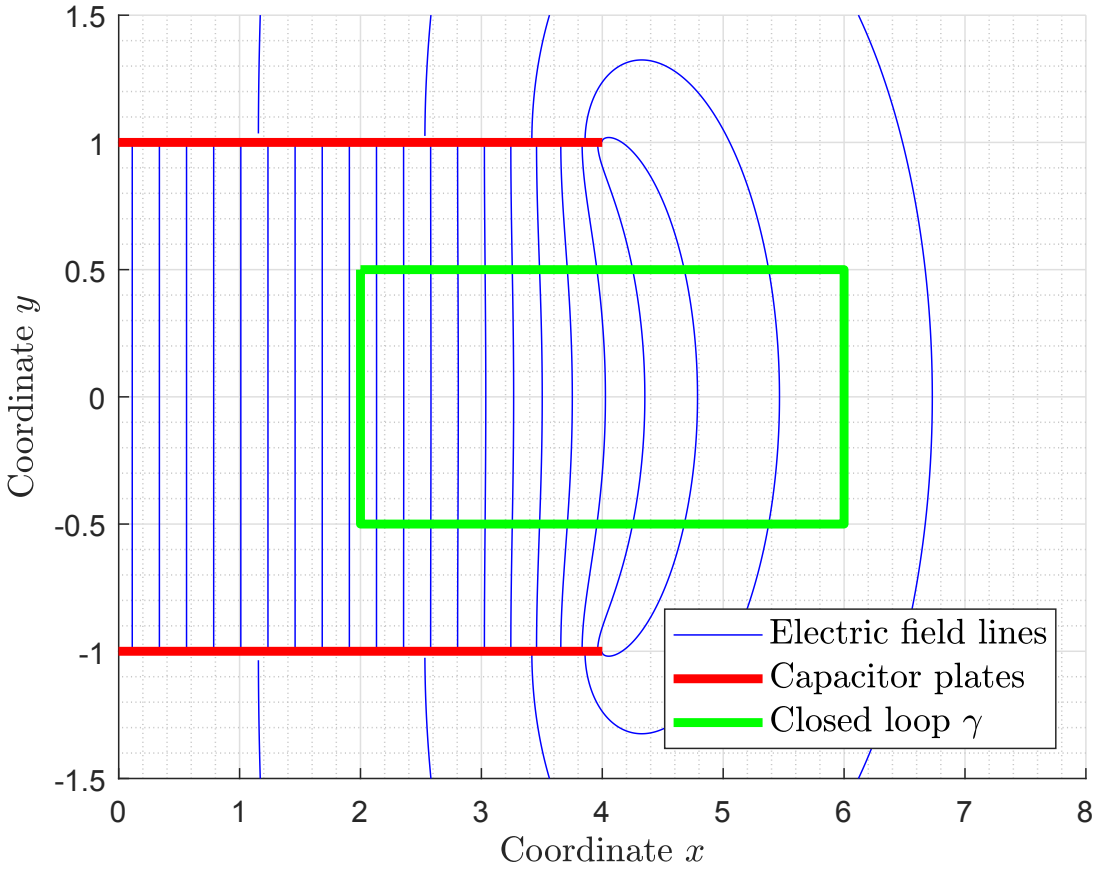


Figure 8 – Fringing fields near the edges of a finitely sized parallel plate capacitor. The capacitor represented has plates of width $w = 8$ separated by a distance of $t = 2$, and with electrostatic potentials $\pm\frac{1}{2}$.

by a distance t along the y dimension, so that we can formally define P_1 and P_2 to be:

$$\begin{aligned} P_1 &= \left\{ (x,y,z) \in \mathbb{R}^3 \,\middle|\, -\frac{w}{2} \leq x \leq \frac{w}{2}, y = \frac{t}{2} \right\} \\ P_2 &= \left\{ (x,y,z) \in \mathbb{R}^3 \,\middle|\, -\frac{w}{2} \leq x \leq \frac{w}{2}, y = -\frac{t}{2} \right\} \end{aligned} \tag{59}$$

We set the electrostatic potentials on P_1 and P_2 to be respectively $\frac{V}{2}$ and $-\frac{V}{2}$. The problem that we now aim to solve is the following: what is the capacitance of the capacitor formed by the plates P_1 and P_2? This problem could, of course, be solved numerically by discretizing space on a lattice $\mathbb{L}$ and casting Poisson's equation into a matrix equation on $\mathbb{L}$. However, following our philosophy that complicated numerical procedures are not

appropriate for the purpose of compact modeling, we will instead attempt to solve this problem analytically.

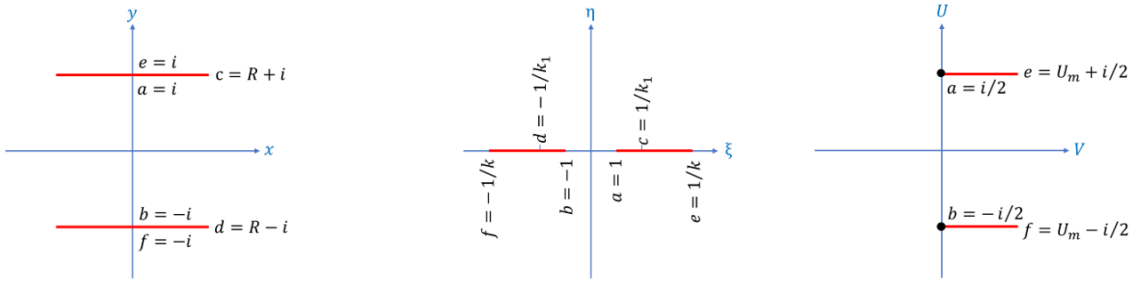
Calculating analytically the capacitance of a finite parallel plate capacitor, taking fringing fields into consideration, is a problem that has been approached by Maxwell [81], who used a number of simplifying approximations. Thomson [82] later suggested an exact approach to the problem making use of conformal mappings, which was worked out by Love and Bromwich [83], and promoted by Palmer [84]. It is this approach that we will expose. We will ultimately obtain an exact analytical expression for the capacitance under investigation, expressible entirely in terms of the ratio $R = \frac{w}{t}$ and the permittivity ε.

As a reminder, a harmonic function is a function satisfying Laplace's equation, and that can therefore be expressed in terms of a scalar potential. Conformal mappings are angle-preserving transformations between subsets of the complex plane $\mathbb{C}$. Given a harmonic function f defined on a connected, open subset of $\mathbb{C}$, the composition of any conformal mapping with f is also harmonic. This fact makes conformal mappings particularly useful tools for solving problems in electrostatics with convoluted geometries. Indeed, with an appropriate choice of conformal mapping, such problems can be cast into geometrically simpler forms. The solution can then be mapped back to the original space using the inverse of the chosen conformal mapping. It is this line of thought that we will follow.

Conformal mappings and Schwarz-Christoffel are introduced in greater depth in Appendix B. Furthermore, our calculation will involve elliptic integrals, the basic properties of which are exposed in Appendix C.

3.1.1.2 Definition of the complex domains to be used for our calculation

The first step for the purpose of our calculation in nondimensionalization. We set the width of the two plates of the capacitor under investigation (defined in Eq. 59) to be $w = 2R$, their separation to be $t = 2$, and the potential on the plates to be $\pm\frac{1}{2}$. Under these assumptions, the capacitor can be represented by the $Z = x + iy$ plane, as show in Fig. 9a.



(a) Domain $Z = x + iy$ (b) Domain $\zeta = \xi + i\eta$ (c) Domain $W = V + iU$

Figure 9 – Definition of the three complex domains that shall be mapped to each other through Schwarz-Christoffel transformations for the purpose of our calculation. Polygon edges are represented by red lines, and polygon vertices are labelled by the letters a-f. The black dots in (c) depict vertices a and b. The domain Z represents real space, and the domain W represents potential-flux space.

By virtue of the reflection symmetry of the system along the y axis, we can further simplify the problem by only considering the right half of the Z plane (namely the $x \geq 0$ half). Thus, the capacitor plates can be viewed as a polygon with vertices a-f that has collapsed such that the sides ac and ec on the one hand, and the sides bd and fd on the other hand, overlap. In the Z plane, fringing field lines are present, making the calculation of the capacitance of the capacitor a geometrically complex problem.

The $\zeta = \xi + i\eta$ plane is illustrated in Fig. 9b, and will be the domain of definition of our Schwarz-Christoffel mappings (see Eq. 136). The real axis of ζ will be mapped to the polygonal edges of the capacitor in Z. The precise meaning of the constants k and k_1 will be made explicit later.

Finally, the $W = V + iU$ plane is illustrated in Fig. 9c. V and U respectively represent electric potential and electric flux. U_m represents half of the total amount of electric flux between the two plates of the capacitor. In the Z plane, field lines and equipotential lines cross each other at right angle, as is well known from elementary electrostatics [85]. Likewise, these lines also cross each other at right angle in the W plane, but they possess the additional property of being straight lines. Our goal will thus be to obtain a conformal mapping from

Z to W, compute the investigated capacitance in W, and map the problem back to Z using the inverse mapping.

3.1.1.3 Schwarz-Christoffel mapping from ζ to Z

The Schwarz-Christoffel transformation $f_1 \colon \zeta \to Z$ can be found using Fig. 9a and Fig. 9b. In the notation of Eq. 136, the interior polygonal angles α, β, γ, δ, ϵ, and ϕ corresponding to vertices a, b, c, d, e, and f respectively are:

$$\begin{cases} \alpha = \frac{\pi}{2} \\ \beta = \frac{\pi}{2} \\ \gamma = 2\pi \\ \delta = 2\pi \\ \epsilon = \frac{\pi}{2} \\ \phi = \frac{\pi}{2} \end{cases} \tag{60}$$

Thus, from Eq. 136, we obtain:

$$\begin{aligned} f_1(\zeta) &= A_1 \int^{\zeta} \frac{w^2 - \frac{1}{k_1^2}}{\sqrt{\left(w^2 - \frac{1}{k^2}\right)(w^2 - 1)}} dw \\ &= A_1 \left[k_1^2 E(\phi, k) - \left(k_1^2 - k^2\right) F(\phi, k) \right] + B_1 \end{aligned} \tag{61}$$

where $\sin\phi = \zeta$. We now impose appropriate boundary conditions to find the values of the constants A_1 and B_1:

$$f_1(0) = 0 \implies B_1 = 0 \tag{62}$$

$$f_1(\pm 1) = \pm i \implies A_1 = \frac{i}{k_1^2 E(k) - (k_1^2 - k^2) K(k)} \tag{63}$$

$$f_1\left(\pm \frac{1}{k}\right) = \pm i \implies k_1^2 = \frac{K'(k)}{E'(k)} k^2 \tag{64}$$

The final boundary condition is $f_1\left(\pm\frac{1}{k_1}\right) = R \pm i$, from which it follows that:

$$\begin{aligned} R &= \frac{K'(k)\,E'(\beta,k) - E'(k)\,F'(\beta,k)}{[E'(k) - K'(k)]\,K(k) + K'(k)\,E(k)} \\ &= \frac{2}{\pi}\left[K'(k)\,E'(\beta,k) - E'(k)\,F'(\beta,k)\right] \end{aligned} \tag{65}$$

where we have used Legendre's relation (Eq. 145), and where:

$$\sin^2(\beta) = \frac{K'(k) - E'(k)}{(1-k^2)\,K'(k)} \tag{66}$$

Thus, given a value of the plate width to plate separation ratio R, the elliptic modulus k can be computed using Eq. 65 and Eq. 66.

3.1.1.4 Schwarz-Christoffel mapping from ζ to W

The Schwarz-Christoffel transformation $f_2 \colon \zeta \to W$ can similarly be calculated to be:

$$\begin{aligned} f_2(\zeta) &= A_2 \int^{\zeta} \frac{1}{\sqrt{\left(w^2 - \frac{1}{k^2}\right)(w^2-1)}} dw \\ &= A_2 F(\phi,k) + B_2 \end{aligned} \tag{67}$$

where $\sin\phi = \zeta$. We now impose appropriate boundary conditions to find the values of the constants A_2 and B_2:

$$f_2(0) = 0 \implies B_2 = 0 \tag{68}$$

$$f_2(\pm 1) = \pm\frac{i}{2} \implies A_2 = \frac{i}{2K(k)} \tag{69}$$

The inverse mapping $f_2^{-1} \colon W \to \zeta$ of f_2 then is:

$$f_2^{-1}(W) = -\sin\left[\mathrm{am}\left(2iK(k)\,W, k\right)\right] \tag{70}$$

where am denotes the Jacobi amplitude function, defined as the solution $\mathrm{am}(u,k) = \phi$ of the equation $F(\phi,k) = u$.

3.1.1.5 Field lines

Using the results of Sec. 3.1.1.3 and Sec. 3.1.1.4, the fringing field lines in real space can easily be calculated. Indeed, it suffices to map straight vertical lines in W to Z through the mapping $f_1 \circ f_2^{-1}$. This procedure was carried out for four different values of the capacitors plates width to plates separations ratios R, and the results are plotted in Fig. 10. We note that the fringing of the field near the edges of the capacitor becomes more pronounced as R is made smaller, as expected.

3.1.1.6 Exact capacitance calculation

From Fig. 9c, it is apparent that the point e in the W domain has for real part half of the total flux between the plates of the capacitor, and for imaginary part half of the potential difference between the two plates. Fig. 9b indicates that the value of ζ that is mapped to e by f_2 is $\zeta = \frac{1}{k}$. We deduce from Eq. 67 that in the W plane, the coordinates of the point e are:

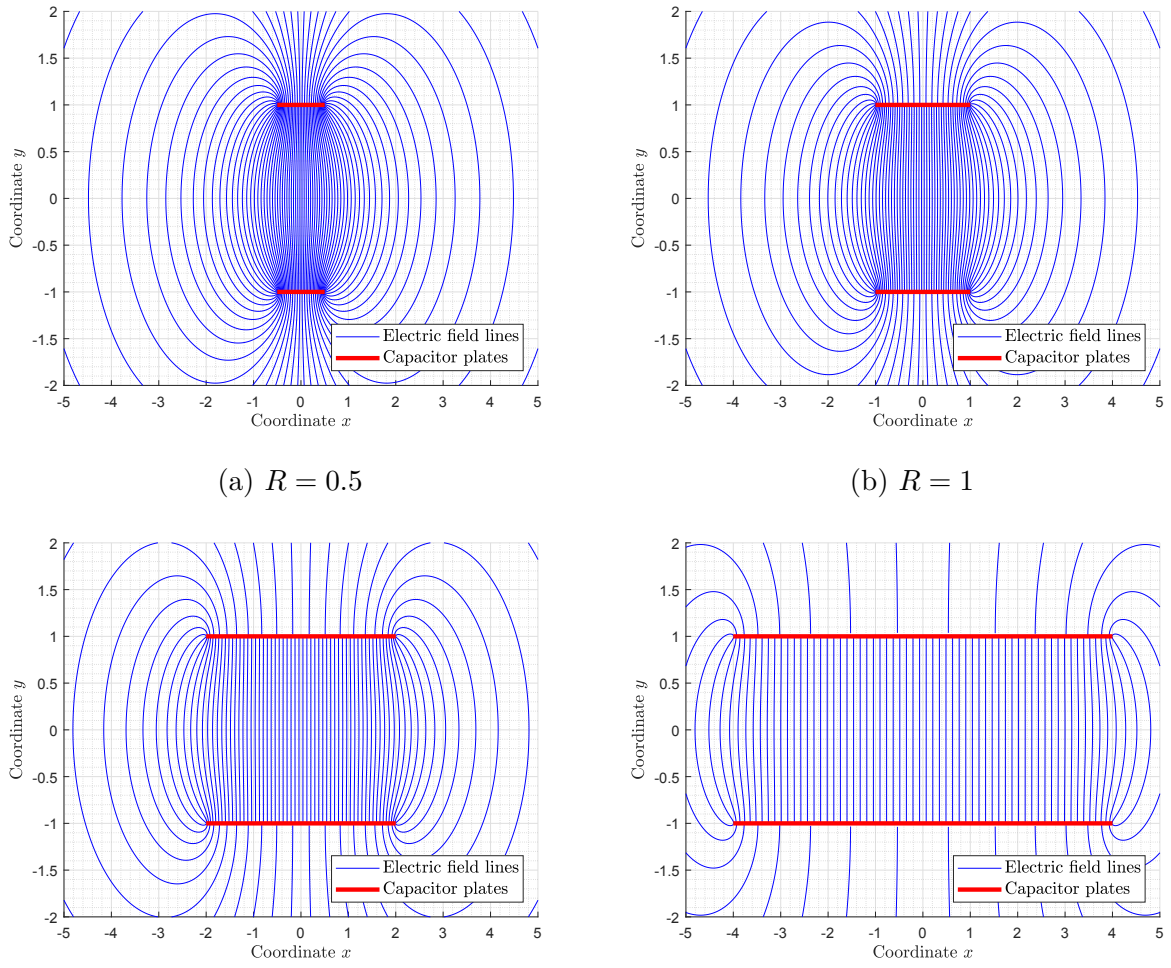
$$e = \frac{K'(k)}{2K(k)} + \frac{i}{2} \tag{71}$$

From Gauss' law and by definition of the capacitance, it follows that the capacitor under investigation has capacitance (per unit length along z):

$$C_{\text{Exact}} = \varepsilon \frac{K'(k)}{K(k)} \tag{72}$$

This formula takes into account all fringing field effects, and is thus exact. To summarize the procedure of finding the capacitance of a finitely size parallel plate capacitor, one must:

1. Compute the plate width to plate separation ratio R.
2. Compute the value of the elliptic modulus k using Eq. 65 and Eq. 66.
3. Compute the capacitance using Eq. 72.



(a) $R = 0.5$ (b) $R = 1$

(c) $R = 2$ (d) $R = 4$

Figure 10 – Field lines of finitely sized capacitors calculated using Schwarz-Christoffel transformations. The calculations were repeated for four different values of the capacitor plate width to plate separation ratio R.

3.1.1.7 Approximative formulas for the capacitance

While Eq. 72 is exact and analytical, it involves elliptic integrals, which can be difficult to evaluate. As such, a number of physicists and engineers have suggested approximations of 72 in terms of elementary functions. Maxwell [81] suggested the following approximation, valid in the large R limit:

$$C_{\text{Maxwell}} = \varepsilon R \left[1 + \frac{1}{\pi R} \{ 1 + \log(1 + \pi R) \} \right] \tag{73}$$

Bromwich and Love [83] suggested the following similar but improved approximation, valid again in the large R limit:

$$C_{\text{Bromwich}} = \varepsilon R \left[1 + \frac{1}{\pi R}\{1 + \log(2\pi R)\}\right] \tag{74}$$

Finally, Palmer [84] suggested the following approximation, valid in the small R limit:

$$C_{\text{Palmer}} = \frac{\pi\varepsilon}{\log\left(\frac{4}{R}\right)} \tag{75}$$

An approximation C_{Approx} of C_{Exact} (Eq. 72) valid for all values of R must satisfy the two following properties:

$$\lim_{R\to 0} C_{\text{Approx}} = 0 \tag{76}$$

$$\lim_{R\to\infty} (C_{\text{Approx}} - C_{\text{Infinite plates}}) = 0 \tag{77}$$

where $C_{\text{Infinite plates}}$ is the capacitance obtained without considering fringing field effects (as in Eq. 58):

$$C_{\text{Infinite plates}} = \varepsilon R \tag{78}$$

Note that none of the three approximations presented in Eq. 73, Eq. 74, and Eq. 75 simultaneously satisfy both of the boundary conditions of Eq. 76 and Eq. 77. To remedy to this issue, we propose the following model for C_{Approx}:

$$C_{\text{Fitted}} = \varepsilon R \left(1 + \frac{a}{R^b}\right) \tag{79}$$

where $a \in \mathbb{R}$ and $0 < b < 1$. We have computed the best-fit values of the parameters a and b using a linear fitting algorithm. The resulting approximation is surprisingly accurate, as can be seen in Fig. 11 and Tab. 1. In particular, the absolute percentage difference between C_{Fitted} and C_{Exact} (namely $\left|\frac{C_{\text{Exact}} - C_{\text{Fitted}}}{C_{\text{Exact}}}\right|$) is no larger than $\sim 1\%$ for any value of R, and of the order of 0.1% for most values of R.

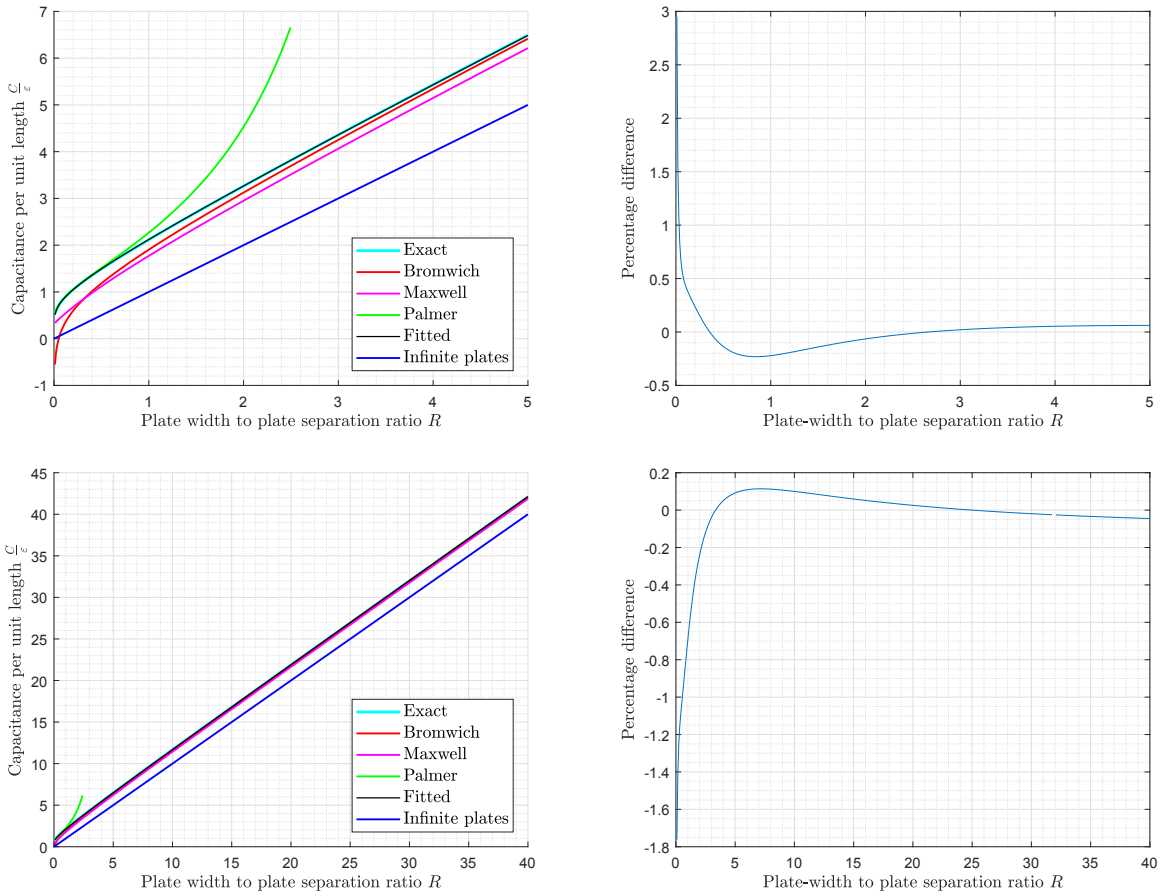


Figure 11 – The exact capacitance C_{Exact} (Eq. 72) was fitted over the range $0 \leq R \leq 5$ in (a) and (b), and over the range $0 \leq R \leq 40$ in (c) and (d), resulting in an approximation C_{Fitted} (Eq. 79). The various approximations C_{Maxwell}, C_{Bromwich}, and C_{Palmer} (Eq. 73, Eq. 74, and Eq. 75) of the exact fringing field capacitance C_{Exact} are plotted in (a) and (c), and compared to C_{Fitted}. The percentage differences $\frac{C_{\text{Exact}} - C_{\text{Fitted}}}{C_{\text{Exact}}}$ are plotted in (b) and (d).

Fit Range	a	b	Adjusted r^2
$0 \leq R \leq 5$	1.1205 ± 0.0002	0.8243 ± 0.0001	$1 - 3.5297 \times 10^{-6}$
$0 \leq R \leq 40$	1.1305 ± 0.0008	0.8306 ± 0.003	$1 - 6.6322 \times 10^{-7}$

Table 1 – Values of the fitting parameters a and b of Eq. 79, and corresponding fits coefficients of determination r^2 adjusted for the number of degrees of freedom. The resulting fits are plotted in Fig. 11.

3.1.1.8 Extensions of the fringing field capacitance formula

In deriving Eq. 72, we have assumed that the plates of the capacitor were finite along one dimension, but infinite along the other. What if the plates were, instead, finite along both dimensions? Consider such a capacitor with rectangular plates. Let w and t denote respectively, as before, the width of the plates of the capacitor and their separation, and l their length. Define $R_1 = \frac{w}{t}$ and $R_2 = \frac{l}{t}$. Then, to a first-order approximation and according to Eq. 79, the capacitance of the rectangular plates capacitor is given by:

$$C = \frac{\varepsilon w l}{t}\left(1 + \frac{a}{R_1^b}\right)\left(1 + \frac{a}{R_2^b}\right) \tag{80}$$

where a and b are defined in Tab. 1. The capacitance in Eq. 80 only takes into account field lines in planes orthogonal to the edges of the capacitor plates, and thus ignores field lines curling around the vertices of the plates. To take into account all fringing fields in the capacitance calculation, the conformal mappings that we have introduced in Sec. B can be extended to have for domains subsets of $\mathbb{C}^2$ rather than $\mathbb{C}$. A procedure similar to that exposed in Sec. 3.1.1.6 can then be carried out to compute the capacitance. Even more complicated geometries can also be handled by means of conformal mappings [86].

In the TB-NEGF simulations upon which we base our compact modeling, periodicity was assumed in the z dimension. Thus, Eq. 59 effectively describes the channel 2DEG of a MOSFET. The results obtained in Sec. 3.1.1.6 and Sec. 3.1.1.7 are therefore appropriate.

We mention two additional potential corrections to Eq. 72 for the modeling of the gate capacitance of MOSFETs. First, we have assumed the capacitor's plates to have infinitesimal thickness. While this assumption is reasonable for the channel 2DEG, it is decidedly inaccurate for the gate electrode. Second, we have assumed the permittivity ε to be constant throughout space. This assumption is inaccurate, as the oxide layer of a MOSFET is typically restricted to the region between the gate electrode and the semiconductor channel. We believe both effects to be relatively unimportant. Indeed, the corrections induced by these effects manifest themselves in the field lines above the top plate and below the bottom plate. As can be seen in Fig. 10, the density of those field lines is comparatively small, regardless

of the value of R.

Finally, we note that we have assumed a two-plate capacitor thus far. Nevertheless, our final goal is to model the gate capacitance of double-gated MOSFETs. Thanks to the linearity of Poisson's equation, this can be achieved by superposition. Assuming the two gates and two oxide layers to be identical, it suffices to multiply Eq. 79 by 2.

3.1.2 Wavefunction penetration

Quantum mechanically, the wavefunctions of charge carriers can penetrate into the oxide layer, thus effectively reducing the thickness t_o of the oxide layer [87]. This effect can be modeled by an "effective penetration length" t_p, a parameter commensurate to the spread of carriers' wavefunctions within the oxide layer, such that:

$$t_o \rightarrow t_o - t_p \tag{81}$$

Within the effective mass approximation [73], the Schrödinger-Poisson algorithm [88] can be used to compute these wavefunctions along axes orthogonal to the channel 2DEG [89]. In cases, where the effective mass approximation in not valid, such as for very thin oxide layers and semiconducting channels, one must rely on more elaborate formalisms (see Sec. 1.2). Regardless of the procedure being used, these wavefunctions are prohibitive to compute within the context of compact modeling. Furthermore, even after such a calculation, it would not be obvious how to extract from wavefunctions a penetration length that accurately models the gate capacitance through Eq. 81.

Note that the t_p depends, in principle, on V_{GS}. Indeed, V_{GS} controls the shape of the inversion layer and the inversion charge density, both of which are directly related to the wavefunctions of carriers in the channel. Moreover, through DIBL, t_p also depends upon V_{DS}.

Wavefunction penetration effects are not of concern for our purposes. Indeed, the TB-NEGF calculations upon which we base our compact modeling restricted the wavefunctions

of charge carriers to lie entirely within the semiconductor channel.

3.1.3 Quantum capacitance

In Sec. 3.1.1 and Sec. 3.1.2, we have exposed how fringing fields and wavefunction penetration effects might influence our classical notion of capacitance. We will hereby refer to the capacitance calculated from the considerations above as the electrostatic capacitance C_E.

However, in the limit where (at least) one of the plates of the capacitor under consideration has a low density of states, the electrostatic capacitance C_E alone is not a suitable descriptor for the gate capacitance C_G. Indeed, as the charge on the plates of the capacitor increases, the negatively charged plate gains high-energy electrons, while the positively charged plate loses low-energy electrons. Thus, in the low density of states limit, the bandstructure as well as the band-filling of the plates must be carefully examined in order to properly describe the charging energy of the capacitor. This inherently quantum phenomenon, which is ultimately due to the Pauli exclusion principle, has the net effect of decreasing the capacitance, as if there were another capacitor in series with C_E. We will refer to the capacitance of this additional virtual capacitor as the quantum capacitance C_Q. The concept of quantum capacitance was first introduced by Luryi [90].

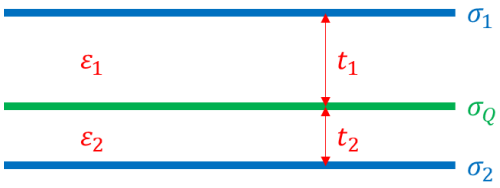


Figure 12 – Diagram of the three-plate capacitor described in Sec. 3.1.3. The top and bottom plates (drawn in blue) are metallic, and the middle plate (drawn in green) has low density of states at the Fermi level.

Consider a capacitor formed by three parallel infinite plane plates. Two of these plates, the bottom and top ones, are metallic, while the middle plate has a low density of states.

The top and middle plates are separated by a dielectric of permittivity ε_1 and thickness t_1, while the middle and bottom plates are separated by a dielectric of permittivity ε_2 and thickness t_2. We further assume that the middle and bottom plates are grounded. This capacitor is illustrated in Fig. 12. According to Eq. 58, the electrostatic capacitances (per unit area) of the capacitors formed by the top and middle plates on the one hand, and the middle and bottom plates on the other hand, are respectively:

$$\begin{aligned} C_1 &= \frac{\varepsilon_1}{t_1} \\ C_2 &= \frac{\varepsilon_2}{t_2} \end{aligned} \tag{82}$$

Now, let σ_1, σ_2, and σ_Q respectively be the areal charge densities on the top, bottom and middle plates. By charge neutrality, one has:

$$\sigma_1 + \sigma_2 + \sigma_Q = 0 \tag{83}$$

This condition may be parametrized by an angle $0 \leq \phi < \pi$ as:

$$\begin{aligned} \sigma_2 &= -\sigma_1 \sin^2(\phi) \\ \sigma_Q &= -\sigma_1 \cos^2(\phi) \end{aligned} \tag{84}$$

The electric field energies (per unit area) [85] above and below the middle plate are respectively U_1 and U_2, where for $i = 1, 2$:

$$U_i = \frac{\varepsilon_i}{2} \int_0^{t_i} \left(\frac{\sigma_i}{\varepsilon_i}\right)^2 dt = \frac{\sigma_i^2 t_i}{2\varepsilon_i} \tag{85}$$

The Fermi degeneracy energy [91], ignoring electronic correlation effects, is given by:

$$U_Q = \frac{\sigma_Q^2}{2\rho(E_F)\, e^2} \tag{86}$$

where $\rho(E_F)$ denotes the density of states (per unit area) of the electron gas in the middle plate at the Fermi level. Note that Eq. 86 is only valid at zero temperature. Assuming that the middle plates is composed of a nondegenerate semiconductor 2DEG, under the free electron approximation, the density of states is given by:

$$\rho = \frac{g_S g_V \overline{m}^\star}{2\pi\hbar^2} \tag{87}$$

where g_S, g_V, and $\overline{m}^\star$ respectively denote the spin degeneracy, valley degeneracy, and density of states effective mass of the 2DEG.

Thus, the total energy (per unit area) of the system is:

$$U = U_1 + U_2 + U_Q = \sigma_1^2 \left[\frac{t_1}{2\varepsilon_1} + \frac{t_2}{2\varepsilon_2} \sin^4(\phi) + \frac{\pi\hbar^2}{g_S g_V \overline{m}^\star e^2} \cos^4(\phi) \right] \tag{88}$$

By energy minimization, we obtain:

$$\frac{dU}{d\phi} = 0 \implies \tan^2(\phi) = \frac{2\pi\hbar^2\varepsilon_2}{g_S g_V e^2 \overline{m}^\star t_2} = \frac{C_2}{C_Q} \tag{89}$$

where:

$$C_Q = \frac{g_S g_V e^2 \overline{m}^\star}{2\pi\hbar^2} \tag{90}$$

Using trigonometric identities, it follows that:

$$\begin{aligned} \sigma_Q &= -\frac{C_Q}{C_2 + C_Q}\sigma_1 \\ \sigma_2 &= -\frac{C_2}{C_2 + C_Q}\sigma_1 \end{aligned} \tag{91}$$

Eq. 91 implies that the circuit under consideration can be viewed as the equivalent three capacitor circuit illustrated in Fig. 13.

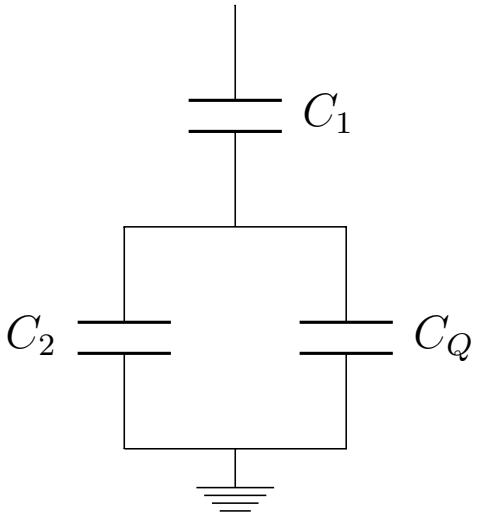


Figure 13 – Circuit diagram equivalent to the three-plate capacitor illustrated in Fig. 12.

Eq. 90 assumes a constant density of states, as well as an electronic dispersion relation well-described by an effective mass tensor. Several materials of interest in nanotechnologies, such as graphene, do not satisfy these properties, and thus have quantum capacitances [92] that cannot be described by Eq. 90. The assumptions underlying Eq. 90 can easily be relaxed, following a derivation almost identical to ours, leading to a more general expression for the quantum capacitance:

$$C_Q = e^2 \rho \left(E_F \right) \tag{92}$$

Eq. 86 can further be generalized to non-zero temperatures, leading to the following expression for the quantum capacitance:

$$C_Q = e^2 \int_{-\infty}^{+\infty} \left(-\frac{\partial f \left(E - E_F \right)}{\partial E} \right) \rho \left(E \right) dE \tag{93}$$

where f denotes the Fermi-Dirac distribution $f \left(E \right) = \frac{1}{2} \left[1 - \tanh \left(\frac{E}{2} \right) \right]$.

We note that in a MOSFET, the 2DEG in the semiconductor channel has low density of states, being nondegenerate. The nondegenerate approximation only fails for extremely high doping densities in the channel. Thus, the concept on quantum capacitance is important for the modeling of MOSFETs. From Fig. 13, setting C_1 to be the electrostatic gate capacitance C_E, and taking the limit $C_2 \to 0$, we find that the net gate capacitance of a MOSFET is given by:

$$C_G^{-1} = C_E^{-1} + C_Q^{-1} \tag{94}$$

It is thus apparent from Eq. 82 and Eq. 92 that quantum capacitance effects only become important in MOSFETs when either:

- The capacitor oxide layer has very high permittivity or very small thickness, so that the electrostatic capacitance C_E is large.
- The density of states in the semiconductor channel at the Fermi level E_F is very low, so that the quantum capacitance C_Q is low.

3.2 Threshold voltage

There does not exist a precise and universal definition for the threshold voltage V_T of a transistor. However, there is a universal, albeit somewhat ambiguous, notion of what V_T should be. In the subthreshold regime, the dependence of the drain current upon the gate voltage is exponential, whereas it is algebraic above threshold. The threshold voltage is the value of the gate voltage at the crossover between these two regimes. Ultimately, V_T is used to model the (absolute value of the) charge Q in the channel of a MOSFET above threshold through the relation:

$$Q = C_G \left(V_{GS} - V_T \right) \tag{95}$$

where C_G is the gate capacitance, and $V_{GS} > V_T$ the gate voltage.

In this section, we introduce 7 threshold voltage extraction methods [93] utilizing MOSFET transfer characteristics. We consider the first five of these methods to be arbitrary and prone to numerical instabilities, and will hereby refer to them as numerical methods. It is nevertheless important to introduce these methods as they are very commonly used in the semiconductor industry. On the other hand, the last two methods are based on physical arguments, and are more stable under deviations from exact transfer characteristics. We will hereby refer to these two methods as physics-driven methods.

3.2.1 Numerical methods

3.2.1.1 Constant current method

The constant current method [94] is the simplest threshold voltage extraction method, and arguably the most commonly used method by the semiconductor industry. It consists in defining the threshold voltage V_T as the value of the gate voltage corresponding to specified, arbitrary value of the drain current I_{CC}:

$$V_T = I_{DS}^{-1} \left(I_{CC} \right) \tag{96}$$

Each researcher may use a different value of I_{CC}. However, to make the constant current method less arbitrary, the following value of I_{CC}, depending on the device geometry, has been suggested [95]:

$$I_{CC} = \frac{10^{-7}\ \text{A}}{L} \tag{97}$$

where L represent the channel length. It is this value of the constant current that we will use in our analysis.

The constant current method is illustrated in Fig. 14, where it has been applied to the transfer characteristic of a 5 nm ML-BP with transport direction ZD, and biased at $V_{DS} = 1$ mV. The extracted value of the threshold voltage is:

$$V_T = 1.25 \pm 0.03\ \text{V} \tag{98}$$

3.2.1.2 Maximal slope method

The maximal slope method [96], also known as the extrapolation in the linear regime method, consists in finding the tangent to the transfer characteristic with maximal slope, and then finding the intersection of this tangent with the gate voltage axis. The threshold voltage is then defined as the gate voltage at this intersection point.

The maximal slope method is illustrated in Fig. 15, where it has been applied to the transfer characteristic of a 5 nm ML-BP with transport direction ZD, and biased at $V_{DS} =$ 1 mV. The extracted value of the threshold voltage is:

$$V_T \sim 0.74\ \text{V} \tag{99}$$

Note that the value of the threshold voltage in Eq. 99 is quoted without uncertainties. Indeed, V_T is extracted using numerical differentiation, a procedure which leads to very high uncertainties.

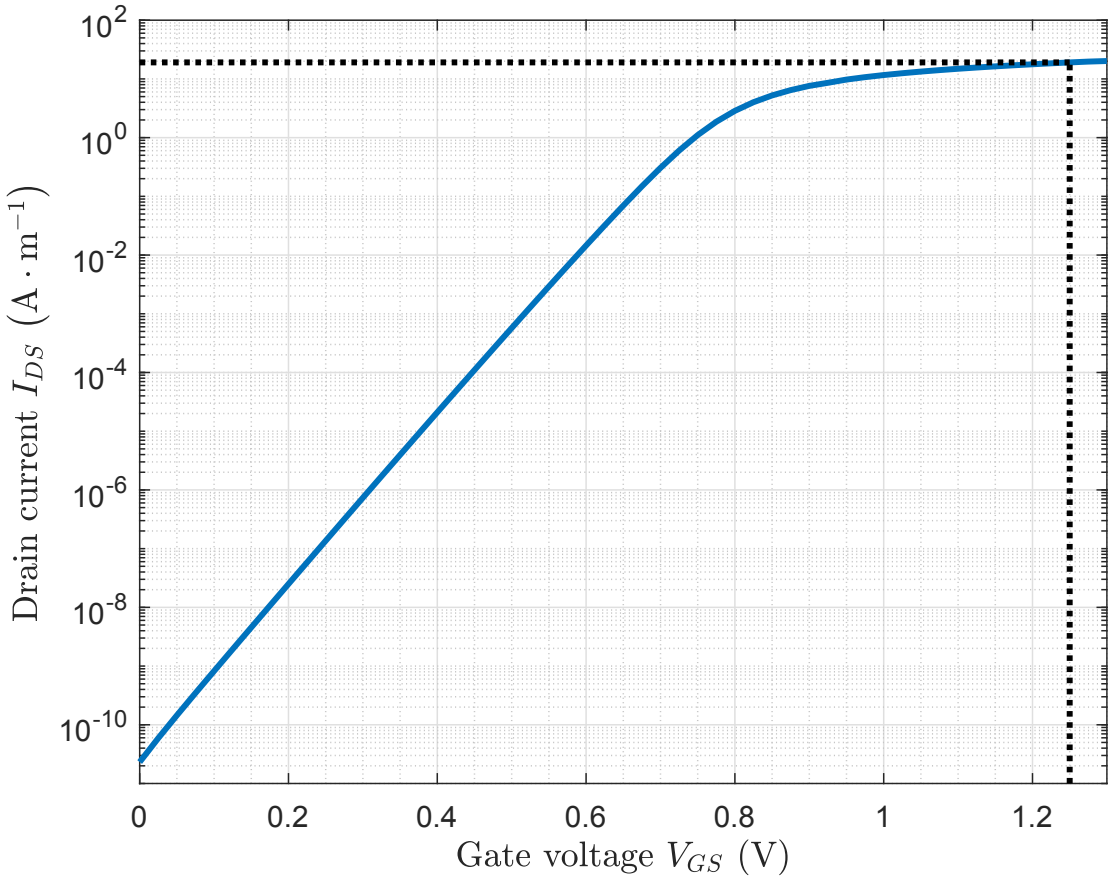


Figure 14 – Constant current method for threshold voltage extraction. The blue curve represents the transfer characteristic of a 5 nm channel ZD ML-BP MOSFET biased at $V_{DS} = 1$ mV. The horizontal dotted line represents the constant current $I_{CC} = 10\ \mathrm{A}\cdot\mathrm{m}^{-1}$, and the vertical dotted line represents the associated gate voltage, corresponding to the extracted threshold voltage.

3.2.1.3 Transconductance method

The transconductance method [95], also known as the transconductance extrapolation method in the linear regime, is similar to the maximal slope method. It consists in finding the tangent to the transconductance (defined as $g_m = \frac{\partial I_{DS}}{\partial V_{GS}}$) with maximal slope, and then finding the intersection of this tangent with the gate voltage axis. The threshold voltage is then defined as the gate voltage at this intersection point.

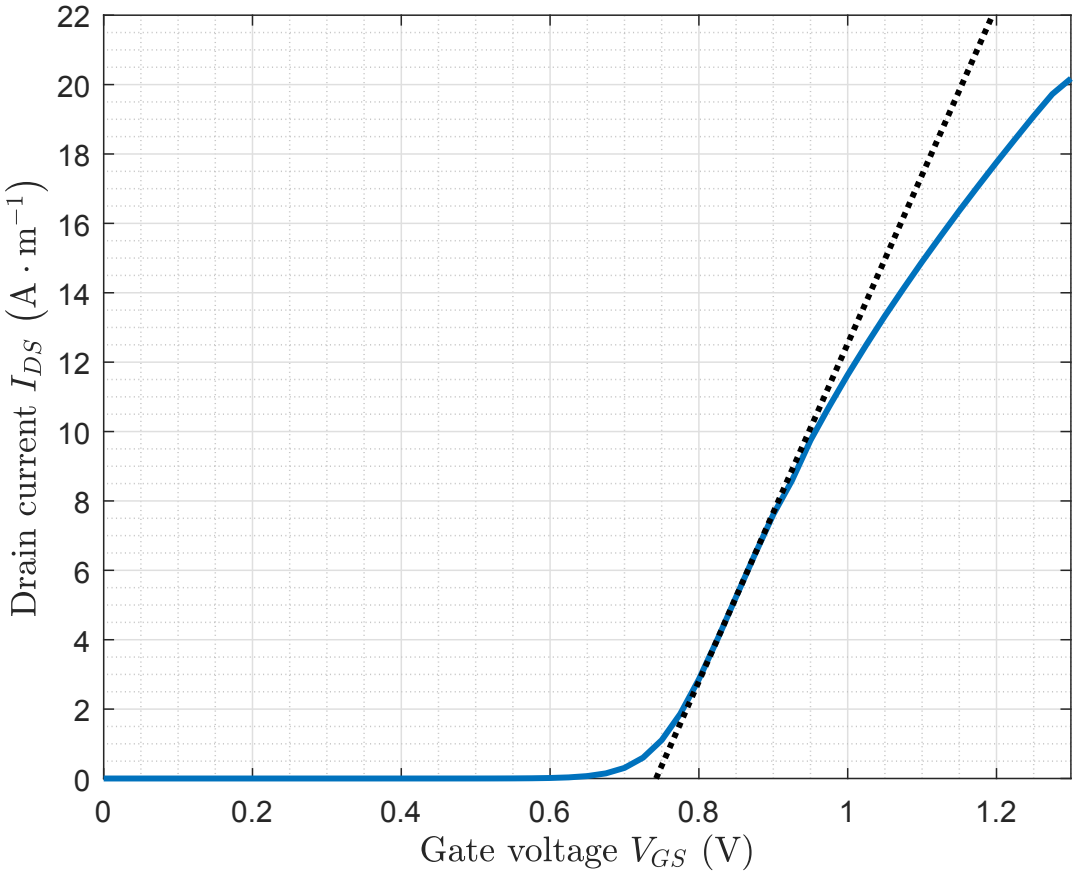


Figure 15 – Maximal slope method for threshold voltage extraction. The blue curve represents the transfer characteristic of a 5 nm channel ZD ML-BP MOSFET biased at $V_{DS} = 1$ mV. The dotted line represents the tangent to the transfer characteristic where its derivative is maximal. The intersection of the dotted line with the axis of abscissas corresponds to the extracted threshold voltage.

The transconductance method is illustrated in Fig. 16, where it has been applied to the transfer characteristic of a 5 nm ML-BP with transport direction ZD, and biased at $V_{DS} = 1$ mV. The extracted value of the threshold voltage is:

$$V_T \sim 0.68 \text{ V} \tag{100}$$

Note that the value of the threshold voltage in Eq. 100 is quoted without uncertainties, for identical reasons as those mentioned after Eq. 99.

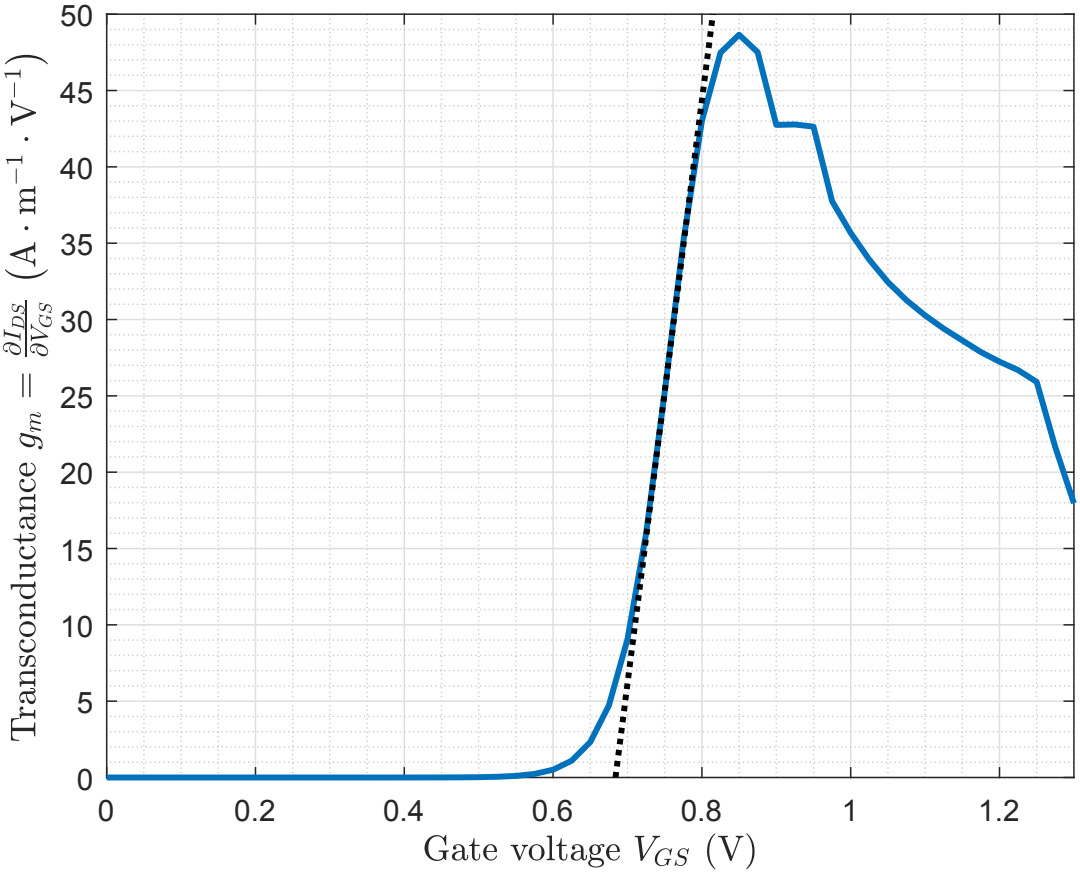


Figure 16 – Transconductance method for threshold voltage extraction. The blue curve represents the transconductance of a 5 nm channel ZD ML-BP MOSFET biased at $V_{DS} = 1$ mV. The dotted line represents the tangent to the transconductance where its derivative is maximal. The intersection of the dotted line with the axis of abscissas corresponds to the extracted threshold voltage.

3.2.1.4 Second derivative method

The second derivative method [97] consists in defining the threshold voltage as the gate voltage at which the second derivative of the transfer characteristic is maximal. It is interesting to note that the second derivative method is guaranteed to produce results distinct from the transconductance method (they would produce identical result only if the second derivative of the transfer characteristic were to diverge to ∞).

The second derivative method is illustrated in Fig. 17, where it has been applied to the transfer characteristic of a 5 nm ML-BP with transport direction ZD, and biased at $V_{DS} = 1$ mV. The extracted value of the threshold voltage is:

$$V_T = 0.75 \pm 0.03 \text{ V} \tag{101}$$

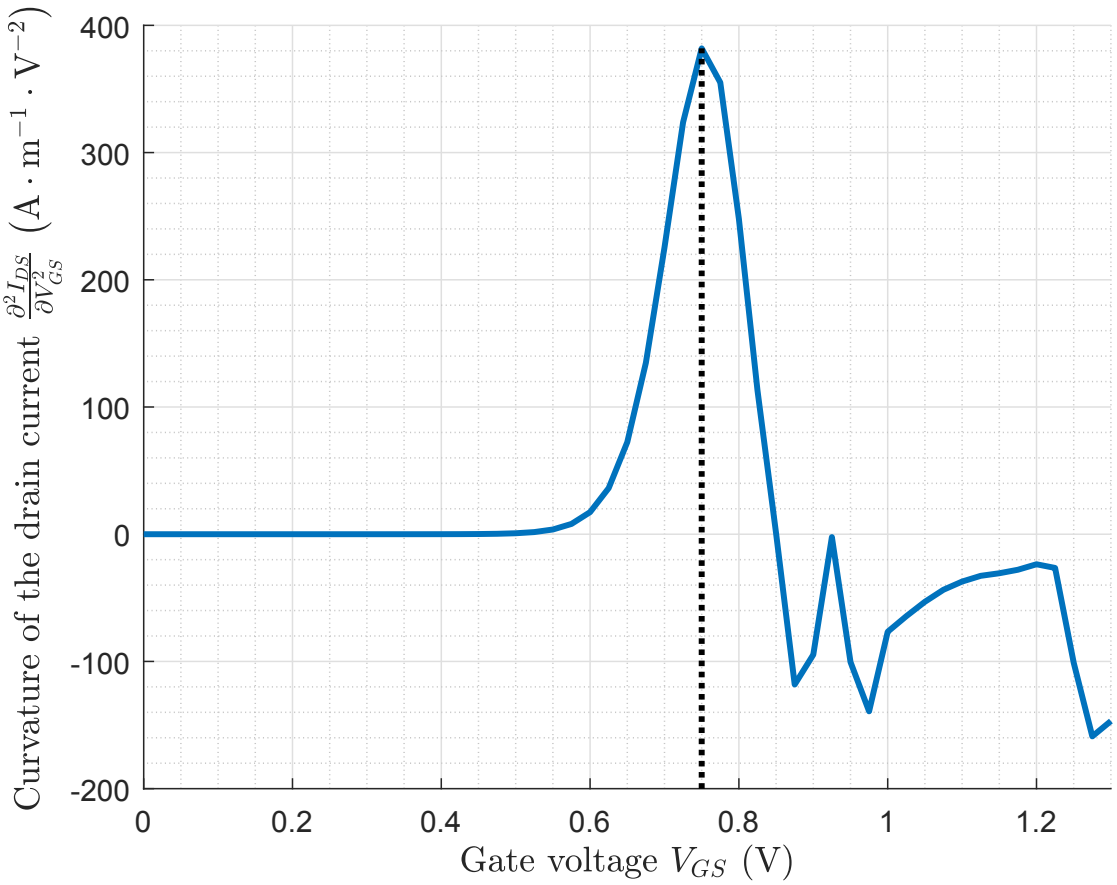


Figure 17 – Second derivative method for threshold voltage extraction. The blue curve represents the second derivative of the transfer characteristic of a 5 nm channel ZD ML-BP MOSFET biased at $V_{DS} = 1$ mV. The dotted line represents the gate voltage at which the second derivative of the transfer characteristic is maximal, corresponding to the extracted threshold voltage.

3.2.1.5 Second derivative of the logarithm method

The second derivative of the logarithm method [98] is similar to the second derivative method. It consists in defining the threshold voltage as the gate voltage at which the second derivative of the logarithm of the transfer characteristic is minimal.

The second derivative of the logarithm method is illustrated in Fig. 18, where it has been applied to the transfer characteristic of a 5 nm ML-BP with transport direction ZD, and biased at $V_{DS} = 1$ mV. The extracted value of the threshold voltage is:

$$V_T = 0.78 \pm 0.03 \text{ V} \tag{102}$$

3.2.2 Physics-driven methods

3.2.2.1 Ballistic mobility method

The ballistic mobility method is a linear optimization-based procedure that we propose for the extraction of threshold voltage from transfer characteristics of nondegenerate MOSFETs for drain voltages V_{DS} much smaller than the thermal voltage $\frac{k_B T}{e}$. It can be thought of as a combination of Karlsson's nonlinear optimization method for threshold voltage extraction [99] and of the ballistic mobility formalism [78]. The idea is the following: from Sec. 2.4.3, the current in a ballistic MOSFETs for gate voltages V_{GS} larger than the threshold voltage V_T and drain voltages V_{DS} smaller than the thermal voltage $\frac{k_B T}{e}$ can be expressed as:

$$I_{DS} = W C_G \left(V_{GS} - V_T\right) \mu_{\text{ball}} \frac{V_{DS}}{L} \tag{103}$$

where $\mu_{\text{ball}} = \frac{e v_T L}{2 k_B T}$ is the ballistic mobility (Eq. 51) and $v_T = \sqrt{\frac{2 k_B T}{\pi m_t^\star}}$ is the unidirectional thermal velocity (Eq. 45).

The drain current is seen to be a linear function of V_{GS} under the assumptions of validity of Eq. 103. We note that the assumption of nondegeneracy is valid under relatively low values of V_{DS} and V_{GS}, as in such cases, the barrier height is high compared to the

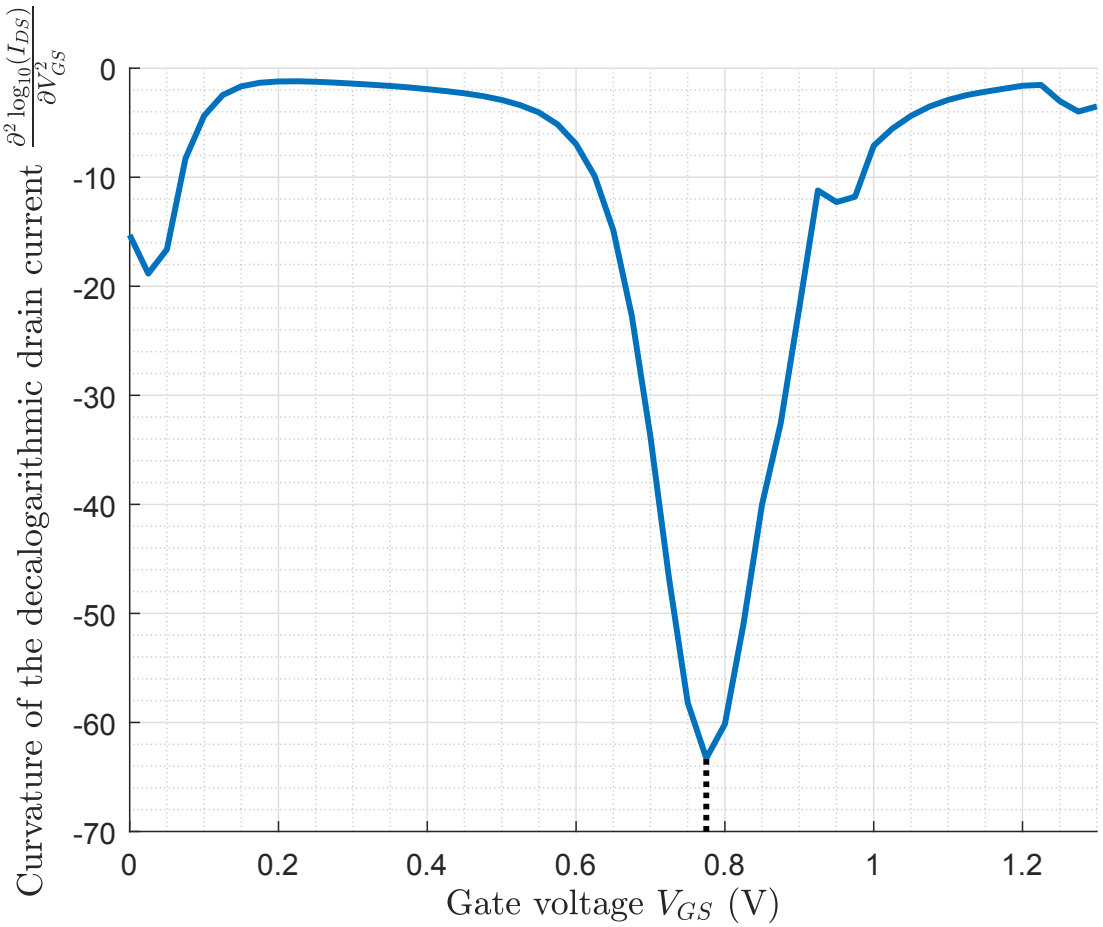


Figure 18 – Second derivative of the logarithm method for threshold voltage extraction. The blue curve represents the second derivative of the base 10 logarithm of the transfer characteristic of a 5 nm channel ZD ML-BP MOSFET biased at $V_{DS} = 1$ mV. The dotted line represents the gate voltage at which the second derivative of the base 10 logarithm of the transfer characteristic is maximal, corresponding to the extracted threshold voltage.

source Fermi level. The drain current I_{DS} should then indeed be a linear function of V_{GS} for $V_{GS} \gtrsim V_T$. Deviations from this linear behaviour should only arise when $V_{GS} \gg V_T$, where the assumption of nondegeneracy fails. This is precisely what we observe in Fig. 19. Thus, by performing a linear fit to transfer characteristics, one can not only extract the threshold voltage V_T, but also the gate capacitance C_G. The extracted values of the latter parameter are presented in Tab. 2.

The ballistic mobility method is illustrated in Fig. 19, where it has been applied to the

Transport direction	Armchair		Zigzag	
Channel length	10 nm	5 nm	10 nm	5 nm
C_G ($\mathrm{mF \cdot m^{-2}}$)	43.2 ± 0.1	41.4 ± 0.2	44.16 ± 0.08	43.60 ± 0.06

Table 2 – Values of the gate capacitance for the four studied ML-BP MOSFETs extracted from their $V_{DS} = 1$ mV transfer characteristics using the ballistic mobility method.

transfer characteristic of a 5 nm ML-BP with transport direction ZD, and biased at $V_{DS} = 1$ mV. The extracted values of the threshold voltage and gate capacitance are respectively:

$$\begin{aligned} V_T &= 0.742 \pm 0.002 \text{ V} \\ C_G &= 43.60 \pm 0.06 \text{ µF} \cdot \mathrm{m}^{-2} \end{aligned} \tag{104}$$

To obtain the desired linear fit, we start by isolating the subset of the transfer characteristic corresponding to values of V_{GS} between 0.7 V and 1 V into a vector D with N entries. We define $\mathbb{D}$ to be the set of all subsets of at least 3 adjacent entries of D. For example, if $D = \{1, 2, 3, 4\}$, then $\mathbb{D} = \{\{1, 2, 3\}, \{2, 3, 4\}, \{1, 2, 3, 4\}\}$. Note that the cardinality of $\mathbb{D}$ is $\frac{1}{2}(N^2 - 3N + 2)$. Linear fits are then performed on all elements of $\mathbb{D}$. The coefficients of determination (adjusted for the number of degrees of freedom) of all of those fits are computed. We define the minimal acceptable coefficient of determination to be $R^2_{\text{min}} = 0.9999$. The desired linear fit is then defined as the element of $\mathbb{D}$ with the greatest number of entries whose linear fit has a coefficient of determination greater than R^2_{min}. If several such elements of $\mathbb{D}$ exist, the one whose linear fit has the greatest coefficient of determination is chosen.

We note that the concept of ballistic mobility is not crucial in the present context. Indeed, Eq. 49 could be used directly, without invoking the ballistic mobility. Nevertheless, the analogy of our method with Karlsson's method is more apparent by using the ballistic mobility. We also note that while our method can be used to extract the value of the gate capacitance C_G, Karlsson's method (being applicable to non-ballistic MOSFET) can only be used to extract the product μC_G of the channel mobility μ and C_G. This is because the exact value of μ_{ball} can be readily calculated from Eq. 51, assuming the MOSFET channel 2DEG to be nondegenerate.

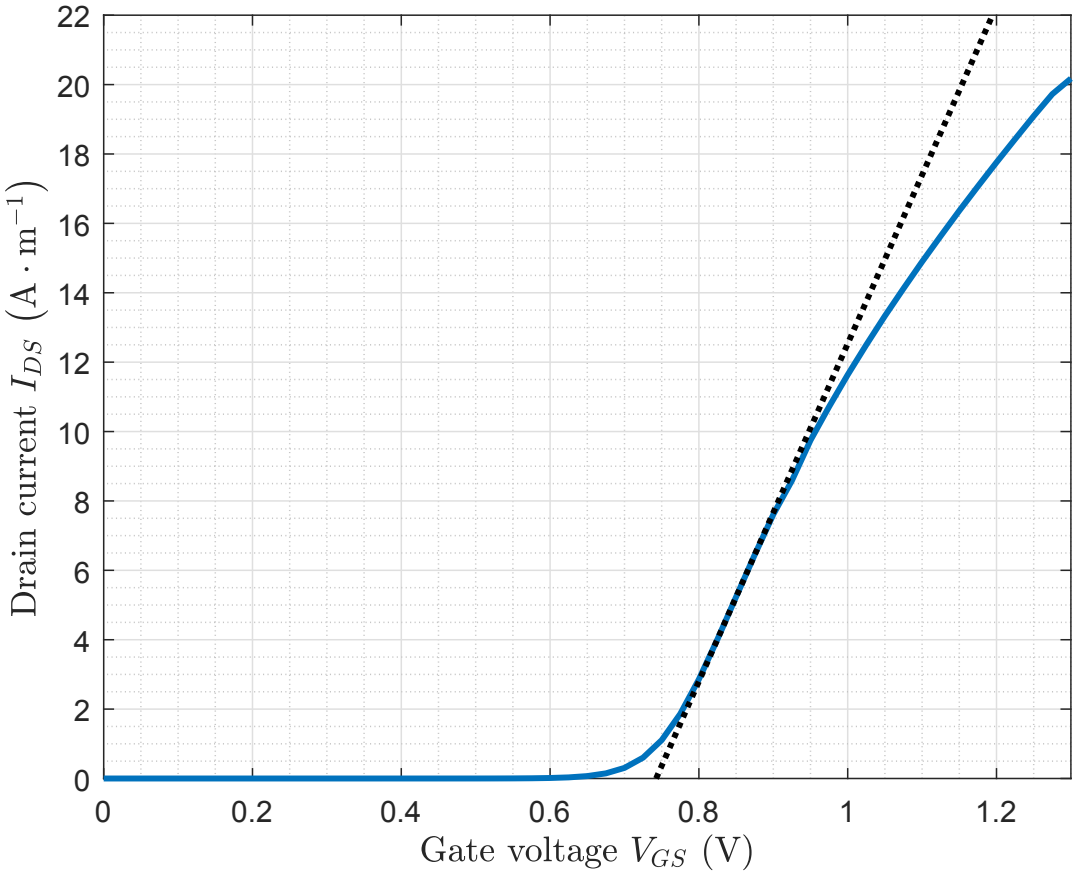


Figure 19 – Ballistic mobility method for threshold voltage extraction. The blue curve represents the transfer characteristic of a 5 nm channel ZD ML-BP MOSFET biased at $V_{DS} = 1$ mV. The dotted line represents the linear fit to the transfer characteristic. Its slope is $\frac{\mu_{\mathrm{ball}} C_G V_{DS}}{L}$, and its intersection with the axis of abscissas is the extracted threshold voltage.

3.2.2.2 Amorphous channel method

The amorphous channel method [93, 100, 101] was originally devised for the extraction of threshold voltage in non-crystalline MOSFETs, for which the saturation drain current in strong inversion can be modeled as:

$$I_{DS_{\mathrm{sat}}} = K\left(V_{GS} - V_T\right)^m \tag{105}$$

where K is a constant empirical parameter with units of $\mathrm{A}\cdot\mathrm{V}^{-m}$, and m is dimensionless exponent. As shown in Sec. 2.2, in the drift-diffusion limit, Eq. 105 best models long-channel

MOSFETs for $m = 2$, and shorter channel MOSFETs where velocity saturation effects are important for $m = 1$. In the case of ballistic transport (Sec. 2.3), Eq. 105 also correctly models the saturation current, with an exponent typically in the range $1 \leq m \leq 2$. As such, the amorphous channel method is also applicable to ballistic MOSFETs.

The idea behind the amorphous channel method is the following: first, consider the following function computed from the transfer characteristic in the saturation regime $I_{DS}(V_{GS})$:

$$H(V_{GS}) = \frac{1}{I_{DS}(V_{GS})} \int_{0}^{V_{GS}} I_{DS}(V_{GS})\, dV_{GS} \tag{106}$$

Since $I_{DS}(V_{GS}) \approx 0$ for $V_{GS} < V_T$, we can write, using Eq. 105, $\forall V_{GS} > V_T$:

$$\begin{aligned} H(V_{GS}) &= \frac{1}{I_{DS}(V_{GS})} \int_{V_T}^{V_{GS}} I_{DS}(V_{GS})\, dV_{GS} \\ H(V_{GS}) &= \frac{V_{GS} - V_T}{m+1} \end{aligned} \tag{107}$$

The function $H(V_{GS})$ is seen to be a linear function of V_{GS} for $V_{GS} > V_T$. Thus, by performing a linear fit of $H(V_{GS})$, the threshold voltage V_T and the exponent m can be computed. The extracted values of the latter parameter are presented in Tab. 3. We note that the extracted exponents m are all consistent with the ballistic transport formalisms presented in Sec. 2.3, in the sense that $1 \leq m \leq 2$. Furthermore, we note that the value of m seems to only depend upon the transport direction, with $m \approx 1.7$ along AD, and $m \approx 2$ along ZD. This is expected, since the MOSFETs under consideration in this thesis operate at the ballistic limit.

With regards to the linear fitting procedure, we use the same method as that exposed in Sec. 3.2.2.1, with two caveats. First, we only perform fits that end at the maximal value of V_{GS}, namely 1.3 V. This is because the approximation of Eq. 105 is more accurate for larger values of V_{GS}. Second, we use $R^2_{\text{min}} = 0.99$.

The amorphous channel method is illustrated in Fig. 20, where it has been applied to the transfer characteristic of a 5 nm ML-BP with transport direction ZD, and biased at

$V_{DS} = 0.6$ V. The extracted values of the threshold voltage and exponent are respectively:

$$\begin{aligned} V_T &= 0.67 \pm 0.01 \text{ V} \\ m &= 2.00 \pm 0.02 \end{aligned} \tag{108}$$

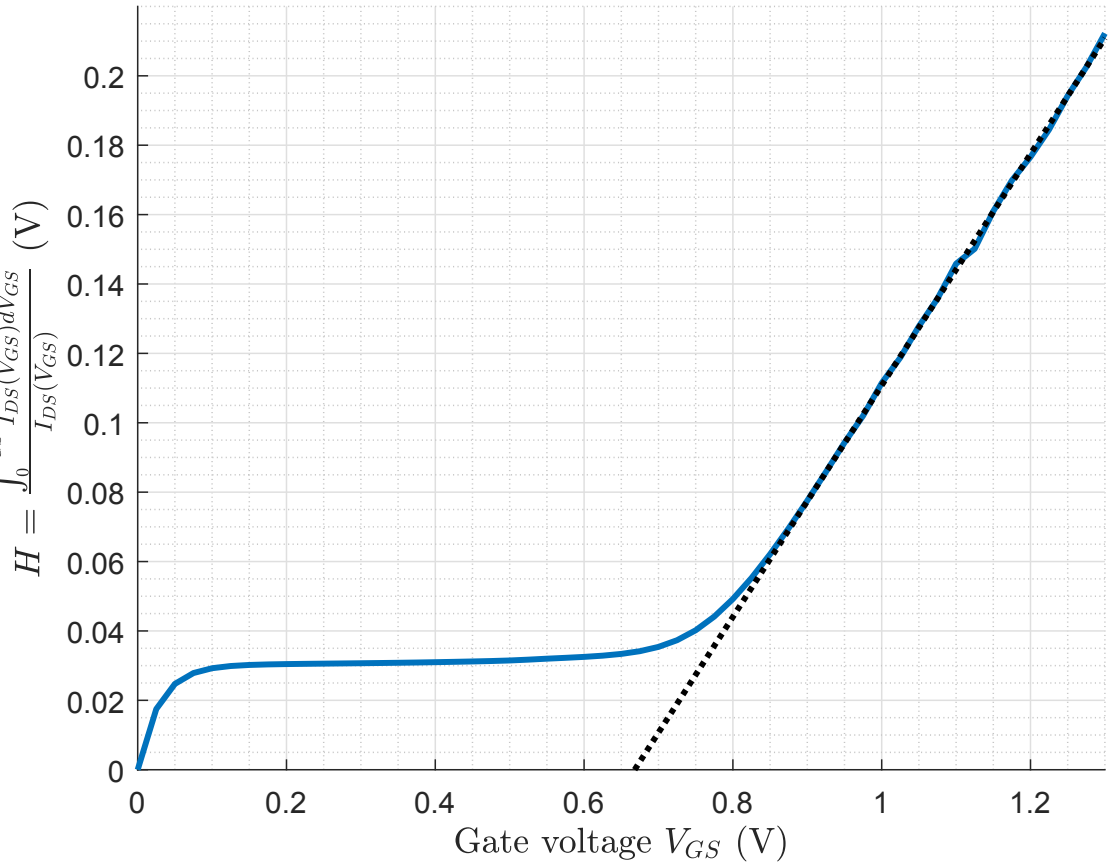


Figure 20 – Amorphous channel method for threshold voltage extraction. The blue curve represents the function $H\left(V_{GS}\right)$ defined in Eq. 106 obtained from the transfer characteristic of a 5 nm channel ZD ML-BP MOSFET biased at $V_{DS} = 0.6$ V. The dotted line represents the linear fit to the function $H\left(V_{GS}\right)$ for $V_{GS} > V_T$. Its slope is $\frac{1}{m+1}$, and its intersection with the axis of abscissas is the extracted threshold voltage.

Transport direction	Armchair		Zigzag	
Channel length	10 nm	5 nm	10 nm	5 nm
Extracted exponent m	1.68 ± 0.03	1.73 ± 0.04	1.97 ± 0.04	2.00 ± 0.02

Table 3 – Values of the exponent m for the four studied ML-BP MOSFETs extracted from their $V_{DS} = 0.6$ V transfer characteristics using the amorphous channel method.

3.2.3 Comparison of the methods and discussion

Except from the constant current method, all of the threshold voltage extraction methods presented in Sec. 3.2.1 rely on the calculation of numerical derivatives. Whether they are obtained from experimental measurements or from *in silico* simulations, transfer characteristics inevitably exhibit some amount of noise. Numerical derivatives are particularly sensitive to such noise. As such, threshold voltages extracted from the methods of Sec. 3.2.1 are inherently prone to inaccuracies. A potential solution to this problem would be to devise a numerical threshold voltage extraction method based on a procedure that is more numerically stable. The integral method [102] is an example of such a method, but was unfortunately found to be wildly unsuccessful in determining the threshold voltages of our devices.

On the other hand, the two physics-driven threshold voltage extraction methods presented in Sec. 3.2.2 rely on much more numerically stable calculations, namely numerical integration and linear fitting. Hence, we believe these methods to be most reliable.

In Tab. 4, we compare the values of the threshold voltage extracted using the 7 different methods that we have presented in Sec. 3.2.1 and Sec. 3.2.2, for the four different ML-BP MOSFETs that we study in this thesis. We note that for all devices, the maximal slope, second derivative, second derivative of the logarithm, and ballistic mobility methods produce results consistent with each other. By law of majority, it is thus likely that these four methods are most accurate in estimating V_T. The fact that the constant current method produces erroneous results is unsurprising, due to the blatant arbitrariness of this method. The

transconductance method is also seen to produce results inconsistent with those obtained from the other methods: as explained in Sec. 3.2.1.4, this method must produce results distinct from the second derivative method; we believe the second derivative method to be more reliable than the transconductance method. Finally, we note that the results of the amorphous channel method differ from those obtained from the other methods. This is expected. Indeed, the amorphous channel method is only applicable for transfer characteristics in the saturation regime (which was taken as $V_{DS} = 0.6$ V for the purpose of this analysis), whereas the other methods are only applicable in the linear regime [93] (which was taken to be $V_{DS} = 1$ mV). Since for higher drain voltages, DIBL reduces the effective threshold voltage, it should come as no surprise that the results obtained from the amorphous channel method are lower than those obtained from the other methods.

To summarize, we advocate for the ballistic mobility method and the amorphous channel method. Indeed, these methods are based on both robust physical derivations and numerical implementations. The ballistic mobility method is further seen to produce results similar to methods that stood the test of time within the semiconductor industry, notably the maximal slope method. Finally, it is important to recall that the ballistic mobility method is only applicable for $V_{DS} \ll \frac{k_B T}{e}$, while the amorphous channel method is only applicable in the saturation regime.

3.3 Subthreshold swing

3.3.1 Definition and theoretical value

The subthreshold swing of a transistor is defined as:

$$S = \left(\frac{\partial \log_{10} I_{DS}}{\partial V_{GS}} \right)^{-1} \tag{109}$$

where we assume the V_{DS} to be fixed. Thus, $\frac{1}{S}$ is simply the slope of the graph of the decalogarithmic drain current as a function of the gate voltage. Note that in the subthreshold regime, namely for $V_{GS} \ll V_T$, this slope is approximately constant, and it is the reciprocal

V_T extraction method	Armchair		Zigzag	
	10 nm	5 nm	10 nm	5 nm
Constant current	0.88 ± 0.03	0.93 ± 0.03	0.98 ± 0.03	1.25 ± 0.03
Maximal slope	~ 0.78	~ 0.76	~ 0.76	~ 0.74
Transconductance	~ 0.71	~ 0.67	~ 0.71	~ 0.68
Second derivative	0.78 ± 0.03	0.78 ± 0.03	0.78 ± 0.03	0.75 ± 0.03
Sec. der. of the log.	0.8 ± 0.03	0.75 ± 0.03	0.80 ± 0.03	0.78 ± 0.03
Ballistic mobility	0.775 ± 0.004	0.755 ± 0.005	0.754 ± 0.002	0.742 ± 0.002
Amorphous channel	0.75 ± 0.02	0.70 ± 0.02	0.71 ± 0.02	0.67 ± 0.01

Table 4 – Threshold voltage extracted from the 7 methods presented in Sec. 3.2.1 and Sec. 3.2.2, for the four different ML-BP MOSFETs that we study. All voltage values are quoted in units of V. The threshold voltages were all extracted from transfer characteristics at $V_{DS} = 1$ mV, except from those for the amorphous channel method, for which $V_{DS} = 0.6$ V.

of this value of the slope which is usually referred to as the subthreshold swing.

The subthreshold swing is a critical parameter which quantifies how to easily it is to turn ON and OFF a transistor, and is therefore commensurate with leakage currents and power dissipations. In MOSFETs, in which the main charge transport mechanism is thermionic emission, the subthreshold swing possesses a physical limit, which can be derived from electrostatics and thermodynamics [103]:

$$S = \log(10) \frac{k_B T}{\alpha_G} \tag{110}$$

where α_G denotes the gate control parameter $\alpha_G = \frac{C_G}{C_\Sigma}$, as introduced in Sec. 2.3.2. Hence, at room temperature, the subthreshold swing of a MOSFET must be greater than 60 mV/dec.

To derive this fundamental switching limit of MOSFETs, we first introduce the barrier height Φ_B, illustrated in Fig. 21, which equals to difference between the top of the barrier energy and the source Fermi level. In the subthreshold regime, $\Phi_B \gg k_B T$. The Fermi-Dirac distribution of electrons in the source can thus be approximated by an exponential

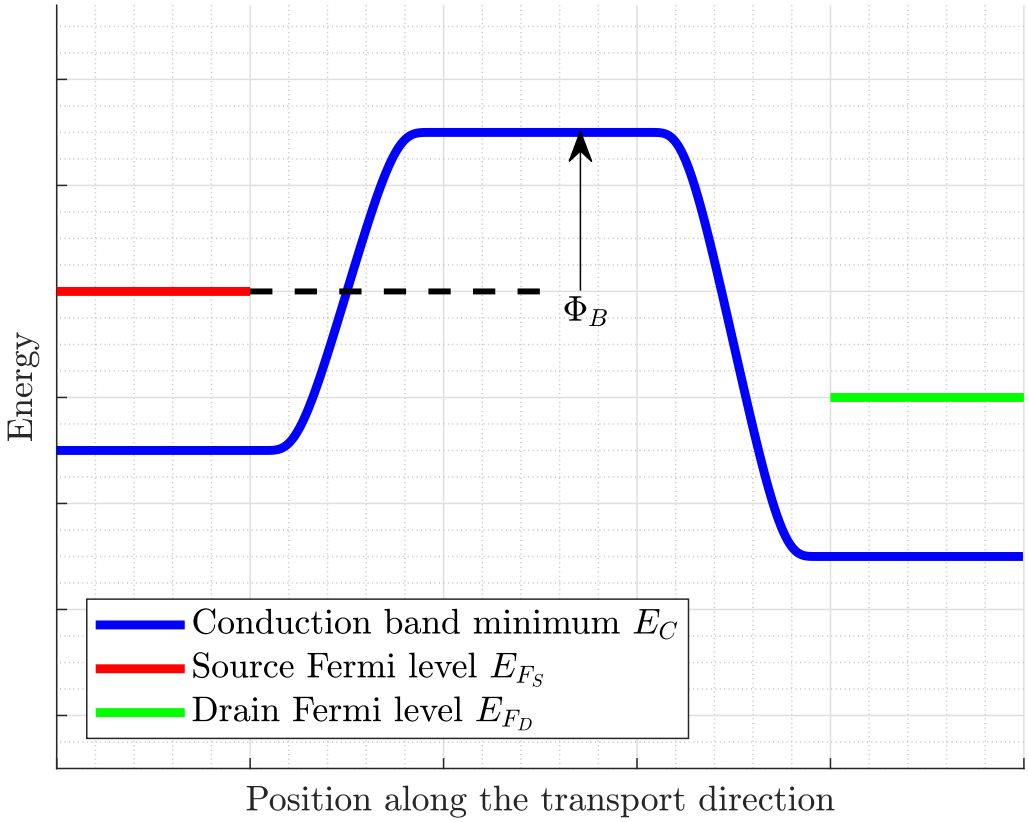


Figure 21 – Illustrative band diagram of a MOSFET in the subthreshold regime and pictorial definition of the barrier height Φ_B.

for energies of the order of Φ_B. Thus, given a slowly varying non-zero density of states for electrons in the source near Φ_B, the drain current, due to thermionic emission, satisfies [104]:

$$I_{DS} \propto \exp\left(-\frac{\Phi_B}{k_B T}\right) \tag{111}$$

We deduce that $\frac{\partial \log_{10} I_{DS}}{\partial \Phi_B} = -\frac{1}{\ln(10) k_B T}$. Furthermore, the barrier height Φ_B is controlled by the gate voltage V_{GS}. In the ideal case where the channel is solely capacitively coupled to the gate, $\frac{\partial \Phi_B}{\partial V_{GS}} = -e$. When this is not the case, instead, $\frac{\partial \Phi_B}{\partial V_{GS}} = -e\alpha_G$. Thus, by the chain rule, we obtain Eq. 110.

3.3.2 Computation from transfer characteristics

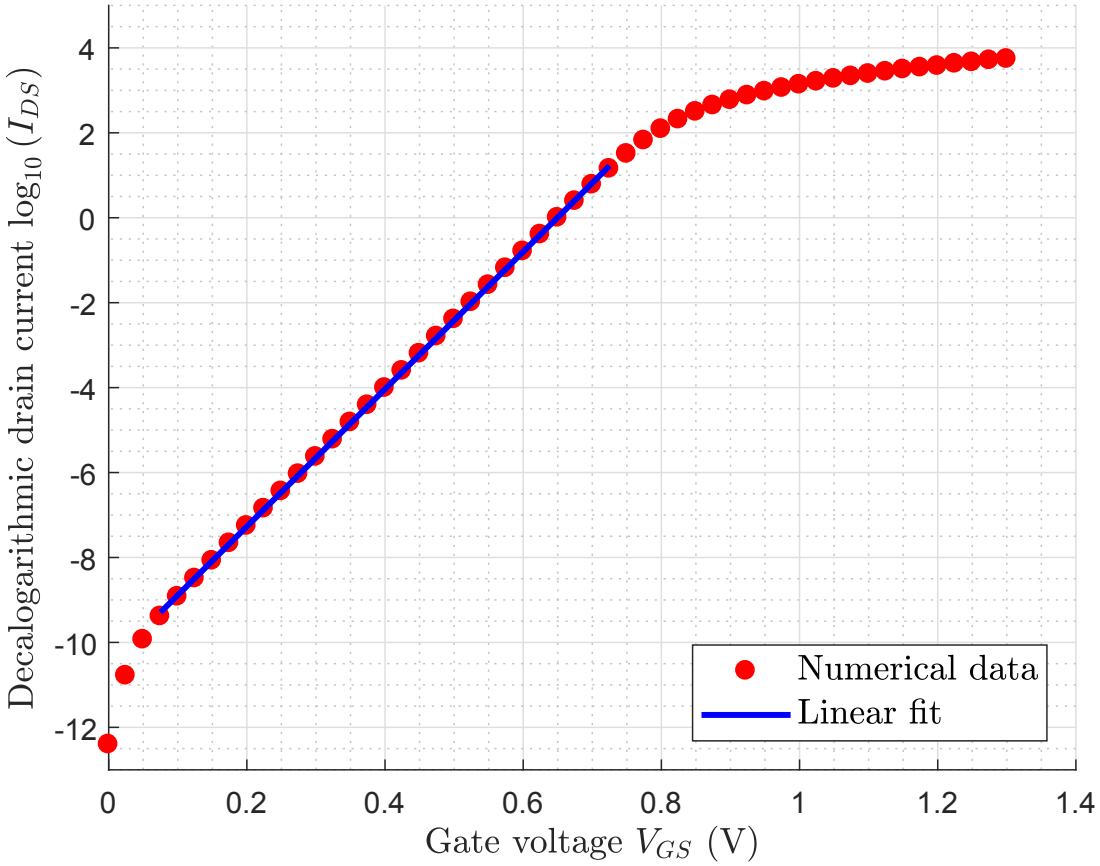
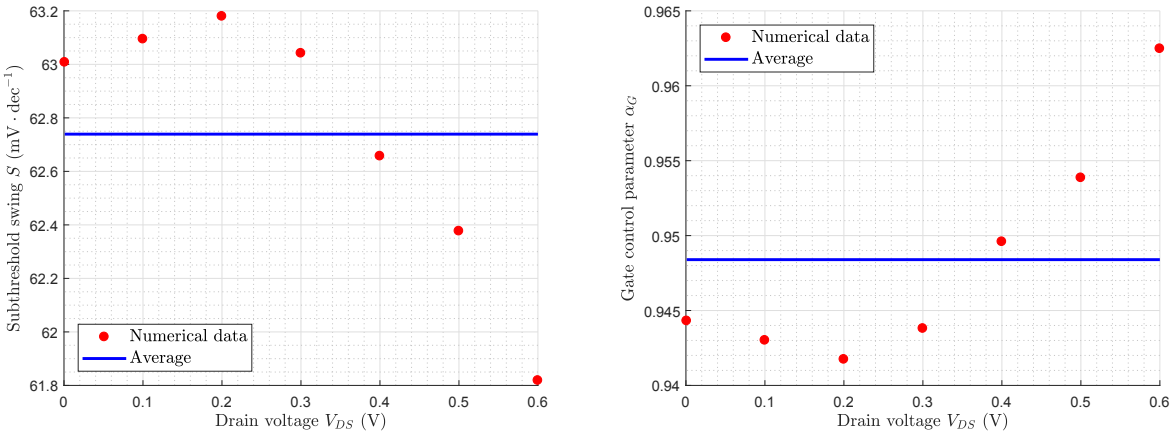


Figure 22 – Transfer characteristic of a 10 nm channel ML-BP MOSFET oriented along ZD at $V_{DS} = 0.6$ V. The red dots represent data computed from TB-NEGF simulations, and the blue line represents the line fitted to the data in the deep subthreshold regime. The slope of the fitted line was used to compute the subthreshold swing S and the gate control parameter α_G.

We extract the subthreshold swing S and the gate control parameter α_G directly from the transfer characteristics, using respectively Eq. 109 and Eq. 110. The linear fitting method used for this purpose is identical to that exposed in Sec. 3.2.2.1, and is illustrated in Fig. 22 for a 10 nm ZD ML-BP MOSFET biased at $V_{DS} = 0.6$ V, for which it yielded:

$$\begin{aligned} S &\approx 61.8 \text{ mV} \cdot \text{dec}^{-1} \\ \alpha_G &\approx 0.963 \end{aligned} \tag{112}$$



(a) Subthreshold swing S

(b) Gate control parameter α_G

Figure 23 – Subthreshold swing S (a) and gate control parameter α_G (b) of a 10 nm ML-BP MOSFET oriented along ZD. The red dots represent the values of S and α_G extracted from transfer characteristics for various values of the drain voltage V_{DS}, and the blue lines represents the means of those extracted values.

In Fig. 23, we plot the extracted values of S and α_G as a function of V_{DS}, for a 10 nm ZD ML-BP MOSFET. Both S and α_G are found to be roughly constant functions of V_{DS}. We also compute the mean of S and α_G across all values of V_{DS}. These means are reported in Tab. 5.

Transport direction	Armchair		Zigzag	
Channel length	10 nm	5 nm	10 nm	5 nm
$S\ \left(\text{mV} \cdot \text{dec}^{-1}\right)$	62.2 ± 0.3	80.1 ± 0.2	62.7 ± 0.5	67.8 ± 0.9
α_G	0.956 ± 0.004	0.743 ± 0.002	0.948 ± 0.008	0.88 ± 0.01

Table 5 – Mean extracted subthreshold swings S and gate control parameters α_G for the four MOSFETs studied.

3.4 Drain-induced barrier lowering

Drain-induced barrier lowering (DIBL) is the phenomenon by which the barrier height of a MOSFET is reduced by increasing the drain voltage V_{DS}. Being due to the existence of

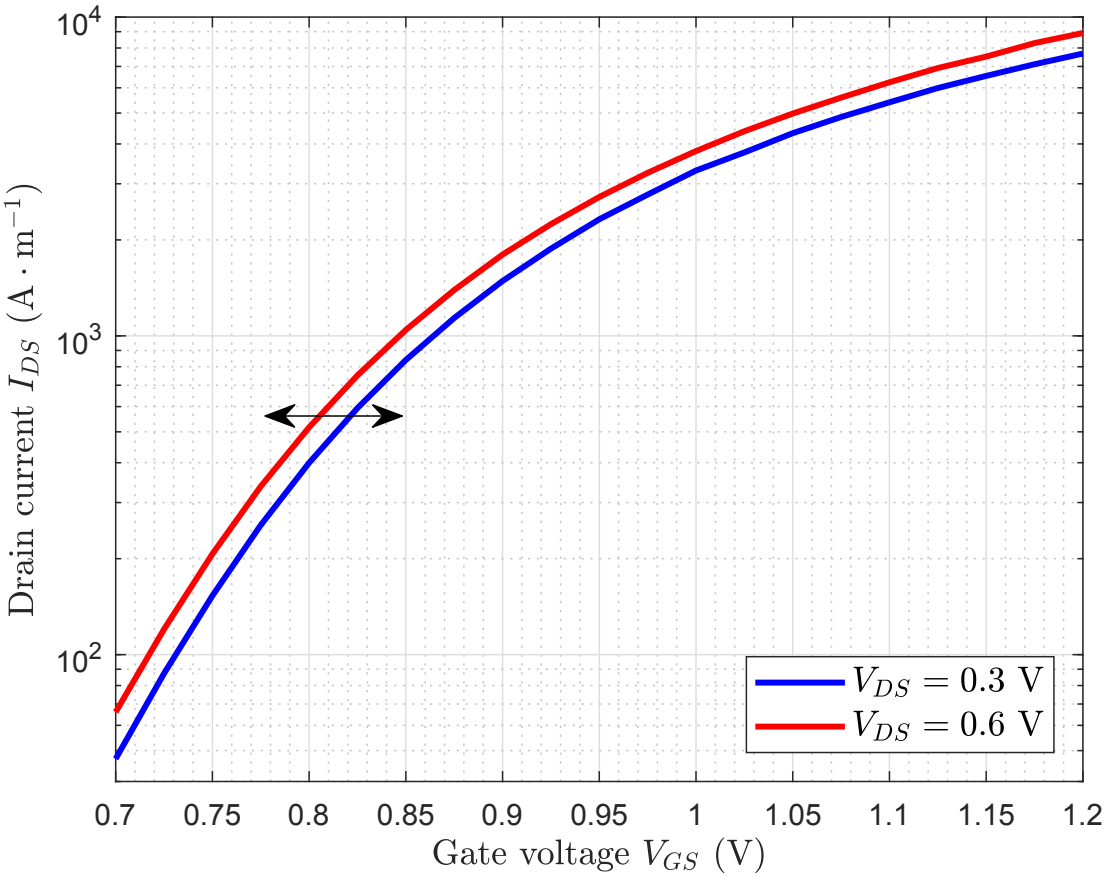


Figure 24 – Transfer characteristics of a 5 nm ML-BP oriented along AD, obtained from TB-NEGF simulations. Note that as V_{DS} is increased, transfer characteristics are shifted to the left. This phenomenon leads to an effective reduction of threshold voltage as V_{DS} is increased, and is known as DIBL.

a sizeable capacitive coupling between the drain electrode and the channel 2DEG, DIBL is typically more significant for MOSFET of small channel lengths. As illustrated in Fig. 24, one of the effects of DIBL is to shift transfer characteristics to the left as V_{DS} is increased. Equivalently, DIBL reduces the threshold voltage, through the relation:

$$V_T = V_{T_0} - \delta V_{DS} \tag{113}$$

where V_{T_0} is the limit of the threshold voltage as $V_{DS} \to 0$, and δ is the DIBL parameter.

The charge at the top of the barrier then becomes $Q = C_G\,(V_G - V_{T_0} + \delta V_D)$. Hence, it

is easy to see that:

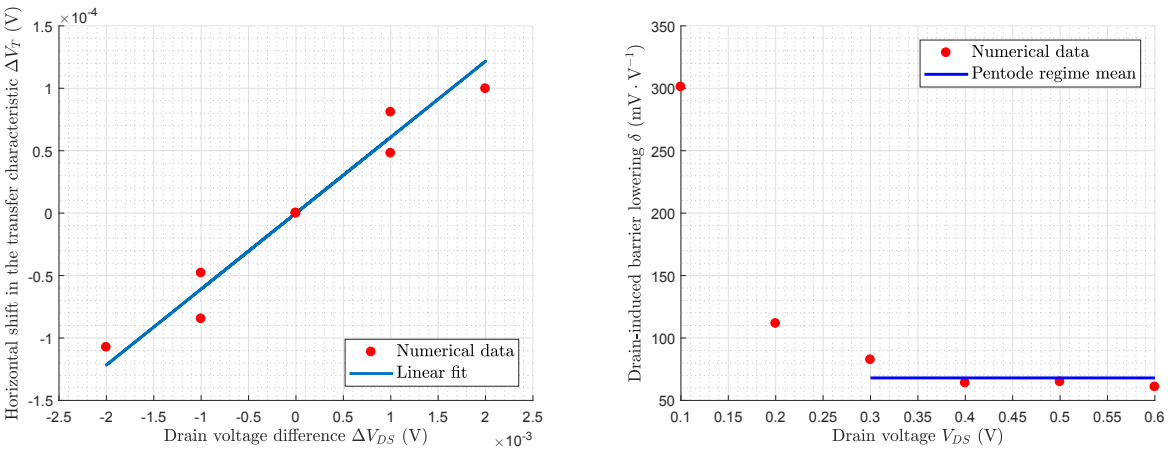
$$\delta = \frac{C_D}{C_G} = \frac{\alpha_D}{\alpha_G} \tag{114}$$

where α_D and α_G are respectively the drain and gate control parameters, as defined in Sec. 2.3.2.

To compute the value of the DIBL parameter δ at a particular value of V_{DS}, we use the following procedure. First, we obtain, from TB-NEGF simulations, three sets of transfer characteristics in the range 0.7 V $\leq V_{GS} \leq$ 1.2 V, one with fixed drain voltage V_{DS}, and the two others with fixed drain voltages $V_{DS} \pm 1$ mV. Second, we consider members of any pair of this set of three characteristics, and shift them horizontally as to obtain the maximal amount of overlap between them. This was achieved using a golden-section search algorithm. The optimal value of the horizontal shift ΔV_T is then recorded, together with the drain voltage difference ΔV_{DS} between the two characteristics under consideration. Third, we plot ΔV_T as a function of ΔV_{DS}, as illustrated in Fig. 25a. We model ΔV_T as a linear function of ΔV_{DS} with 0 intercept, namely: $\Delta V_T = \delta \Delta V_{DS}$. Finally, the extracted δ is what we define as the value of the DIBL parameter at the considered value of V_{DS}, which we plot in Fig. 25b. We computed the mean δ_{sat} of the DIBL parameter in the saturation regime, namely between $V_{DS} = 0.3$ V and $V_{DS} = 0.6$ V. The values of δ_{sat} are reported in Tab. 6.

We note, perhaps surprisingly, that the DIBL parameter δ is far from being a constant function of V_{DS}, as can be seen in Fig. 25b. Indeed, δ appears to decrease with V_{DS}, and tends to a constant in the saturation regime. Note that the three other MOSFET studied in this thesis display precisely the same behaviour. We conclude that Eq. 113 cannot accurately describe the inversion charge of a MOSFET. To remedy to this problem, we suggest the following model for the threshold voltage:

$$\begin{aligned} V_T &= V_{T_0} - \int_0^{V_{DS}} \delta dV_{DS} \\ \delta &= \delta_{\text{lin}} \mathbb{I}_{\left[0, V_{DS_{\text{sat}}}\right]}(V_{DS}) + \delta_{\text{lin}} \mathbb{I}_{\left]V_{DS_{\text{sat}}}, \infty\right[}(V_{DS}) \end{aligned} \tag{115}$$



(a) a (b) a

Figure 25 – DIBL of a 5 nm ML-BP MOSFET oriented along AD. In (a), we illustrate the extraction procedure for δ at $V_{DS} = 0.6$ V. In (b), we plot the extracted values of δ as a function of V_{DS}, and display the mean value of δ in the pentode regime, namely for $0.3 \text{ V} \leq V_{DS} \leq 0.6$ V.

where for an interval I, $\mathbb{I}_I$ is defined as:

$$\mathbb{I}_I\left(V_{DS}\right) = \begin{cases} 1 & \text{for } V_{DS} \in I \\ 0 & \text{for } V_{DS} \notin I \end{cases} \tag{116}$$

To motivate Eq. 115, we note that the effect of DIBL on current-voltage characteristics of MOSFETs is most manifest in the positive slope that it induces in the output characteristics in the saturation regime. To properly model this slope, the saturation regime value of the DIBL parameter δ_{sat} must be used. However, δ_{sat} is smaller than the value of the DIBL parameter in the linear regime. This is why we choose to model DIBL with the additional parameter δ_{lin} in the linear regime. Indeed, without δ_{lin}, the inversion charge would be systematically underestimated.

To compute δ_{lin}, we do not simply take the average of the extracted values of δ in the linear regime. Indeed, in the linear regime, I_{DS} depends on V_{DS} not merely through the DIBL effect, but also and more importantly through current saturation. As such, the values of δ extracted in the linear regime using the procedure described in this section are unreliable.

Transport direction	Armchair		Zigzag	
Channel length	10 nm	5 nm	10 nm	5 nm
δ_{sat} $(\text{mV} \cdot \text{V}^{-1})$	42 ± 2	68 ± 10	51 ± 7	67 ± 21
$\alpha_{D_{\text{sat}}}$	0.040 ± 0.002	0.051 ± 0.007	0.049 ± 0.007	0.059 ± 0.018
δ_{lin} $(\text{mV} \cdot \text{V}^{-1})$	49 ± 54	132 ± 71	95 ± 58	181 ± 39
$\alpha_{D_{\text{lin}}}$	0.046 ± 0.052	0.098 ± 0.052	0.090 ± 0.055	0.159 ± 0.034

Table 6 – Extracted values of δ_{sat} and δ_{lin} for the four MOSFETs studied in this thesis. The extraction procedures of these two parameters are described in Sec. 3.4. We also list the corresponding values of the saturation and linear regimes drain control parameters $\alpha_{D_{\text{sat}}}$ and $\alpha_{D_{\text{lin}}}$. Note that we have defined $\alpha_{D_{\text{sat}}} = \delta_{\text{sat}}\alpha_G$, where α_G is listed in Tab. 5. The parameter $\alpha_{D_{\text{lin}}}$ was defined analogously.

Instead, to compute δ_{lin}, we set $V_{DS_{\text{low}}} = 1$ mV, $V_{DS_{\text{sat}}} = 0.3$ V, and $V_{DS_{\text{high}}} = 0.6$ V. We then compute V_{T_0} using the ballistic mobility method, as described in Sec. 3.2.2.1. We also compute the value of the threshold voltage at $V_{DS_{\text{high}}}$, which we denote as $V_{T_{\text{high}}}$, using the amorphous channel method (Sec. 3.2.2.2). Eq. 115 can then be used to compute the value of δ_{lin}:

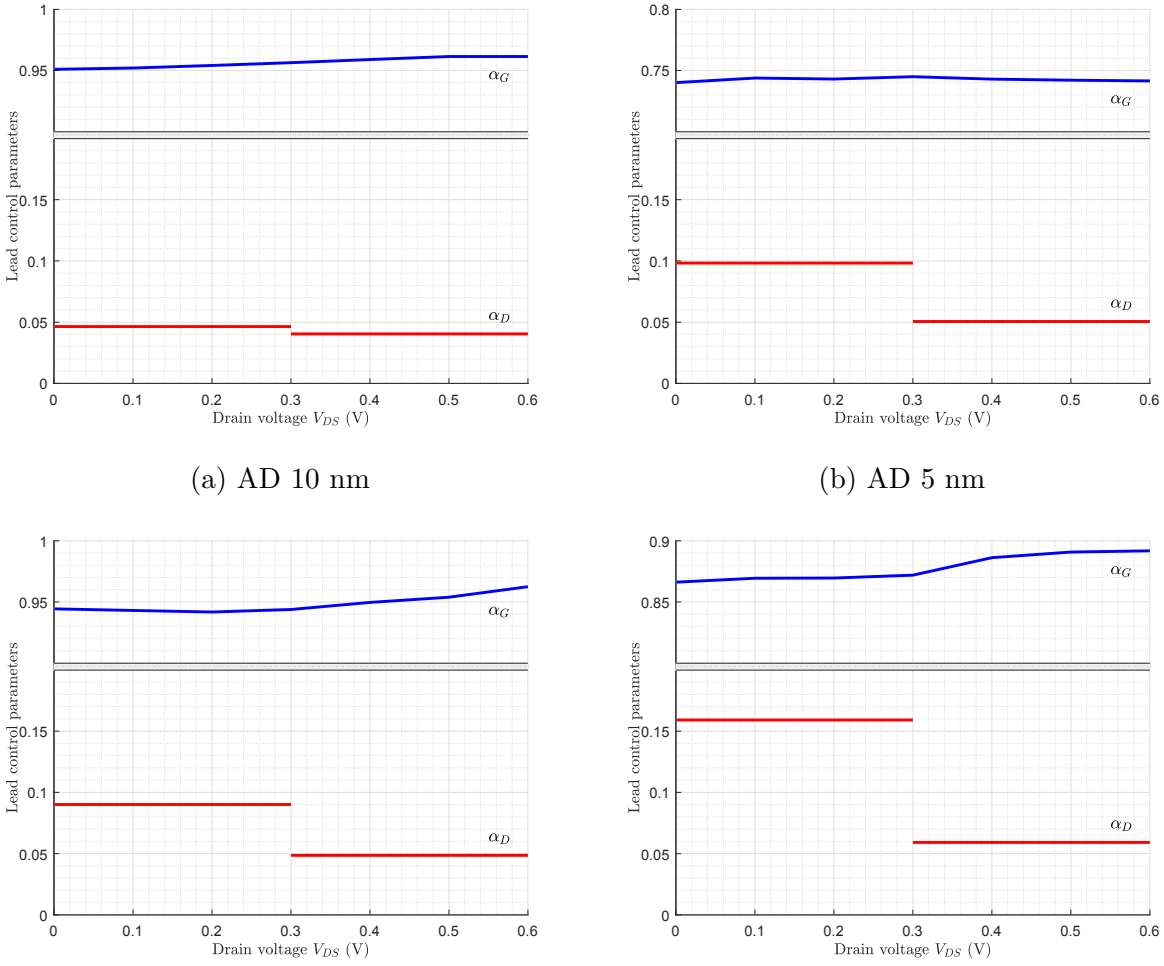
$$\delta_{\text{lin}} = -\frac{V_{T_{\text{high}}} - V_{T_0} + \delta_{\text{sat}}\left(V_{DS_{\text{high}}} - V_{DS_{\text{sat}}}\right)}{V_{DS_{\text{sat}}}} \tag{117}$$

This procedure was carried out; the resulting values of δ_{lin} are listed in Tab. 6.

3.5 Lead control parameters

Using the methodologies of Sec. 3.3 and Sec. 3.4, we are now able to plot the gate and drain control parameters α_G and α_D as a function of the drain voltage V_{DS}. This is shown in Fig. 26.

We note, as previously stated, that α_G is high and roughly independent of V_{DS}. On the other hand, α_D strongly depends upon V_{DS}. Why might that be the case? We believe that the floating source effect to be the culprit for this peculiar behaviour (Sec. 2.3.2.1).



(a) AD 10 nm

(b) AD 5 nm

(c) ZD 10 nm

(d) ZD 5 nm

Figure 26 – Gate and drain control parameters plotted as a function of V_{DS} for the four MOSFETs studied in this thesis. The extraction methods for α_G and α_D are respectively described in Sec. 3.3 and Sec. 3.4. We note that α_G is roughly constant and α_D decreasing.

Indeed, as shown in Fig. 27, in ballistic MOSFETs, as the drain voltage increases, the top of the potential barrier is pushed closer and closer to the source electrode. Therefore, the capacitive coupling between the source and the channel 2DEG, and hence α_D decreases as V_{DS} is increased.

We conclude that the floating source effect is an important phenomenon to consider in the modeling of ballistic or nearly ballistic MOSFETs. Indeed, it leads to:

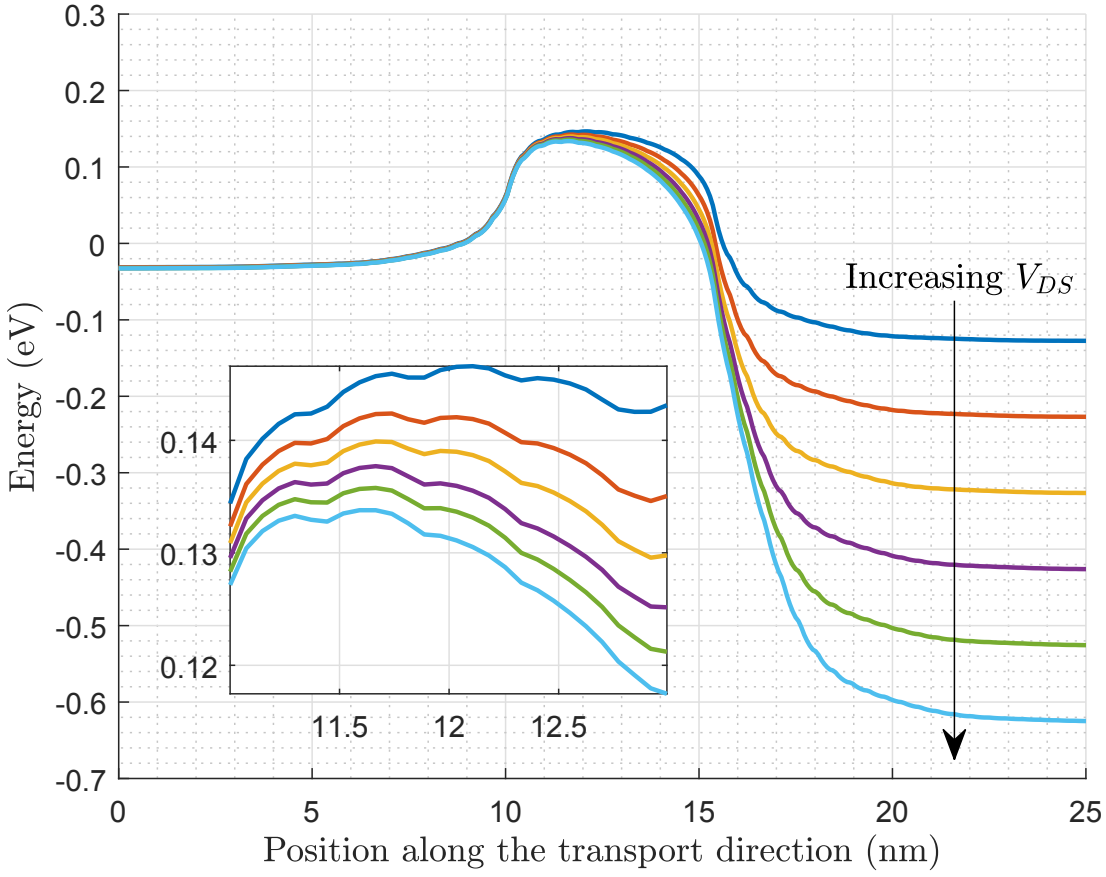


Figure 27 – Band diagrams of a 5 nm channel ML-BP MOSFET oriented along AD, calculated using TB-NEGF simulations. The gate voltage is fixed at $V_{GS} = 0.8$ V, and the drain voltage is swept from $V_{DS} = 0.1$ V to $V_{DS} = 0.6$ V in steps of $\Delta V_{DS} = 0.1$ V. These band diagrams are magnified near the top of the barrier in the inset. We note that the position of the top of the barrier shifts towards the source as V_{DS} is increased. For clarity, the band diagrams presented in this figure were smoothed using a moving average filter.

- A reduction of the conduction band minimum in the source electrode as V_{GS} is increased, as shown in Sec. 2.3.2.1,
- The emergence of a dependence of the DIBL parameter upon V_{DS}.

4 Compact Modeling

In this section, we combine the analytical models that we presented in Sec. 2 with the parameters extraction methods that we developed in Sec. 3 to produce compact models for the four ML-BP MOSFETs under investigation in this thesis. We find that the Landauer-Büttiker model is rather unsuccessful, due to its failure to consider the floating source effect. On the other hand, the capacitor model is surprisingly accurate in predicting both transfer and output characteristics. We also present the virtual source model, and show that it has good predictive power at an extremely low computational cost.

4.1 Landauer-Büttiker model

To compute I_{DS} in the Landauer-Büttiker model for given values of V_{DS} and V_{GS}, we use the algorithm presented in Sec. 2.3.1, namely:

1. Parametrize the inversion layer charge as:
   $Q_0 = -C_G \left(V_{GS} - V_{T_0} + \int_0^{V_{DS}} \delta dV_{DS} \right)$
   Note that DIBL is treated as in Sec. 3.4.

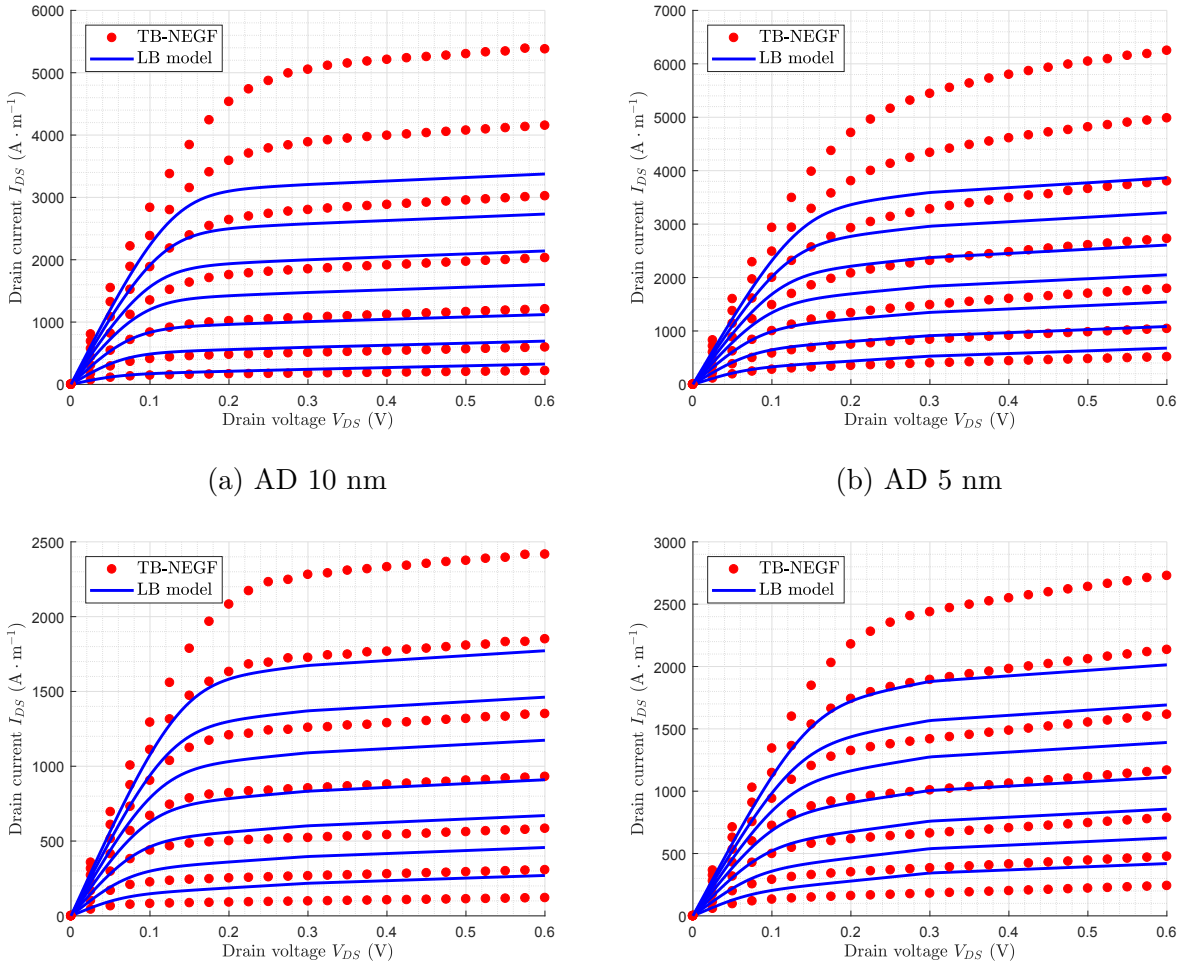
2. Compute η_{F_S} from Eq. 23:
   $Q_0 = -e\frac{N_2}{2} \left[\mathscr{F}_0 \left(\eta_{F_S} \right) + \mathscr{F}_0 \left(\eta_{F_D} \right) \right]$. Note that $\eta_{F_D} = \eta_{F_S} - \frac{eV_{DS}}{k_B T}$

3. Compute I_{DS} from Eq. 24:
   $$I_{DS} = W \left| Q_0 \right| \left[v_T \frac{\mathscr{F}_{\frac{1}{2}}\left(\eta_{F_S} \right)}{\mathscr{F}_0\left(\eta_{F_S} \right)} \right] \left[\frac{1 - \frac{\mathscr{F}_{\frac{1}{2}}\left(\eta_{F_D} \right)}{\mathscr{F}_{\frac{1}{2}}\left(\eta_{F_S} \right)}}{1 + \frac{\mathscr{F}_0\left(\eta_{F_D} \right)}{\mathscr{F}_0\left(\eta_{F_S} \right)}} \right]$$

Since η_{F_S} can only be computed if Q_0 is known, the Landauer-Büttiker can only be used to compute the I_{DS} in the inversion regime.

The Landauer-Büttiker, as presented in this thesis, is a 7 parameters compact model. Two of those parameters, the density-of-states effective mass $\overline{m}^\star$ and the effective mass along the transport direction $m_t^\star$, pertain to the atomistic properties of the semiconductor

channel material, and were computed from DFT [58]. Four of those parameters pertain to the electrostatic properties of the device. The gate capacitance C_G and the threshold voltage V_{T_0} were computed using the ballistic mobility method introduced in Sec. 3.2.2.1. DIBL was taken into account using the formalism of Sec. 3.4, and is parametrized by two parameters, δ_{sat} and δ_{sat}. The last parameter of the Landauer-Büttiker model is the device temperature, which was set to be $T = 300$ K.



(a) AD 10 nm

(b) AD 5 nm

(c) ZD 10 nm

(d) ZD 5 nm

Figure 28 – Output characteristics of the four ML-BP MOSFETs studied in this thesis. The gate voltage is swept from $V_{GS} = 0.8$ V to $V_{GS} = 1.1$ V in steps of $\Delta V_{GS} = 0.05$ V. We compare the results obtained from TB-NEGF simulations to those obtained from the Landauer-Büttiker (LB) model introduced in Sec. 2.3.1.

In Fig. 28, we compare the output characteristics of the four ML-BP MOSFETs under investigation in this thesis calculated from TB-NEGF simulations and from the Landauer-Büttiker model. The drain current I_{DS} tends to be vastly underestimated in the Landauer-Büttiker model, especially for high values of V_{DS} and V_{GS}. This is not surprising, as the Landauer-Büttiker model does not consider the floating source effect, which increases I_{DS} and becomes increasingly important under higher biases (see Sec. 2.3.2.1 and Sec. 3.5). Perhaps more surprising is how drastic are the corrections due to the floating source effect.

While the Landauer-Büttiker model is too unreliable for compact modeling purposes, it nevertheless seems to predict the linear regime drain current to a reasonable degree of accuracy. Furthermore, the scaling of I_{DS} upon $(V_{GS} - V_T)$ as predicted from the Landauer-Büttiker model seems to roughly match the predictions of the more elaborate TB-NEGF simulations.

4.2 Capacitor model

To compute I_{DS} in the capacitor model for given values of V_{DS} and V_{GS}, we use the algorithm presented in Sec. 2.3.2, namely:

1. Guess the value of the self-consistent potential U_{SCF}. To this end, we use a value of $U_{SCF} = -10^{-19}$ J in our algorithm.

2. Compute the induced charge density ΔQ_0 using U_{SCF} and Eq. 32:
$U_{SCF} = -e\left(\alpha_D V_{DS} + \alpha_G V_{GS}\right) - \frac{1}{C_S+C_D+C_G}\Delta Q_0$

3. Compute the self-consistent potential U_{SCF} using ΔQ_0 and Eq. 28:
$\Delta Q_0 = -\frac{eN_2}{2}\mathscr{F}_0\left(\frac{E_{F_S}-U_{SCF}}{k_BT}\right) - \frac{eN_2}{2}\mathscr{F}_0\left(\frac{E_{F_D}-U_{SCF}}{k_BT}\right) + eN_2\mathscr{F}_0\left(\frac{E_F-U_{SCF_{\text{eq}}}}{k_BT}\right)$

4. Repeat Steps 2 and 3 until convergence is attained. The chosen criterion for convergence is that U_{SCF} changes by less than 10^{-26} J from one iteration to the next.

5. Repeat Steps 2 to 4 with various values of E_{F_S} until one such value that guarantees charge neutrality in the source electrode is found. The charge neutrality condition is described by Eq. 34:

$N_{D_S} = N_2 \log\left[\frac{1+\exp\left(\frac{E_{F_S}-E_{C_S}}{k_BT}\right)}{1+\exp\left(\frac{E_{F_S}-U_{SCF}}{k_BT}\right)}\right] + \frac{N_2}{2}\log\left[1+\exp\left(\frac{E_{F_S}-U_{SCF}}{k_BT}\right)\right]$
$+ \frac{N_2}{2}\log\left[1+\exp\left(\frac{E_{F_S}+eV_{DS}-U_{SCF}}{k_BT}\right)\right]$

To do so, we increase E_{F_S} in steps of 10^{-23} J until the right-hand side of the above equation is greater than or equal to N_{D_S}.

6. Compute the current using Eq. 33:
$I_{DS} = We\frac{N_2}{2}v_T\left[\mathscr{F}_{\frac{1}{2}}\left(\frac{E_{F_S}-U_{SCF}}{k_BT}\right) - \mathscr{F}_{\frac{1}{2}}\left(\frac{E_{F_D}-U_{SCF}}{k_BT}\right)\right]$

As a preliminary step, we compute $E_F - U_{SCF_{\text{eq}}}$ by finding its value such that the capacitor model presented above best fits the $V_{DS} = 1$ mV transfer characteristic between $V_{GS} = 0.4$ V and $V_{GS} = 0.6$ V. The conduction band minimum in the source can they be calculated as:

$$E_{C_S} - U_{SCF_{\text{eq}}} = E_F - U_{SCF_{\text{eq}}} - k_BT\log\left[\exp\left(\frac{N_{D_S}}{N_2}\right) - 1\right] \tag{118}$$

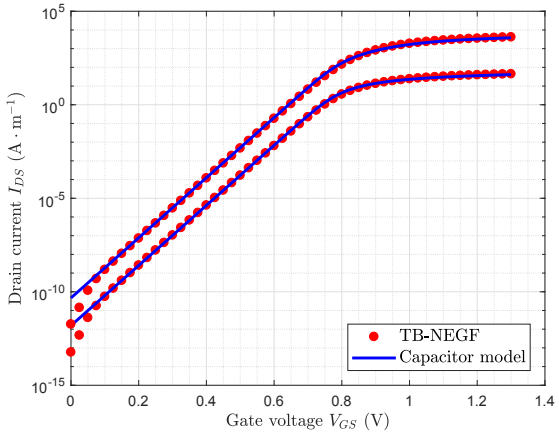
Furthermore, the sum of the source, drain, and gate capacitances can be calculated as:

$$C_S + C_D + C_G = \frac{C_G}{\alpha_G} \tag{119}$$

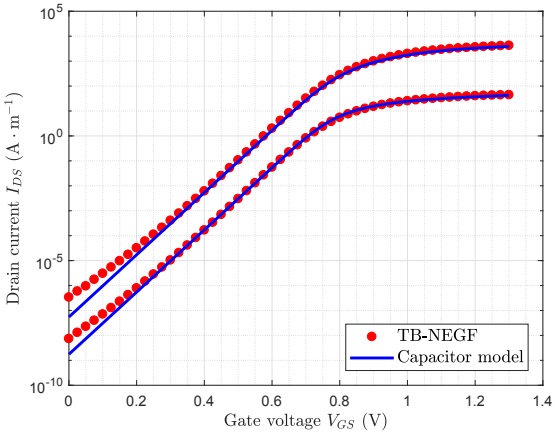
where C_G is the electrostatic capacitance of the gate electrodes.

The capacitor model, as presented in this thesis, is an 8 parameters compact model. One of these parameters specifies the design of the device: the source doping density N_{D_S}. Three of those parameters are atomistic: the density-of-states effective mass $\overline{m}^\star$, the effective mass along the transport direction $m_t^\star$, and the equilibrium barrier height with respect to the Fermi level $E_F - U_{SCF_{\text{eq}}}$. Three of those parameters pertain to the electrostatic properties of the device: the gate capacitance C_G, the gate control parameter α_G, and the drain control parameter α_D. The gate capacitance C_G was computed from Eq. 79 by considering all fringing field effects. The gate control parameter α_G was computed using the method exposed in Sec. 3.3. Finally, the drain control parameter was set to $\alpha_D = \alpha_{D_{\text{lin}}}$ (as defined in Sec. 3.4). Indeed, we found that in the capacitor model, having a V_{DS}-dependent DIBL parameter made little difference. Thus, by virtue of Occam's razor, we favour a one-variable parametrization

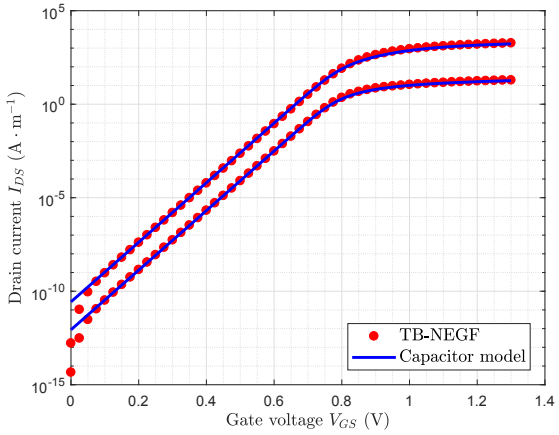
of DIBL. The last parameter of the capacitor model is the device temperature, which was set to be $T = 300$ K.



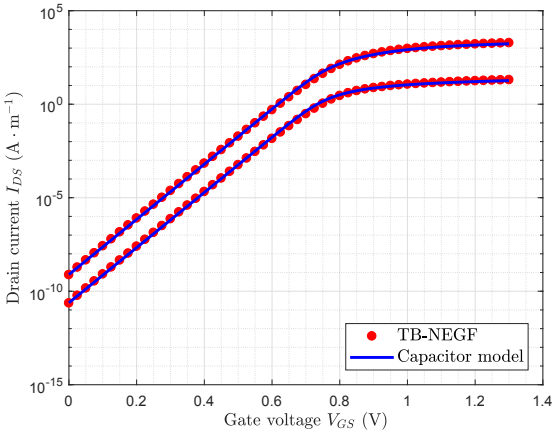
(a) AD 10 nm



(b) AD 5 nm



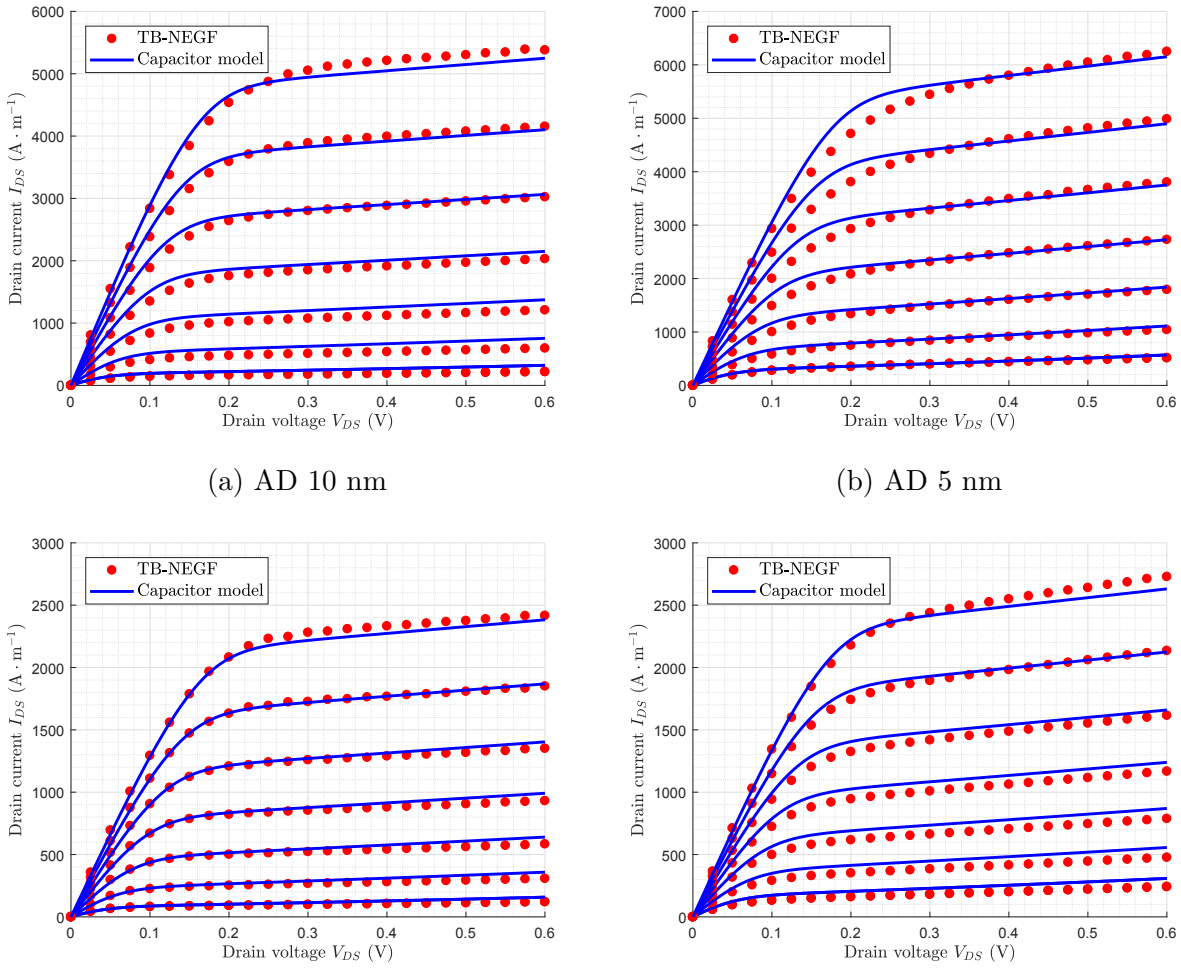
(c) ZD 10 nm



(d) ZD 5 nm

Figure 29 – Transfer characteristics of the four ML-BP MOSFETs studied in this thesis. The drain voltage is set at $V_{DS} = 1$ mV and $V_{DS} = 0.1$ V. We compare the results obtained from TB-NEGF simulations to those obtained from the capacitor model introduced in Sec. 2.3.2.

In Fig. 29 and Fig. 30, we respectively compare the transfer and output characteristics of the four ML-BP MOSFETs under investigation in this thesis calculated from TB-NEGF simulations and from the capacitor model. The drain current I_{DS} is modeled surprisingly well within the capacitor model, both in the subthreshold and inversion regimes, and both in the linear and saturation regimes. By describing electrostatics and Fermi-Dirac statistics self-



(a) AD 10 nm

(b) AD 5 nm

(c) ZD 10 nm

(d) ZD 5 nm

Figure 30 – Output characteristics of the four ML-BP MOSFETs studied in this thesis. The gate voltage is swept from $V_{GS} = 0.8$ V to $V_{GS} = 1.1$ V in steps of $\Delta V_{GS} = 0.05$ V. We compare the results obtained from TB-NEGF simulations to those obtained from the capacitor model introduced in Sec. 2.3.2.

consistently, while taking into account the floating source effect, we have accurately modeled ballistic MOSFETs. Furthermore, the slope of the transfer characteristics in the subthrehsold regime is well-described by the capacitor model. This shows that the α_G extraction method that we presented in Sec. 3.3 is reliable. Likewise, the fact that the slopes of the output characteristics in the saturation regime are well described by the capacitor model shows that the α_D extraction method that we presented in Sec. 3.4 is reliable.

A concern that one might have is whether the drain current I_{DS} can be calculated quickly within the capacitor model. Indeed, the algorithm described at the beginning of this section contains a self-consistent calculation nested into an iterative procedure. However, calculations within the capacitor model are fast enough for the purpose of compact modeling. To give a sense of scale, to calculate all of the TB-NEGF-based data presented in Fig. 29 and Fig. 30, approximately 3 core-years of computing time were needed. On the other hand, approximately 15 core-seconds were needed to obtain the corresponding capacitor-model-based data.

Finally, we mention a caveat to the discussion above. The electrostatic gate capacitance computed from Eq. 79, namely $C_G = \frac{2\varepsilon}{t}\left(1+\frac{a}{\left(\frac{L}{t}\right)^b}\right)$, lead to poor fits of the capacitor model to the TB-NEGF data. Instead, we used one third of the above value, namely $C_G = \frac{2\varepsilon}{3t}\left(1+\frac{a}{\left(\frac{L}{t}\right)^b}\right)$. We believe this mismatch between our theoretical prediction for the gate capacitance and its best-fit value to be at least partly due the way that Poisson's equation is handled in the TB-NEGF simulations. Indeed, the space grid onto which Poisson's equation is solved in these simulations terminates at the exterior boundaries of the oxide layers. As a result, much fewer field lines are considered within the simulations than those that were considered in deriving Eq. 79.

4.3 Virtual source model

We now come back to the virtual source (VS) model, the compact model that we commended in the introduction of this thesis, and cherished for its distinctively bottom-up character. This model [2–4] is continuous from the drift-diffusion to the ballistic limit, and is hence an important conceptual guide in designing MOSFETs approaching the ballistic limit.

In essence, the VS model for the output characteristics of a MOSFET is constructed as follows. First, an expression for the linear regime drain current $I_{DS_{\text{lin}}}$ is obtained. In the case of nondegenerate ballistic MOSFETs, Eq. 49 and Eq. 56 can be used to this end:

$$I_{DS_{\text{lin}}} = C_G\left(V_{GS} - V_{T_0} + \delta V_{DS}\right)\frac{\mu_{\text{ball}} V_{DS}}{L} \tag{120}$$

Second, an expression for the saturation regime current $I_{DS_{\text{sat}}}$ is obtained. For ballistic MOSFETs, Eq. 24 can be used to this end:

$$\begin{aligned} I_{DS_{\text{sat}}} &= C_G \left(V_{GS} - V_{T_0} + \delta V_{DS}\right) v_I \\ v_I &= v_T \frac{\mathscr{F}_{\frac{1}{2}}\left(\eta_{F_S}\right)}{\mathscr{F}_0\left(\eta_{F_S}\right)} \end{aligned} \tag{121}$$

The parameter v_I is known as the injection velocity of charge carriers. In Fig. 31a, we plot $\frac{v_I}{v_T}$ as a function of $\eta_{F_S} = \frac{E_{F_S} - E_C}{k_B T}$.

Third, the drain voltage V_{sat} at which the linear regime current of Eq. 120 equals the saturation regime current of Eq. 121 is calculated. In the present situation, this voltage, known as the saturation voltage, can easily be computed to be:

$$V_{\text{sat}} = \frac{L v_I}{\mu_{\text{ball}}} \tag{122}$$

Fourth, given $\beta > 0$, we define the saturation function as:

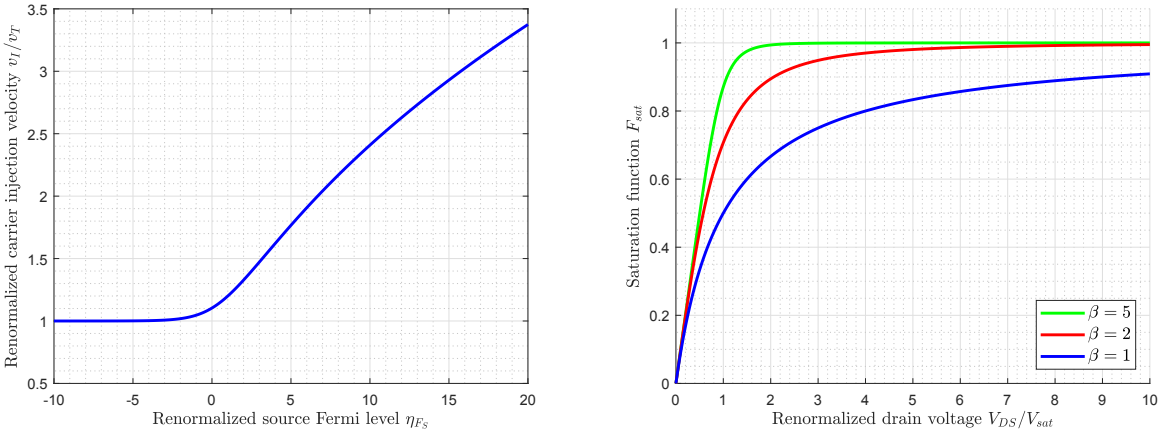
$$F_{\text{sat}}\left(V_{DS}\right) = \frac{\frac{V_{DS}}{V_{\text{sat}}}}{\left[1 + \left(\frac{V_{DS}}{V_{\text{sat}}}\right)^{\beta}\right]^{\frac{1}{\beta}}} \tag{123}$$

The velocity of charge carriers is then expressed as $v_I F_{\text{sat}}\left(V_{DS}\right)$. The saturation function F_{sat} describes the transition of the velocity of charges carriers from their linear regime value of $\frac{\mu_{\text{ball}} V_{DS}}{L}$ to their saturation regime value of v_I, as can be seen by examining the behaviour of F_{sat} in the limits $V_{DS} \to 0$ and $V_{DS} \to \infty$. The parameter β is phenomenological, and quantifies the sharpness of the transition from the linear regime to the saturation regime. In Fig. 31b, we plot F_{sat} for various values of β.

Overall, in the VS model, the drain current in the inversion regime is expressed as:

$$I_{DS} = C_G \left(V_{GS} - V_{T_0} + \delta V_{DS}\right) v_I F_{\text{sat}}\left(V_{DS}\right) \tag{124}$$

As presented in Ref. [4], the VS model injection velocity v_I is left as a free parameter. In this thesis, it is expressed within the Landauer-Büttiker model as $v_I = v_T \frac{\mathscr{F}_{\frac{1}{2}}\left(\eta_{F_S}\right)}{\mathscr{F}_0\left(\eta_{F_S}\right)}$. The exact value of the parameter $\eta_{F_S} = \frac{E_{F_S} - E_C}{k_B T}$ is, however, unknown. It thus remains to express



(a) Injection velocity

(b) Saturation function

Figure 31 – Injection velocity (a) and saturation function (b) for the virtual source model. The injection velocity normalized by the thermal velocity v_T, is plotted as a function of $\eta_{F_S} = \frac{E_{F_S} - E_C}{k_B T}$. The saturation function is plotted for three different values of β as a function of the drain voltage V_{DS} normalized by the saturation voltage V_{sat}.

η_{F_S} as a function of V_{GS}. First, it is known from electrostatics that the barrier height E_C is a linear function of the gate voltage [66, 105]: $E_C \propto \frac{eV_{GS}}{k}$, where $k \approx 1$. We can thus write:

$$\eta_{F_S} = \frac{e}{k_B T} \frac{1}{k} (V_{GS} - V_{T_0}) + \eta_{T_0} \tag{125}$$

where η_{T_0} is a constant. Somewhat arbitrarily, we demand that $\eta_{F_S} = -2$ for $V_{GS} = V_{T_0}$, from which we can deduce that $\eta_{T_0} = -2$. We motivate this choice as follows. The threshold voltage V_{T_0} is defined as the value of V_{GS} at the onset of high conductivity within the MOSFET channel. The drain current through a ballistic MOSFET is proportional to $\mathscr{F}_{\frac{1}{2}}(\eta_{F_S})$ (see Eq. 33). Since this complete Fermi-Dirac integral of order $\frac{1}{2}$ only starts to strongly deviate from 0 at $\eta_{F_S} = -2$, as can be seen in Fig. 33, we require that $\eta_{F_S} = -2$ for $V_{GS} = V_{T_0}$. Setting $\eta_{T_0} = -2$ thus corresponds to the statement that the MOSFET channel becomes conductive when the barrier height E_C is within $2k_B T$ of the source Fermi level E_{F_S}.

Hence, to compute I_{DS} in the virtual source model for given values of V_{DS} and V_{GS}, we use the following algorithm:

1. Compute η_{F_S} from Eq. 125:

$\eta_{F_S} = \frac{e}{k_B T} \frac{1}{k} \left(V_{GS} - V_{T_0}\right) + \eta_{T_0}$

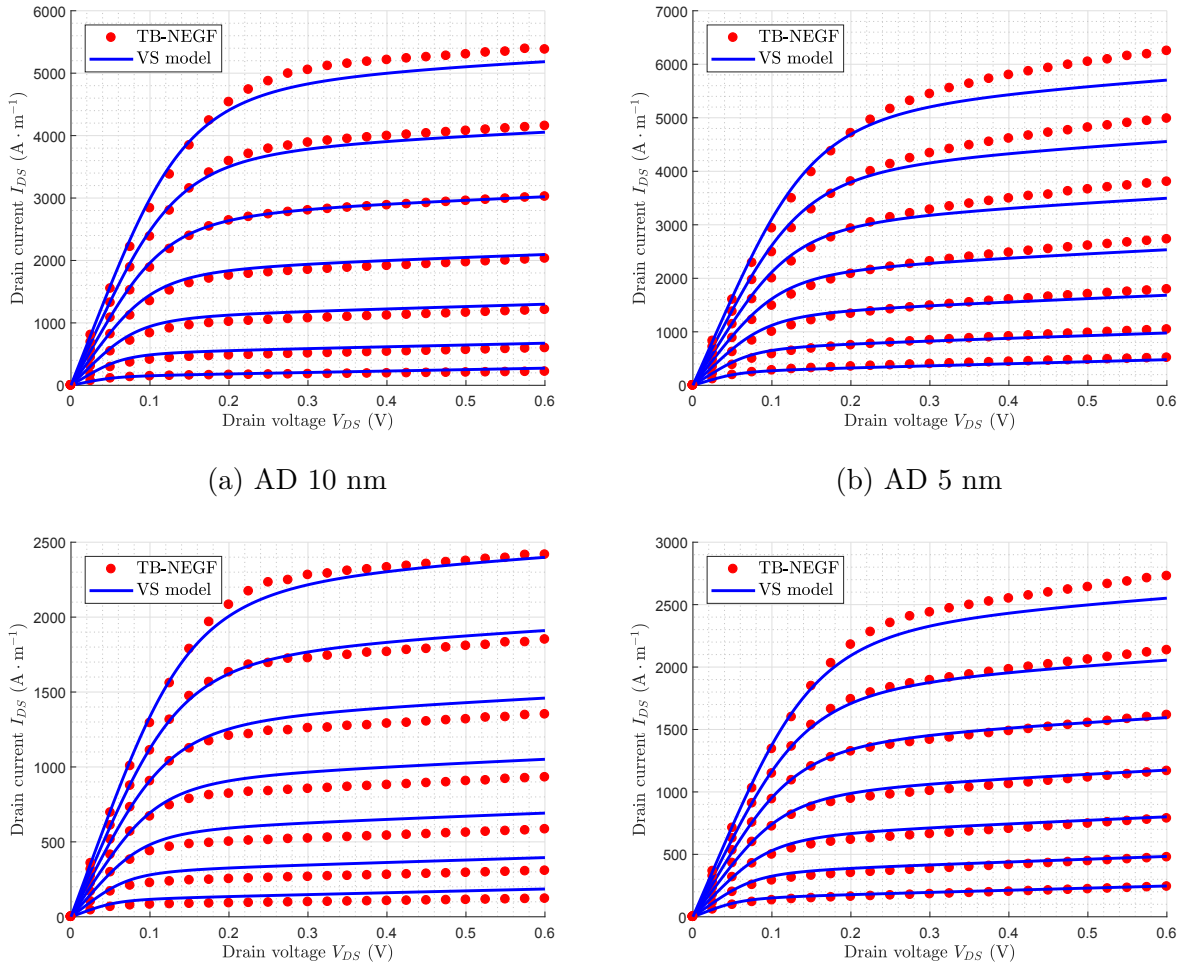
Note that η_{F_S} only depends upon V_{GS} within our parametrization.

2. Compute the injection velocity from Eq. 121:

$v_I = v_T \frac{\mathscr{F}_{\frac{1}{2}}\left(\eta_{F_S}\right)}{\mathscr{F}_0\left(\eta_{F_S}\right)}$

3. Compute the drain current from Eq. 124:

$I_{DS} = C_G \left(V_{GS} - V_{T_0} + \delta V_{DS}\right) v_I F_{\text{sat}} \left(V_{DS}\right)$



(a) AD 10 nm

(b) AD 5 nm

(c) ZD 10 nm

(d) ZD 5 nm

Figure 32 – Output characteristics of the four ML-BP MOSFETs studied in this thesis. The gate voltage is swept from $V_{GS} = 0.8$ V to $V_{GS} = 1.1$ V in steps of $\Delta V_{GS} = 0.05$ V. We compare the results obtained from TB-NEGF simulations to those obtained from the virtual source (VS) model.

The virtual source model, as presented in this thesis, is an 8 parameters compact model. One of those parameters is atomistic: the effective mass along the transport direction $m_t^\star$. Three of those parameters pertain to the electrostatic properties of the device: the gate capacitance C_G, the threshold voltage V_{T_0}, and the DIBL parameter δ. The gate capacitance C_G and threshold voltage V_{T_0} were computed using the ballistic mobility method of Sec. 3.2.2.1. The DIBL parameter was set to $\delta = \delta_{\text{lin}}$ (as defined in Sec. 3.4). Three of these parameters are phenomenological: β, k, and η_{T_0}. The parameter β was set to $\beta = 3$, the parameter k to $k = \frac{3}{4}$, and η_{T_0} to $\eta_{T_0} = -2$, as those values lead to good fits. The last parameter of the virtual source model is the device temperature, which was set to be $T = 300$ K. Note that the channel length L is not a parameter of the virtual source model, as whenever it appears in the equations that define the VS model, it is cancelled out by another factor of L hidden in μ_{ball}.

In Fig. 32, we compare the output characteristics of the four ML-BP MOSFETs under investigation in this thesis calculated from TB-NEGF simulations and from the virtual source model. The drain current I_{DS} is modeled fairly well with the VS model, both in the linear and saturation regimes. While the VS model is not as successful as the capacitor model presented in Sec. 4.2, it has the power of calculating I_{DS} essentially instantaneously. Indeed, the most computationally heavy component of the virtual source model is the calculation of the complete Fermi-Dirac integrals defining the injection velocity (see Eq. 121), which merely needs to be done once for each sampled value of the gate voltage V_{GS}. The virtual source model would thus likely be more appropriate than the capacitor model for simulations of circuits of great scale.

5 Conclusion

We conclude this thesis by a brief summary, followed by the description of a potential future research project in the same vein as the work presented in this thesis.

5.1 Summary

In this thesis, we have developed two accurate bottom-up compact models for ballistic MOSFETs. These models, the capacitor model and the virtual source model, both parametrize current-voltage characteristics using 8 parameters, which stand in striking contrast with the hundreds or thousands of parameters that are found in compact models typically used by the semiconductor industry.

Our models find their foundation in the Landauer-Büttiker formalism of ballistic transport. This formalism has strong physical footing, thus making our models inherently bottom-up. As such, while we have focused our attention on ML-BP MOSFETs, we strongly believe that our models can be used without any major alteration for ballistic MOSFETs composed of various other semiconductors.

In developing our models, we have shown the floating source effect to be of great importance in ballistic MOSFETs. Additionally, we have shown how to obtain the parameters of our models, be it from theoretical arguments, from *ab initio* simulations, or by extraction from current-voltage characteristics.

State-of-the-art MOSFETs have been operating near or at the ballistic limit for more than a decade. Nevertheless, the field of compact modeling, and more generally the field of nanoeletronics, are still to a great extent dominated by ideas stemming from the drift-diffusion transport theory. Concepts such as that of ballistic mobility serve to bridge the gap between these ideas, in term of which scientists and engineers tend to think about MOSFETs, and the actual operating principles of MOSFETs. However, such concepts are fallacious; a

fresh, simple ballistic transport formalism needs to be canonized within the device research community. We believe the Landauer-Büttiker formalism to be the best candidate to take this role. Not only does it provide a clear and coherent description of MOSFETs and many other electronic devices, but it can also, with a few adjustments, be used to build accurate compact models.

5.2 Future research direction

Ever since the invention of the integrated circuit, MOSFETs have been the most prevalent building blocks for digital logic and complex computations. They, along with other semiconductors devices, have undergone an uninterrupted miniaturization over the past fifty years, following a trend known as Moore's law [106]. Now that the channel lengths of MOSFETs are reaching the nanoscale, it is quantum phenomena that dominate their physical properties. Notably, short-channel effects such as DIBL depreciate the power efficiency of MOSFETs. Consequently, researchers have to juggle between these physical constraints, which lead to increased power dissipation, and economic constraints on acceptable power dissipation. This dilemma has hindered the progression of Moore's law over the past five years [107]. An urgent task is thus to devise a device which breaks this power dissipation bottleneck.

The subthreshold swing (SS) is a device parameter which quantifies how difficult it is to turn off a transistor, and is therefore commensurate with leakage currents and power consumption. Electrostatics and thermodynamics dictate that at room temperature, the SS of a MOSFET must be greater than 60 $\mathrm{mV} \cdot \mathrm{dec}^{-1}$. A number of potential successors to the MOSFET with sub-60 $\mathrm{mV} \cdot \mathrm{dec}^{-1}$ SS have been studied, the most promising of which is the tunneling field-effect transistor (TFET) [108]. While thermionic injection is the physical mechanism for electronic transport in MOSFETs, TFETs rely on quantum tunneling. Since quantum tunneling is a temperature-independent process, sub-60 $\mathrm{mV} \cdot \mathrm{dec}^{-1}$ SS can easily be achieved in TFETs [109]. However, the ON-state currents of TFETs are typically ~ 2 orders of magnitude smaller than those of MOSFETs, thus making their adoption on large-scale logic

circuits impractical. Very recently, a potential surrogate to the TFET was introduced. This device, called the Dirac field-effect transistor (DFET), is composed of a monolayer graphene source, an intrinsic monolayer molybdenum disulfide (MoS_2) channel, and a n-doped MoS_2 drain. By virtue of the linear density of states of graphene near its Dirac points, the DFET effectively behaves as a TFET in the OFF-state and a MOSFET in the ON-state. This transistor was shown to break the 60 $mV \cdot dec^{-1}$ SS limit while maintaining a high ON-state current, both experimentally [110] and theoretically [111].

The goals of this research project are twofold. First, the unusual behavior of the DFET is believed to be due to the formation of a gate-modulated Schottky barrier at the graphene/MoS_2 interface, due to the semimetallic character of graphene [111]. I will investigate the universality of this phenomenon by simulating a similar device with another Dirac material in place of graphene in the source (for example, silicene [112]), and another two-dimensional semiconductor in place of MoS_2 in the channel (for example, ML-BP [5]). To do so, I will use first-principles calculations based on density functional theory (DFT) within the nonequilibrium Green's function formalism (NEGF), which has emerged as one of the most powerful and practical methods for predicting nonlinear and nonequilibrium quantum transport properties of nanoelectronic devices [1]. Additionally, the effects of disorder and impurity scattering can be taken into account within the nonequilibrium coherent potential approximation (NECPA) [113]. The NEGF-DFT-NECPA formalisms have been implemented in Nanodsim, a software package developed by Guo *et al.* that I will use for this part of the project.

Second, I will construct a compact model for the DFET. Compact models are analytical parametrizations of the current-voltage characteristics of electronic devices, which are used by the semiconductor industry in large-scale circuit simulations. Such a compact model could then be used to study DFET-based circuits of interest (for example, a static random-access memory cell) to investigate if the use of DFETs does lead to decreased power dissipation at the circuit level. Overall, this research project is one of the first theoretical investigations of a promising successor to the MOSFET, and fits within the research and development of low-power nanoelectronics.

A Complete Fermi-Dirac Integrals

The complete Fermi-Dirac integral of order $i \in \mathbb{Q}$ is defined for $\eta \in \mathbb{R}$ as [114, 115]:

$$\mathscr{F}_i(\eta) = \frac{1}{\Gamma(i+1)} \int_0^\infty \frac{t^i}{1+\exp(t-\eta)} dt \tag{126}$$

Complete Fermi-Dirac integrals of order $-\frac{1}{2}$, 0, and $\frac{1}{2}$ are plotted in the range $-4 \leq \eta \leq 10$ in Fig. 33 for illustrative purposes.

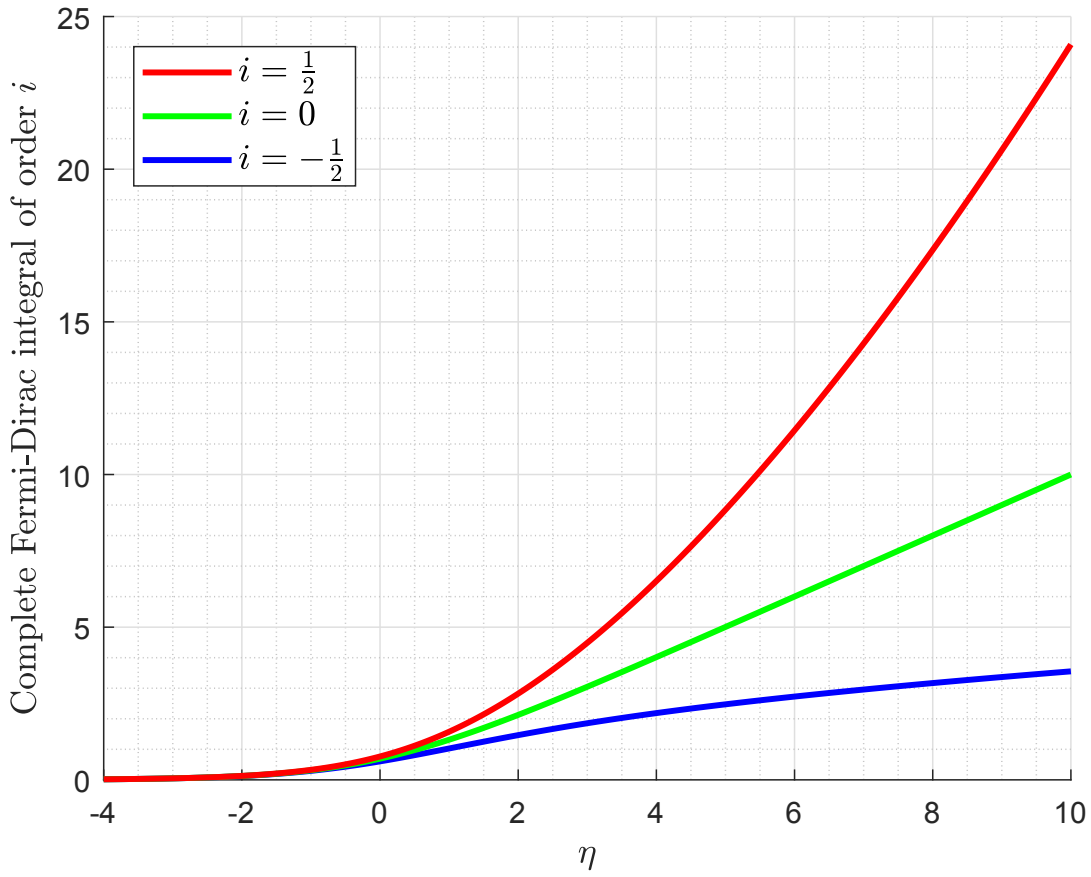


Figure 33 – Plots of the complete Fermi-Dirac integrals $\mathscr{F}_i(\eta)$ of order $i = -\frac{1}{2}$, $i = 0$, and $i = \frac{1}{2}$ as a function of η. The integrals were computed numerically using Eq. 132.

These integrals often arise in semiconductor-related calculations. For example, consider a n-dimensional semiconductor ($n \in \mathbb{N}$) with a parabolic conduction band. The density of

states $\rho_n(E)$ of conduction band electrons (where E denotes energy) satisfies, for $E \geq E_C$ [73]:

$$\rho_n(E) = g_n (E - E_C)^{\frac{n-2}{2}} \theta(E - E_C) \tag{127}$$

where E_C is the conduction band minimum, g_n is a material-specific constant independent of E, and θ the Heaviside step function. Hence, the density n of conduction band electrons is:

$$n = \int_{-\infty}^{+\infty} \frac{\rho_n(E)}{1 + \exp\left(\frac{E-E_F}{k_B T}\right)} dE = g_n \int_{E_C}^{\infty} \frac{(E - E_C)^{\frac{n-2}{2}}}{1 + \exp\left(\frac{E-E_F}{k_B T}\right)} dE \tag{128}$$

where E_F and T are respectively the Fermi level and temperature of the semiconductor. By making the substitution $t = \frac{E-E_C}{k_B T}$ and defining $\eta_F = \frac{E_F - E_C}{k_B T}$, Eq. 128 can be re-expressed in terms of a Fermi-Dirac integral as:

$$n = g_n (k_B T)^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right) \mathscr{F}_{\frac{n-2}{2}}(\eta_F) \tag{129}$$

While the Γ function included in the definitions of complete Fermi-Dirac integrals (Eq. 126) might first seem unnecessary, it guarantees the desirable property that $\forall i \in \mathbb{Q}$ and $\forall \eta \in \mathbb{R}$:

$$\frac{\partial \mathscr{F}_i(\eta)}{\partial \eta} = \mathscr{F}_{i-1}(\eta) \tag{130}$$

The complete Fermi-Dirac integral of order 0 can be computed analytically to be:

$$\mathscr{F}_0(\eta) = \log[1 + \exp(\eta)] \tag{131}$$

There exist analytical approximations to the Fermi-Dirac integrals of order i for $-\frac{1}{2} \leq i \leq \frac{5}{2}$ [116–118], as well as quickly-converging series expansions for $-\frac{1}{2} \leq i \leq \frac{7}{2}$ [119, 120]. However, for greater accuracy, one must resort to numerical methods. A standard approach [121], valid for $i > -1$ and $\eta \lesssim 15$ is to make the substitution $t = \exp(x - e^{-x})$, so that Eq. 126 can be written as:

$$\mathscr{F}_i(\eta) = \frac{1}{\Gamma(i+1)} \int_a^b t\left(1 + e^{-x}\right) \frac{t^j}{1 + \exp(t - \eta)} dx \tag{132}$$

where a and b are chosen in such a way that the integrand in Eq. 132 is almost 0 for $x = a$ and $x = b$. Typical values for the parameters a and b are respectively -5 and 5. The

complete Fermi-Dirac integral as expressed in Eq. 132 has the advantage of being readily computable using the easily implemented trapezoidal rule, despite the fact that the domain of integration of a complete Fermi-Dirac integral is infinite (Eq. 126).

Finally, we examine the asymptotic behaviour of complete Fermi-Dirac integrals in the limit $\eta \ll 0$ (which for obvious reason can be referred to as the nondegenerate limit). In this limit, one has:

$$\mathscr{F}_i(\eta) = \frac{1}{\Gamma(i+1)} \int_0^\infty \frac{t^i}{\exp(t-\eta)} dt \tag{133}$$

and hence, by definition of the Γ function, $\forall i \in \mathbb{Q}$, one obtains:

$$\mathscr{F}_i(\eta) \to e^\eta \quad \text{as} \quad \eta \to -\infty \tag{134}$$

It can also be shown [122] that in the degenerate limit $\eta \gg 0$:

$$\mathscr{F}_i(\eta) \to \frac{\eta^{i+1}}{\Gamma(i+2)} \quad \text{as} \quad \eta \to +\infty \tag{135}$$

B Conformal Mappings and Schwarz-Christoffel Transformations

We expose some important mathematical background on conformal mappings and Schwarz-Christoffel transformations [123].

Definition: Let U and V be open subsets of $\mathbb{C}$. A mapping $f\colon U \to V$ is said to be conformal if $\forall u \in U$, f preserves the oriented angles between all pairs of curves going through u.

The Riemann mapping theorem guarantees the existence of a conformal mapping between the Poincaré upper half-plane $\mathbb{H} = \{\zeta \in \mathbb{C} \mid \mathrm{Im}\,(\zeta) > 0\}$ and the interior of any polygon in the complex plane. Such mappings are called Schwarz-Christoffel mappings.

Definition: Let $S \subset \mathbb{C}$. Then, the closure of S, which we denote as $\mathrm{cl}\,(S)$, is defined to be the unique smallest closed subset of $\mathbb{C}$ containing S.

As an example, if $S = \{z \in \mathbb{C} \mid |z| < 1\}$, then $\mathrm{cl}\,(S) = \{z \in \mathbb{C} \mid |z| \leq 1\}$.

Definition: Let $P \subset \mathbb{C}$ be an open simple polygon with interior angles $(\alpha, \beta, \gamma, \cdots)$. Let $a, b, c, \cdots \in \mathbb{R}$ be such that $a < b < c < \cdots$ and $A \in \mathbb{C}$ a constant. Then, a Schwarz-Christoffel mapping is a mapping $f\colon \mathrm{cl}\,(\mathbb{H}) \to \mathrm{cl}\,(P)$ defined by:

$$f(\zeta) = \int^{\zeta} \frac{A}{(w-a)^{1-\frac{\alpha}{\pi}} (w-b)^{1-\frac{\beta}{\pi}} (w-c)^{1-\frac{\gamma}{\pi}} \cdots} dw \tag{136}$$

The mapping f defined above has the property that it maps the real axis $\mathbb{R}$ to the edges of the polygon $\mathrm{cl}\,(P)$. Furthermore, given appropriate choices of A and of the constant of integration in Eq. 136, f maps $(a, b, c, \cdots)$ to the vertices of $\mathrm{cl}\,(P)$. Finally, the restriction $f|_{\mathbb{H}}$ of f to $\mathbb{H}$ is a conformal mapping from $\mathbb{H}$ to P.

C Incomplete and Complete Elliptic Integrals of the First and Second Kinds

Elliptic integrals often arise in the calculation of conformal mappings and Schwarz-Christoffel transformations. Notations surrounding these integrals are numerous in the literature. For this reason, we wish to set the notation that we will use in this thesis.

The incomplete elliptic integral of the first kind is defined for $-1 < \zeta < 1$ and $0 < k < 1$ as:

$$\begin{aligned} F(\zeta; k) &= \int_0^{\zeta} \frac{1}{\sqrt{(1-w^2)(1-k^2w^2)}} dw \\ &= \int_0^{\phi} \frac{1}{\sqrt{1-k^2\sin^2\theta}} d\theta \coloneqq F(\phi, k) \end{aligned} \tag{137}$$

where $\sin\phi = \zeta$ with $-\frac{\pi}{2} < \phi < \frac{\pi}{2}$. The parameter k is sometimes referred to as the elliptic modulus, or eccentricity.

The complete elliptic integral of the first kind is defined to be the value of the above integral in the limit $\zeta \to 1$, or equivalently $\phi \to \frac{\pi}{2}$:

$$K(k) = \int_0^1 \frac{1}{\sqrt{(1-w^2)(1-k^2w^2)}} dw = \int_0^{\frac{\pi}{2}} \frac{1}{\sqrt{1-k^2\sin^2\theta}} d\theta \tag{138}$$

The incomplete elliptic integral of the second kind is defined for $-1 < \zeta < 1$ and $0 < k < 1$ as:

$$\begin{aligned} E(\zeta; k) &= \int_0^{\zeta} \sqrt{\frac{1-k^2w^2}{1-w^2}} dw \\ &= \int_0^{\phi} \sqrt{1-k^2\sin^2\theta} d\theta \coloneqq E(\phi, k) \end{aligned} \tag{139}$$

where $\sin\phi = \zeta$ with $-\frac{\pi}{2} < \phi < \frac{\pi}{2}$. The complete elliptic integral of the second kind is defined to be the value of the above integral in the limit $\zeta \to 1$, or equivalently $\phi \to \frac{\pi}{2}$:

$$E(k) = \int_0^1 \sqrt{\frac{1-k^2w^2}{1-w^2}} dw = \int_0^{\frac{\pi}{2}} \sqrt{1-k^2\sin^2\theta} d\theta \tag{140}$$

We note that the complete elliptic integral of the second kind has the following mathematical interpretation: given an ellipse with semi-major axis s_a and semi-minor axis s_b (which yield a value of $e = \sqrt{1 - \frac{s_b^2}{s_a^2}}$ for the elliptic eccentricity), then the circumference of the ellipse is given by $4aE(e)$. This fact is the namesake of elliptic integrals.

The complementary incomplete elliptic integrals of the first and second kind are respectively defined for $-1 < \zeta < 1$ and $0 < k < 1$ to be:

$$F'(\zeta; k) = F\left(\zeta; \sqrt{1 - k^2}\right) \tag{141}$$

$$E'(\zeta; k) = E\left(\zeta; \sqrt{1 - k^2}\right) \tag{142}$$

while the complementary complete elliptic integrals of the fist and second kinds are respectively defined for $-1 < \zeta < 1$ and $0 < k < 1$ to be:

$$K'(k) = K\left(\sqrt{1 - k^2}\right) \tag{143}$$

$$E'(k) = E\left(\sqrt{1 - k^2}\right) \tag{144}$$

The parameter $\sqrt{1 - k^2}$ is sometimes called the complementary modulus.

We note that by means of analytic continuation, the domains of definition of the all of the elliptic integrals can be extended to larger regions of $\mathbb{C}$ (in the case of complete elliptic integrals) or $\mathbb{C}^2$ (in the case of incomplete elliptic integrals). For example, the elliptic integrals of the second kind defined in Eq. 139, Eq. 140, Eq. 142, and Eq. 144 are analytic in $\mathbb{C}$ or $\mathbb{C}^2$.

Elliptic integrals cannot be expressed in terms of elementary functions. However, they can readily be calculated using tables [124] or softwares such as MATLAB [125] and Mathematica [126]. In Fig. 34, we plot, for illustrative purposes, the four complete elliptic integrals that were introduced in this appendix.

We finish this section by quoting an important result [127]:

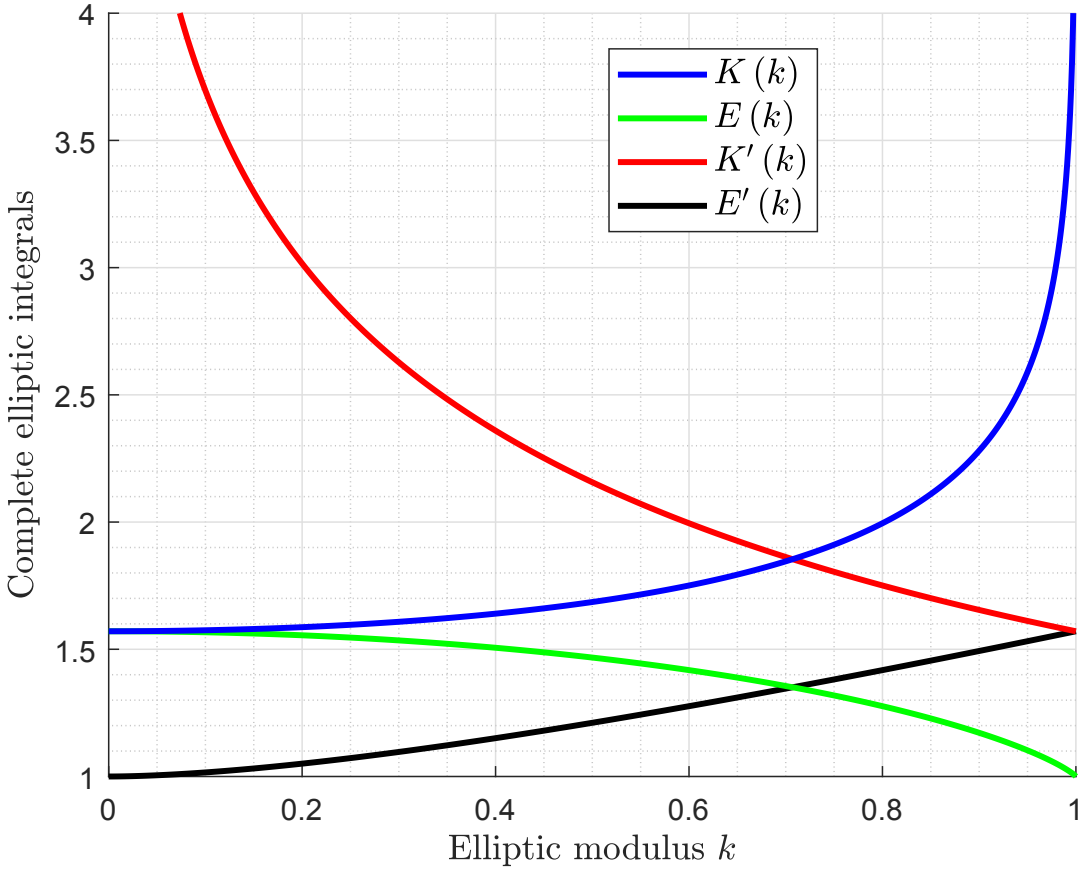


Figure 34 – Plots of the complete elliptic integrals defined in Eq. 138, Eq. 140, Eq. 143, and Eq. 144, as a function of the elliptic modulus k.

Legendre's relation: For any value of $k \in]0,1[$, one has:

$$[E'(k) - K'(k)]\,K(k) + K'(k)\,E(k) = \frac{\pi}{2} \tag{145}$$

After analytic continuation, Legendre's relation still holds for values of $k \in \mathbb{C}$ where the elliptic integrals appearing in Eq. 145 are defined.

References

[1] J. Taylor, H. Guo, and J. Wang. Ab initio modeling of quantum transport properties of molecular electronic devices. *Physical Review B*, 63(24):245407, 2001.

[2] A. Khakifirooz, O. M. Nayfeh, and D. Antoniadis. A simple semiempirical short-channel MOSFET current–voltage model continuous across all regions of operation and employing only physical parameters. *IEEE Transactions on Electron Devices*, 56(8):1674–1680, 2009.

[3] L. Wei, O. Mysore, and D. Antoniadis. Virtual-source-based self-consistent current and charge FET models: From ballistic to drift-diffusion velocity-saturation operation. *IEEE Transactions on Electron Devices*, 59(5):1263–1271, 2012.

[4] M. S. Lundstrom and D. A. Antoniadis. Compact models and the physics of nanoscale FETs. *IEEE Transactions on Electron Devices*, 61(2):225–233, 2014.

[5] L. Li, Y. Yu, G. J. Ye, Q. Ge, X. Ou, H. Wu, D. Feng, X. H. Chen, and Y. Zhang. Black phosphorus field-effect transistors. *Nature Nanotechnology*, 9(5):372–377, 2014.

[6] F. Liu, Y. Wang, X. Liu, J. Wang, and H. Guo. Ballistic transport in monolayer black phosphorus transistors. *IEEE Transactions on Electron Devices*, 61(11):3871–3876, 2014.

[7] F. Liu, Q. Shi, J. Wang, and H. Guo. Device performance simulations of multilayer black phosphorus tunneling transistors. *Applied Physics Letters*, 107(20):203501, 2015.

[8] S. Bohloul, L. Zhang, K. Gong, and H. Guo. Theoretical impurity-limited carrier mobility of monolayer black phosphorus. *Applied Physics Letters*, 108(3):033508, 2016.

[9] P. Drude. Zur elektronentheorie der metalle. *Annalen der Physik*, 306(3):566–613, 1900.

[10] P. Drude. Zur elektronentheorie der metalle; II. Teil. Galvanomagnetische und thermomagnetische effecte. *Annalen der Physik*, 308(11):369–402, 1900.

[11] G. D. Mahan. *Many-particle physics*. Springer Science & Business Media, 2013.

[12] H. Bruus and K. Flensberg. *Many-body quantum theory in condensed matter physics: an introduction*. Oxford University Press, 2004.

[13] G. Gildenblat. *Compact modeling: principles, techniques and applications*. Springer Science & Business Media, 2010.

[14] L. Nagel. SPICE (Simulation Program with Integrated Circuit Emphasis). `https://embedded.eecs.berkeley.edu/pubs/downloads/spice/index.htm`. Accessed on 2017-07-12.

[15] S. K. Saha. *Compact models for integrated circuit design: conventional transistors and beyond*. CRC Press, 2016.

[16] C.-T. Sah. Characteristics of the metal-oxide-semiconductor transistors. *IEEE Transactions on Electron Devices*, 11(7):324–345, 1964.

[17] J. Song, B. Yu, Y. Yuan, and Y. Taur. A review on compact modeling of multiple-gate MOSFETs. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 56(8):1858–1869, 2009.

[18] D. E. E. Rodriguez and A. G. A. Bernal. Development of a bottom-up compact model for Intel's ® high-K 45 nm MOSFET. In *IAENG Transactions on Engineering Technologies*, pages 123–134. Springer, 2013.

[19] *Sentaurus Device User Guide Version D-2010.03*. Synopsys Inc., Mountain View, CA, USA, 2010.

[20] S. Reggiani, M. Valdinoci, L. Colalongo, and G. Baccarani. A unified analytical model for bulk and surface mobility in Si n- and p-channel MOSFETs. In *Proceeding of the 29th European Solid-State Device Research Conference*, volume 1, pages 240–243. IEEE, 1999.

[21] S. Reggiani, M. Valdinoci, L. Colalongo, M. Rudan, G. Baccarani, A. D. Stricker, F. Illien, N. Felber, W. Fichtner, and L. Zullino. Electron and hole mobility in silicon at large operating temperatures. I. Bulk mobility. *IEEE Transactions on Electron Devices*, 49(3):490–499, 2002.

[22] M. E. Peskin and D. V. Schroeder. *An introduction to quantum field theory*. Westview Press, 1995.

[23] A. Korzeniowski, J. L. Fry, D. E. Orr, and N. G. Fazleev. Feynman-kac path-integral calculation of the ground-state energies of atoms. *Physical Review Letters*, 69(6):893, 1992.

[24] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Physical Review*, 136(3B):B864, 1964.

[25] W. Kohn and L. Sham. Quantum density oscillations in an inhomogeneous electron gas. *Physical Review*, 137(6A):A1697, 1965.

[26] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Physical Review*, 140(4A):A1133, 1965.

[27] N. Ashcroft and N. Mermin. *Solid state physics*. Brooks Cole, 1976.

[28] S. Datta. *Electronic transport in mesoscopic systems*. Cambridge University Press, 1997.

[29] Euclid of Alexandria. *Euclid's elements*. Oxyrhynchus Papyri, 300 BC.

[30] H. Haug and A.-P. Jauho. *Quantum kinetics in transport and optics of semiconductors*, volume 2. Springer, 2008.

[31] L. D. Landau and E. M. Lifshitz. *Course of theoretical physics, volume 5: Statistical physics*. Pergamon Press, 1969.

[32] T. Kobayashi, M. Bando, N. Kimura, K. Shimizu, K. Kadono, N. Umezu, K. Miyahara, S. Hayazaki, S. Nagai, Y. Mizuguchi, Y. Murakami, and D. Hobara. Production of a 100-m-long high-quality graphene transparent conductive film by roll-to-roll chemical vapor deposition and transfer process. *Applied Physics Letters*, 102(2):023112, 2013.

[33] K. S. Novoselov, A. K. Geim, S. V. Morozov, D. Jiang, Y. Zhang, S. V. Dubonos, I. V. Grigorieva, and A. A. Firsov. Electric field effect in atomically thin carbon films. *Science*, 306(5696):666–669, 2004.

[34] P. R. Wallace. The band theory of graphite. *Physical Review*, 71(9):622, 1947.

[35] J.-C. Charlier, P. Eklund, J. Zhu, and A. Ferrari. Electron and phonon properties of graphene: their relationship with carbon nanotubes. In *Carbon nanotubes*, pages 673–709. Springer, 2007.

[36] A. K. Geim and K. S. Novoselov. The rise of graphene. *Nature Materials*, 6(3):183–191, 2007.

[37] K. S. Novoselov, V. Fal'ko, L. Colombo, P. Gellert, M. Schwab, and K. Kim. A roadmap for graphene. *Nature*, 490(7419):192–200, 2012.

[38] F. Schwierz. Graphene transistors. *Nature Nanotechnology*, 5(7):487–496, 2010.

[39] Y. Zhang, T.-T. Tang, C. Girit, Z. Hao, M. C. Martin, A. Zettl, M. F. Crommie, Y. R. Shen, and F. Wang. Direct observation of a widely tunable bandgap in bilayer graphene. *Nature*, 459(7248):820, 2009.

[40] B. Lalmi, H. Oughaddou, H. Enriquez, A. Kara, S. Vizzini, B. Ealet, and B. Aufray. Epitaxial growth of a silicene sheet. *Applied Physics Letters*, 97(22):223109, 2010.

[41] M. Dávila, L. Xian, S. Cahangirov, A. Rubio, and G. Le Lay. Germanene: a novel two-dimensional germanium allotrope akin to graphene and silicene. *New Journal of Physics*, 16(9):095002, 2014.

[42] Z. Ni, H. Zhong, X. Jiang, R. Quhe, G. Luo, Y. Wang, M. Ye, J. Yang, J. Shi, and J. Lu. Tunable band gap and doping type in silicene by surface adsorption: towards tunneling transistors. *Nanoscale*, 6(13):7609–7618, 2014.

[43] K. F. Mak, C. Lee, J. Hone, J. Shan, and T. F. Heinz. Atomically thin MoS_2: a new direct-gap semiconductor. *Physical Review Letters*, 105(13):136805, 2010.

[44] B. Radisavljevic, A. Radenovic, J. Brivio, V. Giacometti, and A. Kis. Single-layer MoS_2 transistors. *Nature Nanotechnology*, 6(3):147–150, 2011.

[45] S. B. Desai, S. R. Madhvapathy, A. B. Sachid, J. P. Llinas, Q. Wang, G. H. Ahn, G. Pitner, M. J. Kim, J. Bokor, C. Hu, H.-S. P. Wong, and A. Javey. MoS_2 transistors with 1-nanometer gate lengths. *Science*, 354(6308):99–102, 2016.

[46] H. Fang, S. Chuang, T. C. Chang, K. Takei, T. Takahashi, and A. Javey. High-performance single layered WSe_2 p-FETs with chemically doped contacts. *Nano Letters*, 12(7):3788–3792, 2012.

[47] H. Wang, L. Yu, Y.-H. Lee, Y. Shi, A. Hsu, M. L. Chin, L.-J. Li, M. Dubey, J. Kong, and T. Palacios. Integrated circuits based on bilayer MoS_2 transistors. *Nano Letters*, 12(9):4674–4680, 2012.

[48] M. S. Fuhrer and J. Hone. Measurement of mobility in dual-gated MoS_2 transistors. *Nature Nanotechnology*, 8(3):146–147, 2013.

[49] S. P. Koenig, R. A. Doganov, H. Schmidt, A. Castro Neto, and B. Ozyilmaz. Electric field effect in ultrathin black phosphorus. *Applied Physics Letters*, 104(10):103106, 2014.

[50] Z. Guo, H. Zhang, S. Lu, Z. Wang, S. Tang, J. Shao, Z. Sun, H. Xie, H. Wang, X.-F. Yu, and P. K. Chu. From black phosphorus to phosphorene: basic solvent exfoliation, evolution of Raman scattering, and applications to ultrafast photonics. *Advanced Functional Materials*, 25(45):6996–7002, 2015.

[51] A. Carvalho, M. Wang, X. Zhu, A. S. Rodin, H. Su, and A. H. C. Neto. Phosphorene: from theory to applications. *Nature Reviews Materials*, 1:16061, 2016.

[52] H. Liu, A. T. Neal, Z. Zhu, Z. Luo, X. Xu, D. Tománek, and D. Y. Peide. Phosphorene: an unexplored 2D semiconductor with a high hole mobility. *ACS Nano*, 8(4):4033–4041, 2014.

[53] X. Chen, Y. Wu, Z. Wu, Y. Han, S. Xu, L. Wang, W. Ye, T. Han, Y. He, Y. Cai, and N. Wang. High-quality sandwiched black phosphorus heterostructure and its quantum oscillations. *Nature Communications*, 6:7315, 2015.

[54] D. J. Perello, S. H. Chae, S. Song, and Y. H. Lee. High-performance n-type black phosphorus transistors with type control via thickness and contact-metal engineering. *Nature Communications*, 6, 2015.

[55] D. J. Perello, S. H. Chae, S. Song, and Y. H. Lee. Corrigendum: High-performance n-type black phosphorus transistors with type control via thickness and contact-metal engineering. *Nature Communications*, 7, 2016.

[56] J. Qiao, X. Kong, Z.-X. Hu, F. Yang, and W. Ji. High-mobility transport anisotropy and linear dichroism in few-layer black phosphorus. *Nature Communications*, 5, 2014.

[57] F. Xia, H. Wang, and Y. Jia. Rediscovering black phosphorus as an anisotropic layered material for optoelectronics and electronics. *Nature Communications*, 5:4458, 2014.

[58] A. N. Rudenko and M. I. Katsnelson. Quasiparticle band structure and tight-binding model for single-and bilayer black phosphorus. *Physical Review B*, 89(20):201408, 2014.

[59] H. Ilatikhameneh, T. Ameen, B. Novakovic, Y. Tan, G. Klimeck, and R. Rahman. Saving Moore's law down to 1 nm channels with anisotropic effective mass. *Scientific Reports*, 6, 2016.

[60] Q. Wei and X. Peng. Superior mechanical flexibility of phosphorene and few-layer black phosphorus. *Applied Physics Letters*, 104(25):251915, 2014.

[61] G. Wang, G. Loh, R. Pandey, and S. P. Karna. Out-of-plane structural flexibility of phosphorene. *Nanotechnology*, 27(5):055701, 2015.

[62] F. Liu, Y. Zhou, Y. Wang, X. Liu, J. Wang, and H. Guo. Negative capacitance transistors with monolayer black phosphorus. *npj Quantum Materials*, 1:16004, 2016.

[63] A. Favron, E. Gaufrès, F. Fossard, A.-L. Phaneuf-L'Heureux, N. Y. Tang, P. L. Lévesque, A. Loiseau, R. Leonelli, S. Francoeur, and R. Martel. Photooxidation and quantum confinement effects in exfoliated black phosphorus. *Nature Materials*, 14(8):826, 2015.

[64] L. Hedin. New method for calculating the one-particle Green's function with application to the electron-gas problem. *Physical Review*, 139(3A):A796, 1965.

[65] H. Kroemer. Quasi-electric fields and band offsets: Teaching electrons new tricks. *International Journal of Modern Physics B*, 16(05):677–697, 2002.

[66] Y. Tsividis and C. McAndrew. *Operation and modeling of the MOS transistor*. Oxford University Press, 2011.

[67] S. Hofstein and F. Heiman. The silicon insulated-gate field-effect transistor. *Proceedings of the IEEE*, 51(9):1190–1202, 1963.

[68] P. Y. Yu and M. Cardona. *Fundamentals of semiconductors: Physics and materials properties*. Springer, 2010.

[69] C. Jacoboni, C. Canali, G. Ottaviani, and A. A. Quaranta. A review of some charge transport properties of silicon. *Solid-State Electronics*, 20(2):77–89, 1977.

[70] D. Frank, S. Laux, and M. Fischetti. Monte Carlo simulation of a 30 nm dual-gate MOSFET: How short can Si go? *IEDM Technical Digest*, 553, 1992.

[71] K. Natori. Ballistic metal-oxide-semiconductor field effect transistor. *Journal of Applied Physics*, 76(8):4879–4890, 1994.

[72] M. M. Tai. A mathematical model for the determination of total area under glucose tolerance and other metabolic curves. *Diabetes Care*, 17(2):152–154, 1994.

[73] J. H. Davies. *The physics of low-dimensional semiconductors: An introduction*. Cambridge University Press, 1997.

[74] A. Rahman and M. S. Lundstrom. A compact scattering model for the nanoscale double-gate MOSFET. *IEEE Transactions on Electron Devices*, 49(3):481–489, 2002.

[75] R. Venugopal, Z. Ren, and M. S. Lundstrom. Simulating quantum transport in nanoscale MOSFETs: Ballistic hole transport, subband engineering and boundary conditions. *IEEE Transactions on Nanotechnology*, 2(3):135–143, 2003.

[76] A. Rahman, J. Guo, S. Datta, and M. S. Lundstrom. Theory of ballistic nanotransistors. *IEEE Transactions on Electron Devices*, 50(9):1853–1864, 2003.

[77] J.-P. Colinge and C. A. Colinge. *Physics of semiconductor devices*. Springer Science & Business Media, 2005.

[78] A. Kastalsky and M. Shur. Conductance of small semiconductor devices. *Solid State Communications*, 39(6):715–718, 1981.

[79] M. S. Shur. Low ballistic mobility in submicron HEMTs. *IEEE Electron Device Letters*, 23(9):511–513, 2002.

[80] J. Wang and M. Lundstrom. Ballistic transport in high electron mobility transistors. *IEEE Transactions on Electron Devices*, 50(7):1604–1609, 2003.

[81] J. C. Maxwell. *A treatise on electricity and magnetism*, volume 1. Clarendon press, 1881.

[82] J. Thomson. Notes on recent researches in electricity and magnetism. *Mercury*, 8(10):18, 1893.

[83] A. E. H. Love. Some electrostatic distributions in two dimensions. *Proceedings of the London Mathematical Society*, 2(1):337–369, 1924.

[84] H. B. Palmer. The capacitance of a parallel-plate capacitor by the Schwartz-Christoffel transformation. *Electrical Engineering*, 56(3):363–368, 1937.

[85] D. J. Griffiths. *Introduction to electrodynamics*. Pearson, 2005.

[86] A. Bansal, B. C. Paul, and K. Roy. An analytical fringe capacitance model for interconnects using conformal mapping. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 25(12):2765–2774, 2006.

[87] A. Haque and M. Z. Kauser. A comparison of wave-function penetration effects on gate capacitance in deep submicron n- and p-MOSFETs. *IEEE Transactions on Electron Devices*, 49(9):1580–1587, 2002.

[88] I.-H. Tan, G. Snider, L. Chang, and E. Hu. A self-consistent solution of Schrödinger-Poisson equations using a nonuniform mesh. *Journal of Applied Physics*, 68(8):4071–4076, 1990.

[89] S. Mudanai, L. F. Register, A. F. Tasch, and S. K. Banerjee. Understanding the effects of wave function penetration on the inversion layer capacitance of n-MOSFETs. *IEEE Electron Device Letters*, 22(3):145–147, 2001.

[90] S. Luryi. Quantum capacitance devices. *Applied Physics Letters*, 52(6):501–503, 1988.

[91] D. J. Griffiths. *Introduction to quantum mechanics*. Cambridge University Press, 2016.

[92] J. Xia, F. Chen, J. Li, and N. Tao. Measurement of the quantum capacitance of graphene. *Nature Nanotechnology*, 4(8):505–509, 2009.

[93] A. Ortiz-Conde, F. G. Sánchez, J. J. Liou, A. Cerdeira, M. Estrada, and Y. Yue. A review of recent MOSFET threshold voltage extraction methods. *Microelectronics Reliability*, 42(4):583–596, 2002.

[94] D. K. Schroder. *Semiconductor material and device characterization*. John Wiley & Sons, 2006.

[95] M. Tsuno, M. Suga, M. Tanaka, K. Shibahara, M. Miura-Mattausch, and M. Hirose. Physically-based threshold voltage determination for MOSFETs of all gate lengths. *IEEE Transactions on Electron Devices*, 46(7):1429–1434, 1999.

[96] H. Haddara. *Characterization methods for submicron MOSFETs*, volume 352. Springer Science & Business Media, 2012.

[97] H.-S. Wong, M. H. White, T. J. Krutsick, and R. V. Booth. Modeling of transconductance degradation and extraction of threshold voltage in thin oxide MOSFETs. *Solid-State Electronics*, 30(9):953–968, 1987.

[98] K. Aoyama. A method for extracting the threshold voltage of MOSFETs based on current components. In *Simulation of semiconductor devices and processes*, pages 118–121. Springer, 1995.

[99] P. Karlsson and K. Jeppson. An efficient method for determining threshold voltage, series resistance and effective geometry of MOS transistors. *IEEE Transactions on Semiconductor Manufacturing*, 9(2):215–222, 1996.

[100] G. Merckel and A. Rolland. A compact CAD model for amorphous silicon thin film transistors simulation. I. DC analysis. *Solid-State Electronics*, 39(8):1231–1239, 1996.

[101] G. Merckel and A. Rolland. A compact CAD model for amorphous silicon thin film transistors simulation. II. Transient non-quasi-static analysis. *Solid-State Electronics*, 39(8):1241–1245, 1996.

[102] A. Ortiz-Conde, E. G. Fernandes, J. Liou, M. R. Hassan, F. Garcia-Sanchez, G. De Mercato, and W. Wong. A new approach to extract the threshold voltage of MOSFETs. *IEEE Transactions on Electron Devices*, 44(9):1523–1528, 1997.

[103] S. M. Sze and K. K. Ng. *Physics of semiconductor devices*. John Wiley & Sons, 2006.

[104] K. P. Cheung. On the 60 mV/dec at 300 K limit for MOSFET subthreshold swing. In *International Symposium on VLSI Technology Systems and Applications*, pages 72–73. IEEE, 2010.

[105] M. Miura-Mattausch, H. Ueno, M. Tanaka, H. Mattausch, S. Kumashiro, T. Yamaguchi, K. Yamashita, and N. Nakayama. HiSIM: a MOSFET model for circuit simulation connecting circuit performance with technology. In *International IEDM Electron Devices Meeting*, pages 109–112. IEEE, 2002.

[106] G. E. Moore. Cramming more components onto integrated circuits. *Proceedings of the IEEE*, 86(1):82–85, 1998.

[107] T. N. Theis and P. M. Solomon. In quest of the "next switch": prospects for greatly reduced power dissipation in a successor to the silicon field-effect transistor. *Proceedings of the IEEE*, 98(12):2005–2014, 2010.

[108] A. M. Ionescu and H. Riel. Tunnel field-effect transistors as energy-efficient electronic switches. *Nature*, 479(7373):329–337, 2011.

[109] J. Appenzeller, Y. Lin, J. Knoch, and P. Avouris. Band-to-band tunneling in carbon nanotube field-effect transistors. *Physical Review Letters*, 93(19):196805, 2004.

[110] C. Qiu, F. Liu, M. Xiao, B. Deng, L. Xu, Z. Zhang, J. Wang, H. Guo, H. Peng, and L. Peng. Sub-60 mV/dec carbon nanotube field-effect transistor with a height-tunable Schottky barrier at source. 2017.

[111] F. Liu, C. Qiu, Z. Zhang, L. Peng, J. Wang, and H. Guo. Dirac electrons at the source: breaking the 60 mV/dec switching limit. 2017.

[112] L. Matthes, O. Pulci, and F. Bechstedt. Massive Dirac quasiparticles in the optical absorbance of graphene, silicene, germanene, and tinene. *Journal of Physics: Condensed Matter*, 25(39):395305, 2013.

[113] Y. Zhu, L. Liu, and H. Guo. Quantum transport theory with nonequilibrium coherent potentials. *Physical Review B*, 88(20):205415, 2013.

[114] R. Dingle. The Fermi-Dirac integrals. *Applied Scientific Research*, 6(1):225–239, 1956.

[115] J. Blakemore. Approximations for Fermi-Dirac integrals, especially the function $\mathscr{F}_{\frac{1}{2}}(\eta)$ used to describe electron density in a semiconductor. *Solid-State Electronics*, 25(11):1067–1076, 1982.

[116] D. Bednarczyk and J. Bednarczyk. The approximation of the Fermi-Dirac integral $\mathscr{F}_{\frac{1}{2}}(\eta)$. *Physics Letters A*, 64(4):409–410, 1978.

[117] X. Aymerich-Humet, F. Serra-Mestres, and J. Millan. An analytical approximation for the Fermi-Dirac integral $\mathscr{F}_{\frac{3}{2}}(\eta)$. *Solid-State Electronics*, 24(10):981–982, 1981.

[118] X. Aymerich-Humet, F. Serra-Mestres, and J. Millan. A generalized approximation of the Fermi-Dirac integrals. *Journal of Applied Physics*, 54(5):2850–2851, 1983.

[119] P. Van Halen and D. Pulfrey. Accurate, short series approximations to Fermi-Dirac integrals of order $-\frac{1}{2}$, $\frac{1}{2}$, 1, $\frac{3}{2}$, 2, $\frac{5}{2}$, 3, and $\frac{7}{2}$. *Journal of Applied Physics*, 57(12):5271–5274, 1985.

[120] P. Van Halen and D. Pulfrey. Erratum: Accurate, short series approximations to Fermi-Dirac integrals of order $-\frac{1}{2}$, $\frac{1}{2}$, 1, $\frac{3}{2}$, 2, $\frac{5}{2}$, 3, and $\frac{7}{2}$. *Journal of Applied Physics*, 59(6):2264–2265, 1986.

[121] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical recipes: The art of scientific computing*. Cambridge University Press, 2007.

[122] J. McDougall and E. C. Stoner. The computation of Fermi-Dirac functions. *Philosophical Transactions of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 237(773):67–104, 1938.

[123] E. M. Stein and R. Shakarchi. *Princeton lectures in analysis. II. Complex analysis*. Princeton University Press, 2003.

[124] I. N. Bronshtein, K. A. Semendyayev, G. Musiol, and H. Muehlig. *Handbook of mathematics*. Springer, 2015.

[125] *MATLAB version R2017a*. The MathWorks Inc., Natick, MA, 2017.

[126] *Mathematica version 11.1*. Wolfram Research Inc., Champaign, IL, 2017.

[127] P. Duren. The Legendre relation for elliptic integrals. In *Paul Halmos: Celebrating 50 years of mathematics*, volume 50, pages 305–315. Springer, 1991.