2017

Theoretical and Experimental Analysis of Optical Properties of Defects in GaN:

Ibrahima Castillo Diallo

Theoretical and Experimental Analysis of Optical Properties of Defects in GaN:

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Nanoscience and Nanotechnology at Virginia Commonwealth University.

By:

Ibrahima Castillo Diallo

M.S. in Physics/Applied Physics, Virginia Commonwealth University, 2013

B.S. in Physics, Virginia Commonwealth University, 2011

Major Director:

Dr. Denis O. Demchenko

Associate Professor, Department of Physics

Virginia Commonwealth University

Richmond, Virginia, 23284

July 10, 2017

Acknowledgement

I would like to thank:

- Allah: for EVERYTHING.
- My family: for emotional support and definitely financial support.
- Dr. Denis Demchenko: for kind of not giving up on me throughout the past painful seven years.
- Dr. Michael Reshchikov: for teaching me in a very calm manner and paying my fees this semester or else I could have probably ended up in the streets.
- Dr. Arthur Reber: too many things to thank him about.
- Dr. Tom Mac Mullen: for making me feel like a mentally challenged student every time I had questions to ask.
- Dr. Bishop: for not giving up in me "initially" although I was not the hardest working student.
- Samar and Alarki family: too many things to thank them about.
- Nahla Albarakati: for couple of things on how to pray properly.
- Daniel Guest: for the free rides, the free lessons on the American Caucasian culture and possible first friendship with an American fellow.
- Dr. Anthony Pedicini: for sharing OriginPro with his other broke fellow and depressing conversations about possible depressing future lives as wannabe physicists.
- Cara Frame: for her positive attitude and covering my early 8 AM laboratories.
- Amanda Steck (and Wesley): for covering my labs, seriously editing my statement of purpose for job applications and random interesting conversations, i.e. the digestive system of cows, the singularity of the word "priority", etc…
- Joe Ferguson: for introducing me to the awesome comedian Louie C.K. and a random conversation about the university of Mary Washington.
- Chris Angevine and others: for kind of covering my labs. I do not quite remember why I should be thanking them but they somehow should be here.
- The VCU Center for High Performance Computing
- The National Science Foundation (NSF) Grant (DMR-1410125)

Table of Contents

List of Figures and Tables

Abstract

THEORETICAL AND EXPERIMENTAL ANALYSIS OF OPTICAL PROPERTIES OF DEFECTS IN GaN

By Ibrahima Castillo Diallo, Doctor of Philosophy in Nanoscience and Nanotechnology

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Nanoscience and Nanotechnology at Virginia Commonwealth University.

Virginia Commonwealth University, 2017

Major Director: Denis O. Demchenko, Associate Professor, Department of Physics

In this dissertation, we first present a brief overview of various theoretical approaches used to examine the electronic structure and optical properties of defects in GaN. Using the Heyd-Scuseria-Ernzherof (HSE06) hybrid functional method along with photoluminescence experimental measurements, we analyze the properties of intrinsic defects such as vacancies, interstitials, antisites, and common complexes. By using configurational coordinate diagrams, we estimate the likelihood of defects to be radiative or non-radiative. Our calculations show that gallium vacancies exhibit a large magnetic moment in the neutral charge state and are most likely non-radiative. This is in contrast to nitrogen vacancies, which are probable sources of the experimentally observed green luminescence band (labeled GL2) peaking at 2.35 eV in undoped

GaN. We also investigate the correlation between the observed infrared PL bands created in 2.5 MeV electron-irradiated GaN samples and the formation of native defects. It is found that gallium-nitrogen divacancies are possible sources of the broad PL band peaking at 0.95 eV while interstitial gallium is likely to be responsible for the narrow infrared PL band centered around 0.85 eV, with a phonon fine structure at 0.88 eV.

In addition to native defects, we also investigate the blue luminescence band (BL2) peaking at 3.0 eV that is observed in high-resistivity GaN samples. Under extended ultraviolet (UV) light exposure, the BL2 band transforms into the yellow luminescence (YL) band with a maximum at 2.2 eV. Our calculations suggest that the BL2 band is related to a hydrogen-carbon defect complex, either C_NO_N-H_i or C_N-H_i. The complex creates defect transition level close to the valence band, which is responsible for the BL2 band. Under UV illumination, the complex dissociates, leaving as byproduct the source of the YL band (C_NO_N or C_N) and interstitial hydrogen.

In conclusion, theoretical predictions of thermodynamic and optical transitions of defects in GaN via the HSE06 method are found to be within less than 0.2 eV when compared to experiment. Hence the HSE formalism is a powerful tool for the identification and characterization of the microscopic sources of observed PL bands in GaN.

List of Common Abbreviations

A: Electron Affinity

BL2: Blue Luminescence

CBM/VBM: Conduction Band Minimum/Valence Band Maximum

CCD: Configuration Coordinate Diagram

DFT: Density Functional Theory

e^-, e^*: electron, weakly localized electron

FC: Franck-Condon

Ga_i: Interstitial Ga

Ga_N: Gallium Antisite

GaN: Gallium Nitride

GGA: Generalized Gradient Approximation

h^+: hole

HF: Hartree-Fock

H_i: Interstitial Hydrogen

HK: Hohenberg-Kohn

HSE: Heyd-Scuseria-Ernzherof

HUMO/LUMO: Highest Occupied Molecular Orbital/ Lowest Unoccupied Molecular Orbital

HVPE: Hydride Vapor Phase Epitaxy

IP: Ionization Potential

IR: Infrared

J: Janak

KS: Kohn-Sham

LDA, LSDA: Local Density Approximation, Local Spin Density Approximation

LZ: Lany-Zunger

Mg_{Ga}: Magnesium substituting Gallium

MOCVD: Metalorganic Chemical Vapor Deposition

MP: Makov-Payne

N_{Ga}: Nitrogen Antisite

N_i-N_i: Nitrogen Split Interstitial

ODEPR: Optically Detected Electron Paramagnetic Resonance

ODMR: Optically Detected Magnetic Resonance

O_N: Oxygen substituting Nitrogen

PA: Potential Alignment

PAS: Positron Annihilation Spectroscopy

PBE, PBE0: Perdew-Burke-Ernzerhof, Hybrid Perdew-Burke-Ernzerhof functional

PL: Photoluminescence

PW: Plane Waves

TF, TFD: Thomas-Fermi, Thomas-Fermi-Dirac

V_{Ga}: Gallium Vacancy

$V_{Ga}V_N$: Gallium-Nitrogen divacancy

V_N: Nitrogen Vacancy

XC: Exchange-correlation

YL: Yellow Luminescence

ZPL: Zero-phonon Line

0. Introduction

Gallium nitride (GaN) has emerged as an important wide bandgap semiconductor because of its proven success in the fabrication of light emitting devices (LEDs),[1] lasers,[2] solar cells[3] and its potential for high-temperature/power applications. However, more research still needs to be performed so that high quality and cost effective GaN may be synthesized. An important issue associated with the growth of GaN is the creation of defects, which may prove damaging to the electrical and optical properties of the host material. These point defects include native defects (vacancies, interstitial and antisites), intentional or unintentional impurities and complexes involving various combinations of isolated point defects. In LEDs, the generation of light is hindered by point defect assisted non-radiative recombination (Shockley-Reed-Hall). Furthermore, point defects tend to cause a reduction in responsivity and an increase of noise in detectors.[4] Although tremendous progress has been made in the efficient fabrication of electronic devices based on GaN, a theoretical understanding of the optical and electronic properties of defects in GaN still remains unclear. One of the main reasons is the large amount of contradictory results produced by both theory and experiment. Hence, theoretical analysis of defects in GaN, based on first-principles calculations that could complement experiments and therefore serve as a predictive tool is necessary.

Kohn Sham Density Functional Theory (DFT)[5] has proven to be a prevailing tool for analyzing and understanding defect energetics and electronic structure in semiconductors. Good progress for approximating the crucial exchange-correlation (XC) energy from the Kohn-Sham approach has been made in the last decades. One of the most relevant formalisms for the analysis of the electronic structure of spin systems is the local spin density approximation (LSDA)[6]. Although the construction of the LSDA is based on the uniform electron gas, it has provided

acceptable results in crystal structure, bond lengths and vibrational frequencies[7,8] in both homogenous and inhomogeneous systems. However, the LSDA leads to major drawbacks such as the inability to describe the magnetic configuration of transition metals, the lack of cancelation of self-interaction which is crucial for strongly localized states and the severe underestimation of the band gap in semiconductors and insulators. Such shortcomings have stimulated ideas for the creation of improved functionals such as non-empirical generalized gradient approximations[9] (GGA). Several sophisticated adaptations of GGA have been developed in the last decades[10,11,12,13], but the most commonly used version is the Perdew-Burke-Ernzerhof (PBE)[14] method that employs both the density and its gradient at each point in space. Both GGA and LSDA were derived in the limits of the homogeneous electron gas theory and are therefore expected to be useful for systems with slowly varying charge densities[15]. These formalisms have provided satisfactory results for the computation of molecular binding energies, atomic ionization energies and geometrical structure of molecules and solids. The partial error cancellation in the exchange and correlation energy parts integrated in both first-principles calculations methods provided the accuracy required for DFT to be used in solid states physics as well as in chemistry[7,16]. Nevertheless, the underestimation of the band gap in solids remains one of the major drawbacks of both LSDA and GGA formalisms.[17,18]

To remedy the band gap problem and several other unphysical results of LSDA and GGA, much effort has been put into the improvement of the XC-parameter[19,12,20]. One of the most fruitful approximations in the computation of band gap is described by a combination of Green Function and screened Coulomb interaction, often referred to as the GW method[21]. However, the GW method happens to be computationally expensive for complex systems. Alternative approach that rectified the band gap problem was the construction of a hybrid

functional theory that contains a mixture of a certain amount of non-local Fock exchange and a part of *local/semilocal* LSDA/PBE exchange[22,23,24]. Due to the periodicity of the lattice, that generates a long range Hartree-Fock (HF) exchange interactions, the use of hybrid functional in solid states physics has been inadequate[15,25]. Significant progress into reducing the computational effort of calculating the long range Fock exchange has been achieved by the development of the range-separated Heyd-Scuseria-Ernzherof (HSE03) functional[26]. This newly derived formalism separates the Fock exchange into short-range and long-range components. The short-range exchange energy is made of 25% of HF and 75% of PBE, while the long range exchange energy and correlation term are entirely represented by the semi-local PBE method.[26] Such modifications would cause major corrections to the electronic properties of the system and can therefore be used to compute improved band gaps, bulk moduli and atomization energies of solids including semiconductors and metals[26,8,27,28,29]. An in-depth study of electronic structure of solids has not yet reached its peak with the development of HSE03 formalism. More detailed analysis of energetics of defects in semiconductors[30], vibrational frequencies of lattice[31,32], and optical properties of semiconductors[33,34,35,36,37] have been recently performed with the creation of the HSE06[31] approximation. In the HSE06 formalism, further tuning of the screening parameter is performed in such way that improved agreement with experimental data is obtained.

In the first part of this dissertation, we provide a brief description of the nature of defects we shall be investigating, followed by a literature review of intrinsic defects in GaN and the observed BL2 band in high-resistivity GaN. In sections 2-3, we present an overview of the methodology used to perform first-principles calculations of defects and impurities in semiconductors. In the last section of the dissertation, we investigate the electronics and optics of

intrinsic defects in GaN and the microscopic origin of the BL2 band in high-resistivity GaN samples.

Section 1. Doping and Photoluminescence in GaN

The introduction of defects in a host crystal alters the characteristics of the material in various ways. Because of the multiplicity of imperfections that can occur within the lattice, we will first describe ways of classifying them and later discuss their role in various observed PL bands in GaN.

1.1. Creation of Native Point Defects

In this section, the brief analysis on some characteristics of defects in semiconductors is based on the detailed review written by S. T. Pantelides (1979)[38]. Native defects are intrinsic imperfections that are formed within the "pure" host lattice and can either be:

- point defects, which correspond to the imperfect location of atoms
- planar defects, which describe misplaced planes of atoms
- line defects which correspond to misplaced line of atoms.

Since we are only investigating point defects as lattice-type defects in GaN, the characteristics of either planar or line defects shall not be discussed in this dissertation. Native point defects usually occur in vacant and interstitial sites. In case of vacancies, atoms are missing from their regular atomic site. Interstitial point defects describe extra-atoms that occupy highly symmetric interstitial sites in the crystal. More details regarding the exact location of various vacant and interstitial sites in GaN is given in section 4.1. In addition to native point defects, foreign or external impurities may occur inside the crystal and can be classified in terms of their crystal locations.

1.2. Incorporation of Impurities in the Crystal Lattice

Foreign impurities may occur in either interstitial locations or substitutional sites in which case the impurity replaces the host atom. Substitutional atoms that generally have more valence electrons than the host atoms are called donors since they must donate electrons to the host atoms in order to fulfill local bonding requirements. While on the other hand, substitutional atoms that possess less valence electrons are called acceptors since they must accept electrons from the host atoms in order to bond with their nearest neighbors. Our definition of donors and acceptors is quite general so far, but more descriptive details are given in section 3.4.1.

Defects may be called shallow when their respective energy levels are very close to the conduction or valence band edges. On the contrary, defects are called deep when their respective energy levels are far from the band edges. Shallow and deep defects play an important role in the properties of a given material and will be discussed in the following section.

In addition to isolated defects, one must also notice that it is sometimes possible for defects to interact with one another and form complexes. The simplest situation is a complex pair consisting of two isolated impurities occupying neighboring sites, a vacancy defect and the nearest neighbor impurity, and two vacancies defects in neighboring sites. More details on swirl defect complexes in Si and dark-like defect complexes in GaAs-GaAlAs can be found in reviews written by De Kock (1973)[39] and Petrov and Hartmann (1973)[40], respectively.

1.3. Shallow and Deep Impurities in Semiconductors

Throughout this dissertation, we will be using first-principles electronic structure calculations of defects in GaN bulk. In other words, we are solving the Schrödinger equation in GaN lattice containing a defect, using periodic boundary conditions. The use of periodically repeated supercells[41] provides a physically appealing description of the defect and its closest neighbors. These supercells are composed of numerous primitive unit cells which contain a single defect. Even though the supercell technique precisely describes the local arrangements of bonding between atoms and the defect crystal structure, it unfortunately introduces some drawbacks that need to be corrected such as the divergence of the Coulomb energy for charged defects[41], the possible band-filling error[42] and the potential alignment[42] for charged impurities.

The role that most impurities and defects play in a given semiconductor often depends on the concentration in which they can be incorporated in the material and the kind of localized states they create in the band gap. In fact, there are fundamental differences in defects' properties depending on the proximity of the defects to the band edges.

Shallow levels are characterized by their extreme closeness to the band edges at room temperature. At such levels, impurities have ionization energies comparable to $k_B T$ and therefore will play a crucial role in controlling conductivity. The case of neutral shallow donor impurities in semiconductors requires careful investigation because of its weakly localized characteristic. Based on effective mass theory, the wave functions of shallow defects are Hydrogen-like and thus relatively spread out in real space. However within the supercell formalism, typical supercell is not large enough to completely encompass such widespread wave function. In fact, supercells that could contain a weakly localized wave function of a shallow impurity would contain tens of thousands of atoms and would therefore be computationally

“impossible”. When a shallow defect is computed in relatively small supercell, the impurity band level becomes resonant with the CBM. As a result, the electron located at the impurity level will drop into the conduction band maximum which becomes a delocalized perturbed host state. Our neutral system will therefore be composed of a positive ion surrounded by completely delocalized electron charge which is similar to the case of a positive ion sitting in uniform negatively charged compensating background. In order to correct such unphysical interactions, one must include a special correction scheme, which will be discussed later in Section 3.3.2.1.

On the other hand, deep impurities have localized wave functions and therefore provide levels inside the band gap that could increase the probability of recombination between the electrons and holes. In addition to emitting phonons during the recombination process (see section 3.4.2.2), photons may also be produced and PL could therefore be measured.

Now that we have provided a brief introduction of the type of defects we will be looking into, a literature review describing previous experimental and theoretical results related to native defects in GaN and the observed BL2 band in high-resistivity GaN samples is given in the next two sections.

1.4. Experimental and Theoretical Background on Native Defects in Gallium Nitride

Knowledge of the electronic properties of defects is important to assess their formation during material fabrication and processing. Of particular interest, are the native point defects of GaN, such as vacancies and interstitial defects, which may form naturally during sample growth, or can be formed as a result of electron irradiation.

Some native defects, such as vacancies, have been extensively studied experimentally throughout the past two decades, while others have been less scrutinized. Isolated Ga vacancies[43] (V_{Ga}) and the possible complexes with oxygen donors (V_{Ga}-O_N)[44,45] have been experimentally investigated by positron annihilation spectroscopy (PAS) in bulk GaN crystals and epitaxial GaN samples. Based on these experiments, Saarinen et al.[43] concluded that Ga vacancies are negatively charged in both bulk GaN crystals and layers, playing a major role in electrical compensation of *n*-type GaN. Later, Oila et al.[46] using PAS demonstrated that negatively charged Ga vacancy is the most stable acceptor defect in *n*-type GaN grown by hydride vapor phase epitaxy (HVPE). Optical properties of Ga vacancies and related complexes were also investigated, and an apparent correlation of the yellow luminescence (YL) intensity with the concentration of Ga vacancies was suggested.[43,47,48] However, other experimental studies have shown that vacancies of Ga alone do not account for the YL observed in GaN, with the possibility that carbon-related defects are involved as well.[49,50]

Among other native defects, interstitial Ga has also been extensively studied experimentally. In 2.5 MeV electron-irradiated GaN epilayers, optically detected magnetic resonance (ODMR) signals at 1.5 K were observed in PL bands peaking at ~0.85 eV and ~0.95 eV.[51] Based on the obtained resolved hyperfine structure, it was suggested that the microscopic origin of one of the

ODMR signals was a complex formed by interstitial Ga and another unidentified defect. Further ODMR studies on the 0.85 eV PL band were performed by Buyanova et al.[52] at 2 K and 30 K. It was shown that the defect responsible for the 0.85 eV PL band has its principal symmetry axis along the *c*-axis of wurtzite GaN. Bozdog et al.[53] performed ODMR studies at room temperature on electron–irradiated GaN samples, also observing the two infrared (IR) bands centered at 0.85 eV and 0.95 eV. Two out of the three observed EPR signals revealed strong hyperfine interaction with a single Ga nucleus, suggesting that they are related to the isolated interstitial Ga. More recent ODMR studies on electron-irradiated GaN samples were performed at various temperatures by Watkins et al.[54] and Chow et al.,[55,56] where only the broad 0.95 eV PL band was observed from 4.2 K (*in-situ* irradiation) up to 295 K. Two of the obtained ODMR signals were attributed to the isolated interstitial Ga located in either octahedral or tetrahedral sites in GaN. Above 295 K, the 0.95 eV PL band lost 80-90% of its intensity while the sharp 0.85 eV PL band and its characteristic ODMR signal started emerging. The changes of ODMR signals band were explained by the possible migration of interstitial Ga near the vacant Ga site, from which they were created by electron irradiation. However, despite thorough experimental studies of the effect of electron irradiation on the properties of GaN, the microscopic origin of the near IR PL bands peaking at 0.85 eV and 0.95 eV is still not entirely clear.

Electrical and optical properties of nitrogen vacancies (V_N) and related complexes were also investigated in recent experiments, mostly in Mg-doped *p*-type GaN.[57,58,35] Nitrogen vacancies associated with magnesium acceptors (V_N-Mg_{Ga}), were identified by Hautagankas et al.[57] using positron annihilation spectroscopy (PAS) in Mg-doped GaN. It was suggested that V_N-Mg_{Ga} complexes behave as compensating centers in *p*-type GaN, and that vacancies of Ga and N are abundant in both *n*-type and *p*-type GaN. Using a combination of ODMR and PAS

experiments, Zeng et al.[58] suggested that the red PL band peaking at 1.80 eV in Mg-doped GaN is caused by the donor-acceptor pair recombination, where electrons localized on deep donor V_N-Mg_{Ga} complexes recombine with holes on deep V_{Ga} acceptors. Further investigation of the optical properties of nitrogen vacancies performed by Reshchikov et al.[35] using PL spectroscopy, proposed that V_N is the best candidate for the green luminescence band (labeled GL2) occurring at 2.35 eV in high-resistivity undoped and Mg-doped GaN. However, it was noted that N vacancy is present with relatively low concentrations in both types of samples.

While positron annihilation allows detection of vacancies, the experimental identification and characterization of other types of native defects (interstitials, antisites) have been proven difficult. Other experimental techniques, such as ODMR or PL spectroscopy provide only certain partial information about the nature and properties of these native defects. Therefore, sparse (and in some cases contradictory) experimental data suggests that revisiting the basic questions of native defects in GaN from the theoretical point of view is in order.

Theoretical investigations of native defects in GaN have been performed using various methods, such as tight binding approximation[59], empirical potential methods[60], ab-initio Molecular Dynamics[61], and the density functional theory (DFT).[62,63,64,65] Early atomistic theoretical studies of the electronic structure of vacancies and antisites in GaN were performed by Jenkins and Dow[59] using the tight binding approximation. It was shown that N vacancy is a shallow donor in GaN, also creating a doubly occupied deep level within the band gap. Further analysis on intrinsic defects using DFT was performed in Refs. [62-64,66] where it was found that the most stable native defects in GaN are the compensating defects, i.e. donor nitrogen vacancies in *p*-type GaN and acceptor Ga vacancies in *n*-type GaN. Antisites and interstitial native defects were found to be less favorable. Other theoretical calculations based on the local

density approximation (LDA)[5] using scissor corrections agreed well with previous predictions of the deep acceptor properties of V_{Ga} but unexpectedly predicted the existence of both donor and acceptor states (up to 3- charge state) for nitrogen vacancy.[65] Similar DFT calculations performed in Ref. [67] suggested that V_N should be the dominant defect in both *p*-type and *n*-type GaN annealed samples.

Although less studied than vacancies, antisites and native interstitials were also addressed by theory, producing varying results. By using both DFT and empirical potential methods, Gao et al.[60] obtained the formation energy of neutral antisite N_{Ga} which agreed with the value obtained by Gorczyca et al.[68] where the Green's function technique[69] was used. However, these results were significantly higher than the DFT values obtained by Neugebauer et al.[62] In the case of neutral Ga interstitial, the results obtained using empirical potential methods (Ref. [60]) also differ from previous ab-initio calculations performed in Refs. [62-64, 66**].**

In addition to isolated intrinsic defects, properties of di- and trivacancies in bulk GaN were also investigated. DFT calculations performed in Ref. [65] showed that Ga-N mixed divacancy ($V_{Ga}V_N$) behaved as a deep acceptor center. The calculations yielded the divacancy formation energy lower than that of V_{Ga}, with a substantial binding energy of 2.34 eV (for Fermi energy $E_F > 1.5\text{eV}$). However, more recent generalized gradient approximation (GGA)[9] calculations performed by Gohda et al.[70] and Puzyrev et al.[71] suggested that divacancies display both donor and acceptor properties and are less energetically stable than V_{Ga} in *n*-type GaN (with the energy difference ~1 eV). Trivacancies ($V_{Ga}V_NV_{Ga}$) were also investigated in Ref. [70], where it was shown that $V_{Ga}V_NV_{Ga}$ are unstable in *p*-type GaN but exhibit formation energy identical to that of divacancies in *n*-type GaN.

Calculations of defect properties using local (or semi-local) approximations to the DFT are prone to inaccuracies due to the known underestimation of the band gap. Recent studies of vacancies[72,73,74] in GaN used non-local screened exchange LDA (sX-LDA) and Heyd-Scuseria-Ernzerhof (HSE) hybrid functional, which can circumvent the band gap problem. These studies showed substantial differences in the electronic structure of defects compared to the results of the (semi)local approximations to the DFT. While considered a step forward, the results obtained with these new computational methods are also a subject of interpretation. For example, using HSE calculations, Yan et al.[72] suggested that nitrogen vacancy could be a possible source of the YL band observed in Mg-doped GaN. On the other hand, hybrid functional calculations based on sX-LDA performed by Gillen and Robertson,[73] associated the YL with the 0/- transition level of the gallium vacancy. Recent HSE calculations performed by Lyons et al.[74] proposed that while transitions via 2-/3- level of isolated V_{Ga} are most likely non-radiative, V_{Ga} complexes with oxygen and hydrogen can contribute to the YL band in GaN. Most recent HSE calculations performed in Refs. [75,76] also describe the energetics of native defects in GaN. Both calculations indicated that in Ga-rich conditions, V_N is the most energetically stable native defect.

Overall, both experiment and theory have produced large amounts of widely varying results. On one hand, only limited information for some defects is accessible from experiments, while on the other hand, different theoretical approaches often yield conflicting results. Consequently, many details of the electronic properties of native defects in GaN remain unclear. In addition to native defects, external defects also play a major role in the electrical and optical properties of GaN semiconductors.

1.4. Blue Luminescence (BL2) in GaN

For semi-insulating GaN grown by metalorganic chemical vapor deposition (MOCVD), two defect-related luminescence bands are typically observed: the omnipresent yellow luminescence (YL) band with a maximum at 2.2 eV and a broad band in the blue spectral region. The latter has a maximum at 3.0 eV and is labeled BL2 to distinguish it from the blue luminescence (BL) band with a maximum at 2.9 eV which is assigned to the zinc substituting gallium (Zn_{Ga}) acceptor in undoped or Zn-doped GaN.[4,77] In the literature, the BL and BL2 bands are often undistinguished, and both called "the blue band" because they have similar positions. However, detailed studies reveal unique features which make it possible to reliably distinguish these two defect-related bands.[78,79,80] The BL2 band usually has a characteristic fine structure at low temperatures, with the zero-phonon line (ZPL) at 3.33-3.34 eV and a few phonon replicas corresponding to the LO phonon mode (91 meV) and a local or pseudo-local mode (36 meV). On the other hand, the Zn-related BL band has a ZPL at 3.10 eV. The BL2 band is observed only in high-resistivity or semi-insulating GaN, while the BL band is also observed in conductive GaN samples. With increasing temperature, the BL2 band is quenched at temperatures above 100 K, revealing an activation energy of 0.15 eV. The activation energy is consistent with the ZPL at 3.33 eV for this band and indicates that the transition level for the related defect is located at 0.15 eV above the valence band maximum (VBM). The BL band is quenched at $T > 200$ K with an activation energy of about 0.35 eV. The activation energy is consistent with the ZPL at 3.10 eV and it corresponds to the position of the transition level for the Zn_{Ga} acceptor at 0.35-0.40 eV above the VBM. Comparing these properties, we can conclude that a blue band with a maximum at 3.0 eV observed in high-resistivity or semi-insulating GaN grown by the MOCVD technique,[6-11] was in fact the BL2 band.

An important feature of the BL2 band is that it bleaches during continuous UV illumination, indicating unstable behavior (note that the BL band is stable under these conditions). Simultaneously with the bleaching of the BL2 band under continuous above-bandgap illumination, the intensity of the YL band rises. However, in the samples with the Zn-related BL band, the YL band is always stable. The characteristic transfer of PL intensity from the BL2 band (referred to as the "blue band") to the YL band at low temperature under continuous UV illumination has often been reported[7-9,12-15] and is usually attributed to the metastable nature of the related defects.

Previous predictions suggested that carbon (C) plays a major role in the appearance of the BL2 band, since high-resistivity GaN was typically obtained by compensating shallow donors with C acceptors. However, the BL2 band can also be observed in high-resistivity undoped GaN and GaN doped with Fe.[79] It was proposed that the BL2 band is associated with some defect complex containing hydrogen, and the bleaching is caused by the dissociation of this complex under UV exposure.[78,79] Such an attribution can be supported by the fact that a blue band with a maximum at 3.05 eV in Ref. [86] (supposedly BL2) became much stronger after being treated by hydrogen plasma at 200°C for one hour. Since at the time, the YL band was assumed to be caused by the $V_{Ga}O_N$ complex, the BL2 band was tentatively attributed to the $V_{Ga}O_N$-H complex.[79] However, recent calculations based on hybrid functionals indicate that, depending on the sample growth procedure, the YL band can be caused by the isolated C_N defect or the C_NO_N complex.[33,91,92,93,94] Furthermore, these calculations suggest that neither the isolated V_{Ga} nor the $V_{Ga}O_N$ complex can be responsible for the YL band,[92,93,73] because the related energy levels are much deeper than what was previously suggested from density functional theory (DFT).[48,95]

Overall, due to the incessant improvement of first-principles calculations, a more contemporary theoretical analysis of the behavior of the bleaching of the BL2 band under UVL exposure might be required. In the next section, we will provide a description of various methods used for the analysis of the electronic and optical properties of defects in GaN.

Section 2. Approximations to the many-body Schrödinger equation

2.1. Foundation and Importance

The understanding of the electronic structure and optical properties of defects in semiconductors is based upon theoretical methods of statistical and quantum mechanics. If one wishes to discuss the properties of interacting defects within the bulk, it is natural to consider the time-independent Schrödinger equation for *M* electrons with *P* ions,

$$\hat{H}_{Tot}\Psi_{Tot} = E_{Tot}\Psi_{Tot}, \tag{2.1.1}$$

where E_{Tot} represents the total energy of the system and the many-body wave function $\Psi_{Tot} = \Psi(\vec{r}_1,...,\vec{r}_M,\vec{s}_1,...,\vec{s}_M;\vec{r}_1,...,\vec{r}_P,\vec{s}_1,...,\vec{s}_P)$ gives all the necessary information about the system. The position and spin of the *m*-th electron are respectively denoted by $\vec{r}_m,\vec{s}_m$ and the position and spin components of the *p*-th nuclei are represented by $\vec{r}_p,\vec{s}_p$. The nature of the electronic spin component will further discussed in section 2.2.2. The Hamiltonian of our previous equation describes the correlated motion of the electrons and nuclei in a many-body system that is represented by:

$$\hat{H}_{Tot} = -\frac{\hbar^2}{2m_e}\sum_{m=1}^{M}\nabla_m^2 - \sum_{p=1}^{P}\frac{\hbar^2}{2m_p}\nabla_p^2 + \sum_{m=1}^{M}\sum_{n>m}^{N}\frac{e^2}{\left|\vec{r}_m-\vec{r}_n\right|} - \sum_{m=1}^{M}\sum_{p=1}^{P}\frac{Z_p e^2}{\left|\vec{r}_m-\vec{r}_p\right|} + \sum_{p=1}^{P}\sum_{l>p}^{L}\frac{Z_p Z_l e^2}{\left|\vec{r}_p-\vec{r}_l\right|}, \tag{2.1.2}$$

where electrons are represented with charge $-e$ and mass m_e while the nuclei are denoted by charge $+Z_p e$ and mass m_p. In the above equation, $M = N$ and $P = L$; the use of different letters to denote the maximum number of electrons (*M* or *N*) and the maximum number of ions (*P* or *L*) is for mathematical simplicity that shall become more obvious in the derivation of Koopman's theorem in section 2.2.5. In our molecular Hamiltonian, we suppose that the motion of both the

electrons and the nuclei are treated strictly non-relativistically. Although the relativistic corrections of the kinetic energy are completely neglected, one must not forget that for heavy atoms, relativistic effects play a major role in the structure of the Hamiltonian.[96] Furthermore, the description of $\hat{H}_{Tot}$ will be restricted to a zero temperature formalism.

The terms of this quite complex molecular Hamiltonian describe

- The kinetic energy operators for the electrons $\left(\hat{T}_m\right)$ and ions $\left(\hat{T}_p\right)$:

$$\hat{T}_m = -\frac{\hbar^2}{2m_e}\sum_{m=1}^{M}\nabla_m^2 \tag{2.1.3}$$

$$\hat{T}_p = -\frac{\hbar^2}{2m_p}\sum_{p=1}^{P}\nabla_p^2 \tag{2.1.4}$$

- The potential energy due to electron-electron repulsion and the potential acting on the electrons due to the nuclei, $\hat{U}_{mp}$ and $\hat{U}_{mp}$ respectively:

$$\hat{U}_{mn} = \sum_{m=1}^{M}\sum_{n>m}^{N}\frac{e^2}{\left|\vec{r}_m - \vec{r}_n\right|} \tag{2.1.5}$$

$$\hat{U}_{mp} = \sum_{m=1}^{M}\sum_{p=1}^{P}\frac{Z_p e^2}{\left|\vec{r}_m - \vec{r}_p\right|} \tag{2.1.6}$$

- And the ion-ion nuclear repulsion potential energy, $\hat{U}_{pl}$:

$$\hat{U}_{pl} = \sum_{p=1}^{P}\sum_{l>p}^{L}\frac{Z_p Z_l e^2}{\left|\vec{r}_p - \vec{r}_L\right|} \tag{2.1.7}$$

A rather rough estimate of the computational complexity of the many-body Schrödinger equation is to visualize the fairly vast scale of our resulting Hamiltonian operator. For a typical system, the number of electrons is approximately ten times greater than the number of ions and the total amount of *P* ions is quite close to Avogadro's number, where

$P \approx N_A \approx 6.02 \times 10^{23} mol^{-1}$. Hence the total number of variables is to the order of 10^{24}. For *M* electrons and *P* ions, the many-body wave function reaches the degree of freedom of *4M+4P* and therefore the computation of the full many-body wave function remains impossible for real systems with more than few electrons. Even though analytical solutions of the Schrödinger equations can be solved for few very simple systems[97], our cases of interest involve optical properties of systems that contain tremendous amount of electrons and also thermodynamic transition levels associated with deep and shallow defects. A complete description of such systems with quantum mechanics is quite complex and thus one requires the use of an approximate, more simplified representation of our initial system.

A convenient way to reduce the scale of our Hamiltonian is to make use of the Born-Oppenheimer approximation[98], where the nuclear motion can be separated from the electronic motion. Basically in our system, since the nuclei are much heavier than the electrons $(\mathrm{m}_p \gg m_e)$, the inverse mass of the *p*-th nuclei $\frac{1}{m_p}$ becomes extremely small and hence the kinetic energy operator for ions $\hat{T}_p = -\sum_{p=1}^{P} \frac{\hbar^2}{2m_p} \nabla_p^2$ becomes the only negligible term in our many-body Hamiltonian. In the Born-Oppenheimer approximation, the electrons will organize themselves as if the ions were static and those fixed ions will not affect the states of the electrons except as a potential $\hat{U}_{mp}$. Consequently, the interaction potential between "fixed" *m*-th and *p*-th ions will become a constant classical electrostatic potential. Hence, by neglecting the kinetic energy of ions and setting their potential as a constant electrostatic potential, the many-body Hamiltonian becomes the electronic Hamiltonian $\hat{H}_{elec}$, in which the position of the nuclei are only *parameters*[7] where:

$$\hat{H}_{elec} = \hat{T}_m + \hat{U}_{mn} + \hat{U}_{mp}, \tag{2.1.8}$$

By a *parametric dependence*, we imply that the electrons can instantly adjust to any modifications in the nuclear configurations. This means that if one is interested in modifying the nuclear positions in any type of calculations, then it is necessary to add the ion-ion repulsion energy $\hat{U}_{pl}$ to the electronic Hamiltonian in order to calculate the total energy of the new structure.[99] In order to avoid the messiness of the units of $\hat{H}_{elec}$, we adopt Hartree atomic units $\hbar = e = m_e = 4\pi\varepsilon_0 = 1$, then the kinetic energy operator for electrons becomes:

$$\hat{T}_m = -\frac{1}{2}\sum_{m=1}^{M}\nabla_m^2 \tag{2.1.9}$$

The electron-electron interaction potential $\hat{U}_{mn}$ and the potential acting on the electrons due to nuclei $\hat{U}_{mp}$ are expressed as:

$$\hat{U}_{mn} = \sum_{m=1}^{M}\sum_{n>m}^{N}\frac{1}{\left|\vec{r}_m - \vec{r}_n\right|} \tag{2.1.10}$$

$$\hat{U}_{mp} = \sum_{m=1}^{M}\sum_{p=1}^{P}\frac{Z_p}{\left|\vec{r}_m - \vec{r}_p\right|} \tag{2.1.11}$$

Even though we have marginally reduced the number of variables in the general Hamiltonian, the obtained electronic Hamiltonian still achieves frightening proportions.

One of the earliest and traditional formalisms that approximates the many-body wave function Ψ_{Tot} was derived by Hartree[100] in 1928 who rewrote Ψ_{Tot} as a product of single particle functions, i.e,

$$\Psi_{Tot}(\vec{r}_1, \vec{r}_2, ..., \vec{r}_N) = \Psi(\vec{r}_1)\Psi(\vec{r}_2)....\Psi(\vec{r}_N)$$

Here we are not accounting for electron spins yet and more details regarding the Hartree approximation is given in the next section. Each one of the obtained wave functions $\Psi_m(\vec{r}_m)$ satisfies a one-electron Schrodinger equation, and the Hartree Hamiltonian yields:

$$\hat{H}_{Har} = -\frac{1}{2}\sum_{m=1}^{M}\nabla_m^2 + \sum_{m=1}^{M}\sum_{n>m}^{N}\int \frac{\left|\Psi_m(\vec{r}_m)\right|^2 \left|\Psi_n(\vec{r}_n)\right|^2}{\left|\vec{r}_m - \vec{r}_n\right|} d\vec{r}_m d\vec{r}_n - \sum_{m=1}^{M}\sum_{p=1}^{P}\frac{Z_p}{\left|\vec{r}_m - \vec{r}_p\right|} \tag{2.1.12}$$

The second term in the above equation describes the Hartree energy:

$$\hat{U}_{mn}^{Har} = \sum_{m=1}^{M}\sum_{n>m}^{N}\int \frac{\left|\Psi_m(\vec{r}_m)\right|^2 \left|\Psi_n(\vec{r}_n)\right|^2}{\left|\vec{r}_m - \vec{r}_n\right|} d\vec{r}_m d\vec{r}_n, \tag{2.1.13}$$

that arises solely from the electrostatic repulsion between electrons (Coulomb's integral). Further analysis of the Hartree formalism and the peculiar structure of its wave function are postponed until it can be thoroughly discussed in the next chapter. Although the Hartree model drastically decreases the complexity of the many-body Hamiltonian, it does violate the anti-symmetry principle of fermions (see section 2.2.1). By implementing the Pauli spin rule into the Hartree method, the Hartree-Fock (HF) approach is developed in which the wave function is approximated by a combination of anti-symmetric one-electron wave functions (section 2.2.2). Nevertheless, the number of variables from such mathematical construct remained very large.

The very first approach based on electronic density calculations in a many-body system was performed by Thomas[101] and Fermi[102] in the late twenties. In this scheme, the motion of electrons is completely uncorrelated and their corresponding kinetic energies can be described as a "functional" of a local density based on free electrons density in a homogeneous electron gas.[97] However, in the original Thomas-Fermi (TF) method, the exchange and correlation energies among electrons were completely neglected. In 1930, the local exchange effects[103] were proposed by Dirac and became incorporated into the TF original formalism. In the mid-fifties,

Slater[104] proposed a simplification of the HF potential in the $X\alpha$ scheme by setting up an average local potential field based on the uniform electron gas.[97] Countless improvements of the HF method have been made in the last decades and have been essential to the development to modern density functional theory (DFT). In fact, traditional HF wave functions can be used to compute fairly precise results for smaller systems, *providing benchmarks for further developing density functionals, which can hence allow precise electronic calculations of larger systems*.[105] However, one must wait until the mid-sixties to finally obtain a formalism that does not start with too crude approximations[106], and yet provides certain equilibrium between accuracy and computational cost.

The first thorough and complete proof of the existence of DFT was given by Hohenberg and Kohn in 1964.[107] These authors demonstrated that the ground state electron density of a system contains all the information included within its ground state many-electrons wave function.[107] In other words, all characteristics of the systems can be considered as *functionals* of the ground state density.7 Furthermore, they argue that, for any given external potential $\hat{U}_{ext}$ (which corresponds to the potential acting on the electrons due to nuclei or $\hat{U}_{mp}$) if a "universal functional" for the total energy of our system were to be known, one would be able to obtain the exact ground state energy of multi-electron system by minimizing the total energy of the system, with respect to the ground state density.[7] Such method would allow much larger systems to be solved by ab-initio methods, while retaining much of their precision. Nonetheless, precision is a quite relative term. Even though theoretically, DFT is an exact theory, its actual performance relies on the quality of the approximate density functionals employed.

2.2. The Hartree-Fock (HF) method

2.2.1. Overview of the Hartree approximation

One of the very first approaches to the multi-electrons problem was proposed by Hartree (1928)[100], in which he assumed that each electron is subjected to a field arising from the averaged charge density of each other electron. A very thorough and intuitive description of the Hartree method is given by Slater in his first volume of *Quantum Theory of Atomic Structure* (pps. 189, 213). In Slater's book, it is argued that the sum of the charge densities of the remaining electrons surrounding the nucleus is approximately spherically symmetric. As a result, the potential arising from the average charge density of electrons and the nucleus is also spherically symmetric. In other words, in the Hartree approach, every electron moves in an averaged spherical potential field arising from the nucleus and the remaining electrons in the system. Such approach is called the central-field approximation and is described in any advanced electromagnetism and quantum mechanics book. In order to solve the Schrödinger equation of an electron moving in an averaged spherical potential, we first compute the total charge density created from the other electrons surrounding the nucleus. Then, we calculate the potential arising from the total charge density and we finally ensure that the obtained potential matches the initial one that was used in the calculation (self-consistent procedure). However, in the case of electrons subjected to a self-consistent field, the motion of the electrons must be solved quantum mechanically which therefore over-complicates the solution of the problem. Hartree circumvented this problem by using the method of iterations in which:

- A trial wave function hopefully closed to the final one is initially used.
- From the trial wave function, the charge density is defined.

- Insert the charge densities into the Hamiltonian and solve for the potential.
- Use the obtained potential in the Schrödinger equation and calculate new wave functions.
- Start the process again until convergence is obtained.

Before implementing the SCF approach via the iteration method, one needs to inquire about the construction of a wave function of an electron that is subjected to a central potential field. Here, each electron interacts with one another only in an average manner. Therefore the probability density of an electron must be a product of the probability distributions of the remaining electrons. As a result, a trial wave function can be written as a product of one-electron orbitals of the form:

$$\Psi_{Trial}(\vec{r}_1,\vec{r}_2,...,\vec{r}_M)=\prod_{m=1}^{M}\Psi(\vec{r}_m)=\Psi(\vec{r}_1)\Psi(\vec{r}_2)\ldots\Psi(\vec{r}_M) \qquad (2.2.1.1)$$

However, electrons are indistinguishable spin particles (fermions), and by exchanging two electrons, the wave function must correspondingly switch sign according to the anti-symmetry principle. The simple product form in the Hartree wave function is in fact unacceptable since interchanging the indices of electrons does not yield the negative of the wave function. Even though the Hartree approach is not quite realistic enough for multi-electrons systems, it is briefly included in this section just to describe the basic foundations of one-electron approaches.

2.2.2. Slater Determinant

In 1930, Fermi-Dirac statistics[108,109] became incorporated into the Hartree formalism by Fock and Slater[104], in which the total wave function of an *M*-electron system is approximated by an anti-symmetrized product of *I* orthonormal spin orbitals $\psi_i(\vec{\tau}_m)$[110]. The newly derived formalism became known as the HF method and the spin orbitals $\psi_i(\vec{\tau}_m)$ represent one-electron wave functions that result from the product of a spatial orbital $\varphi(\vec{r}_m)$ with a spin part $|m_s\rangle = \begin{cases} |\alpha(s_m)\rangle \\ |\beta(s_m)\rangle \end{cases}$. Here α and β are referred as the "up-spin" and "down-spin", respectively. By incorporating a Slater determinant[111] into the Hartree approximation, Fock proposed an antisymmetric *M*-electron wave function in the form of:

$$\Psi_{HF}(\vec{\tau}_1,\vec{\tau}_2,...,\vec{\tau}_M) = \frac{1}{\sqrt{M!}}\begin{vmatrix} \psi_1(\vec{\tau}_1) & \psi_2(\vec{\tau}_1) & \cdots & \psi_I(\vec{\tau}_1) \\ \psi_1(\vec{\tau}_2) & \psi_2(\vec{\tau}_2) & \cdots & \psi_I(\vec{\tau}_2) \\ \vdots & \vdots & & \vdots \\ \psi_1(\vec{\tau}_M) & \psi_2(\vec{\tau}_I) & \cdots & \psi_I(\vec{\tau}_M) \end{vmatrix} \quad (2.2.2.1)$$

Here $\vec{\tau}_m$ indicates both spatial $(\vec{r}_m)$ and spin coordinates of the *m*-th electron. The coefficient in front of the matrix represents a normalization factor and within the matrix, we notice that *M* electrons can occupy *I* orbitals without specifying which electron is in which spin orbital. For any arbitrary number of electrons, the Slater determinant expression can be shown to satisfy the conditions of the Pauli Exclusion Principle. Although we have satisfied the anti-symmetry principle with a Slater determinant, nothing has been said so far about the form of the spin orbitals. In order to obtain the best set of wave functions that minimize the total energy of the electronic Hamiltonian, one can subject the Slater determinant to the variational principle. In the

following section, we describe the variational principle which plays a major role in the Hartree-Fock formalism.

2.2.3. The Variational Principle

We initially start by remembering that any wave function in Hilbert space can be written as an infinite linear combination of basis functions $\left(\Psi_j\right)$:

$$\Psi_{HF}(\vec{\tau}_1,\vec{\tau}_2,...,\vec{\tau}_M)=\sum_{j=1}^{\infty}c_j\Psi_j, \tag{2.2.3.1}$$

where c_j are the expansion coefficients.

The expectation value of the total energy $\langle E_{HF}\rangle$ associated with the above HF wave function is given by:

$$\langle E_{HF}\rangle=\frac{\langle\Psi_{HF}|\hat{H}_{elec}|\Psi_{HF}\rangle}{\langle\Psi_{HF}|\Psi_{HF}\rangle} \tag{2.2.3.2}$$

By substituting Eq. 2.2.3.1 into the above equation, we obtain:

$$\langle E_{HF}\rangle=\frac{\left\langle\sum_j^{\infty}c_j\Psi_j\left|\hat{H}_{elec}\right|\sum_j^{\infty}c_j\Psi_j\right\rangle}{\left\langle\sum_j^{\infty}c_j\Psi_j\left|\sum_j^{\infty}c_j\Psi_j\right.\right\rangle}=\frac{\sum_j^{\infty}c_j^{*}c_j\langle\Psi_j|\hat{H}_{elec}|\Psi_j\rangle}{\sum_j^{\infty}c_j^{*}c_j} \tag{2.2.3.3}$$

$$\Leftrightarrow\langle E_{HF}\rangle=\frac{\sum_j^{\infty}|c_j|^2\langle\Psi_j^{*}|\hat{H}_{elec}|\Psi_j\rangle}{\sum_j^{\infty}|c_j|^2}=\sum_j^{\infty}|c_j|^2\langle\Psi_j^{*}|\hat{H}_{elec}|\Psi_j\rangle=\sum_j^{\infty}|c_j|^2\langle E_j\rangle \tag{2.2.3.4}$$

where $\langle E_j\rangle$ is the expectation value of the *j*-th eigenstate of the electronic Hamiltonian for a properly normalized HF wave function.

By using the same procedure described in Ref. [112], we expand the expression of the average HF energy $\langle E_{HF} \rangle$ by explicitly including its $j = 0$ term (ground state energy) in the summation:

$$\langle E_{HF} \rangle = |c_0|^2 \langle E_0 \rangle + \sum_{j=1}^{\infty} |c_j|^2 \langle E_j \rangle \tag{2.2.3.5}$$

Here the expansion coefficient $\left(|c_0|^2\right)$ corresponding to the true ground-state $\langle E_0 \rangle$ can be obtained from the definition of the normalization condition:

$$\sum_{j}^{\infty} |c_j|^2 = 1 \Leftrightarrow |c_0|^2 + \sum_{j=1}^{\infty} |c_j|^2 = 1$$
$$\Leftrightarrow |c_0|^2 = 1 - \sum_{j=1}^{\infty} |c_j|^2$$

By substituting the above equation in Eq. 2.2.3.5, we obtain:

$$\langle E_{HF} \rangle = \left[1 - \sum_{j=1}^{\infty} |c_j|^2 \right] \langle E_0 \rangle + \sum_{j=1}^{\infty} |c_j|^2 \langle E_j \rangle$$
$$\Leftrightarrow \langle E_{HF} \rangle = \langle E_0 \rangle - \sum_{j=1}^{\infty} |c_j|^2 \langle E_0 \rangle + \sum_{j=1}^{\infty} |c_j|^2 \langle E_j \rangle$$
$$\Leftrightarrow \langle E_{HF} \rangle - \langle E_0 \rangle = \sum_{j=1}^{\infty} |c_j|^2 \left[\langle E_j \rangle - \langle E_0 \rangle \right]$$

$$\Leftrightarrow \langle E_{HF} \rangle - \langle E_0 \rangle \geq 0, \tag{2.2.3.6}$$

since $\sum_{j=1}^{\infty} |c_j|^2 \geq 0$ and we are subtracting the lowest value of the sum of the E_j terms from $\sum_{j=0}^{\infty} |c_j|^2 \langle E_j \rangle$. Therefore Eq. 2.2.3.6 can be re-written as:

$$\langle E_{HF} \rangle \geq \langle E_0 \rangle \tag{2.2.3.7}$$

This means that regardless of the shape of the trial wave function, the expectation value of the Hamiltonian operator subjected to the variational principle is always greater than, or equal to the

true ground state of the many-body system. Such result is quite encouraging since it appears that any variations in the trial wave function which minimize its corresponding energy $\langle E_j \rangle$ are automatically bringing the HF energy $\langle E_{HF} \rangle$ closer to the exact ground state energy $\langle E_0 \rangle$. However, the practicality of the variational principle depends strongly on the ability to make an acceptable initial guess of the unknown wave function, the symmetry of the system and various other physical properties.[113] Now that we have explained the variational principle, let us apply it to the HF Hamiltonian and analyze the components of the resultant expectation value.

2.2.4. The Hartree-Fock Hamiltonian

After lengthy mathematical derivations that can be found in most advanced quantum chemistry books, we define the expectation value of the HF energy as:

$$\begin{aligned}\left\langle E_{HF}\right\rangle = E_{HF} &= \left\langle \Psi_{HF}(\vec{\tau}_1,\vec{\tau}_2,...,\vec{\tau}_N)\middle|\hat{H}_{elec}\middle|\Psi_{HF}(\vec{\tau}_1,\vec{\tau}_2,...,\vec{\tau}_N)\right\rangle \\ \Leftrightarrow E_{HF} &= \sum_{i=1}^{I}\underbrace{\left\langle \psi_i(\vec{\tau}_m)\middle|\hat{h}_m\middle|\psi_i(\vec{\tau}_m)\right\rangle}_{H_{ii}} \\ &\quad + \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J}\underbrace{\left[\left\langle \psi_i(\vec{\tau}_n)\psi_j(\vec{\tau}_n)\middle|\frac{1}{r_{mn}}\middle|\psi_i(\vec{\tau}_m)\psi_j(\vec{\tau}_m)\right\rangle\right]}_{J_{ij}} \\ &\quad - \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J}\underbrace{\left\langle \psi_i(\vec{\tau}_m)\psi_j(\vec{\tau}_n)\middle|\frac{1}{r_{mn}}\middle|\psi_j(\vec{\tau}_m)\psi_i(\vec{\tau}_n)\right\rangle}_{K_{ij}},\end{aligned} \tag{2.2.4.1}$$

The maximum number of orbitals *I* is equal to *J* and $r_{mn} = |\vec{r}_m - \vec{r}_n|$ denotes the distance between the *m*-th and *n*-th electron. Here, $\hat{h}_m = -\frac{\nabla_m^{\,2}}{2} - \sum_{K=1}\frac{Z_K}{|\vec{r}_m - \vec{R}_K|}$ is the one-electron Hamiltonian that corresponds to the energy of a single electron moving in the average field of the nuclei alone. The J_{ij} term or the Coulomb integral: $J_{ij} = \left\langle \psi_i(\vec{\tau}_n)\psi_j(\vec{\tau}_n)\middle|\frac{1}{r_{mn}}\middle|\psi_i(\vec{\tau}_m)\psi_j(\vec{\tau}_m)\right\rangle$, (2.2.4.2)

describes the average electrostatic repulsion energy between electrons occupying the *i*-th spin orbitals $\psi_i(\vec{\tau}_n)$ and the *j*-th spin orbital $\psi_j(\vec{\tau}_m)$. The last component K_{ij}, called the exchange integral, is purely quantum mechanical and prevents electrons with identical spins to occupy the same orbital. Furthermore, for $i = j$, the K_{ij} term cancels out the J_{ij} term which consequently eliminates the spurious self-interaction term that appears in the approximations to the DFT (see sections 2.6.3-2.6.4).

From Eq. 2.2.4.1, it appears that the HF energy values is dependent on the form of the spin orbital $\psi(\vec{\tau})$ $\left(\text{or } \psi^*(\vec{\tau})\right)$ instead of the variable τ which hence makes $E_{HF}\left[\psi(\vec{\tau})\right]$ $\left(\text{or } E_{HF}\left[\psi^*(\vec{\tau})\right]\right)$ a functional of $\psi(\vec{\tau})$ or $\left(\psi^*(\vec{\tau})\right)$. By allowing small arbitrary variations in the spin orbital $\left(\psi^*(\vec{\tau}) \to \psi^*(\vec{\tau}) + \delta\psi^*(\vec{\tau})\right)$, the HF energy functional $E_{HF}\left[\psi^*(\vec{\tau})\right]$ can be written as:

$$
\begin{aligned}
E_{HF}\left[\psi^*(\vec{\tau}) + \delta\psi^*(\vec{\tau})\right] &= \left\langle \psi(\vec{\tau}) + \delta\psi(\vec{\tau}) \middle| \hat{H}_{elec} \middle| \psi(\vec{\tau}) \right\rangle \\
&= \underbrace{\left\langle \psi(\vec{\tau}) \middle| \hat{H}_{elec} \middle| \psi(\vec{\tau}) \right\rangle}_{E_{HF}\left[\psi^*(\vec{\tau})\right]} + \left\langle \delta\psi(\vec{\tau}) \middle| \hat{H}_{elec} \middle| \psi(\vec{\tau}) \right\rangle \\
\Leftrightarrow E_{HF}\left[\psi^*(\vec{\tau}) + \delta\psi^*(\vec{\tau})\right] &= E_{HF}\left[\psi^*(\vec{\tau})\right] + \left\langle \delta\psi(\vec{\tau}) \middle| \hat{H}_{elec} \middle| \psi(\vec{\tau}) \right\rangle \\
\Leftrightarrow E_{HF}\left[\psi^*(\vec{\tau}) + \delta\psi^*(\vec{\tau})\right] - E_{HF}\left[\psi^*(\vec{\tau})\right] &= \left\langle \delta\psi(\vec{\tau}) \middle| \hat{H}_{elec} \middle| \psi(\vec{\tau}) \right\rangle \\
\Leftrightarrow \delta E_{HF}\left[\psi^*(\vec{\tau})\right] &= \left\langle \delta\psi(\vec{\tau}) \middle| \hat{H}_{elec} \middle| \psi(\vec{\tau}) \right\rangle \qquad (2.2.4.3)
\end{aligned}
$$

where $\delta E_{HF}\left[\psi^*(\vec{\tau})\right]$ is the first linear variation of the functional $E_{HF}\left[\psi^*(\vec{\tau})\right]$. Here we have chosen small arbitrary variations on $\psi^*(\vec{\tau})$ instead of $\psi^*(\vec{\tau})$ for mathematical simplicity. Had we decided to select variations on $\psi^*(\vec{\tau})$ instead of $\psi^*(\vec{\tau})$, the end results would be identical.

By setting $\delta E_{HF}\left[\psi^*(\vec{\tau})\right] = 0$, for small arbitrary variations, we are seeking the set of spin orbitals $\psi^*(\vec{\tau})$ that yields a minimum energy. However, any arbitrary variations of the form $\psi^*(\vec{\tau}) \to \psi^*(\vec{\tau}) + \delta\psi^*(\vec{\tau})$ do not necessarily ensure the orthonormality of the spin orbitals and possibly violate the use of HF energy functional which is a function of orthonormal spin orbitals.

Therefore, one must constrain the first linear variation of the energy functional $\left(\delta E_{HF}\left[\psi\left(\vec{\tau}\right)\right]\right)$ to the orthonormality conditions or in other words:

$$\delta E_{HF}\left[\psi\left(\vec{\tau}\right)\right]=0 \text{ is subjected to } \Omega_{ij}=\left\langle\psi_i\left(\vec{\tau}_m\right)\middle|\psi_j\left(\vec{\tau}_m\right)\right\rangle=\delta_{ij} \tag{2.2.4.4}$$

Whenever a minimum is sought and subjected to a specific constraint, the method of Lagrange multipliers is generally used in which a linear combination of the variations of the functionals E_{HF} and Ω_{ij} is simultaneously equal to zero. In our case, we obtain:

$$c_E\cdot\delta E_{HF}\left[\psi\left(\vec{\tau}\right)\right]+\delta\sum_{i=1}^{I}\sum_{j=1}^{J}c_{ij}\Omega_{ij}=0$$

$$\Leftrightarrow c_E\cdot\delta E_{HF}\left[\psi\left(\vec{\tau}\right)\right]+\delta\sum_{i=1}^{I}\sum_{j=1}^{J}c_{ij}\left\langle\psi_i\left(\vec{\tau}_m\right)\middle|\psi_j\left(\vec{\tau}_m\right)\right\rangle=0, \tag{2.2.4.5}$$

where c_E and c_{ij} are arbitrary multipliers attached to the $\delta E_{HF}\left[\psi\left(\vec{\tau}\right)\right]$ and Ω_{ij}, respectively. Here, we are enforcing the arbitrarity of the variations of the linear combinations of both E_{HF} and Ω_{ij} functionals, and at the same time automatically incorporating the orthonormality conditions.

For mathematical simplicity, we divide the above equation by c_E which yields:

$$\frac{1}{c_E}\left\{c_E\cdot\delta E_{HF}\left[\psi^*\left(\vec{\tau}\right)\right]+\delta\sum_{i=1}^{I}\sum_{j=1}^{J}c_{ij}\left\langle\psi_i\left(\vec{\tau}_m\right)\middle|\psi_j\left(\vec{\tau}_m\right)\right\rangle\right\}=0$$

$$\Leftrightarrow\delta E_{HF}\left[\psi^*\left(\vec{\tau}\right)\right]+\delta\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{c_{ij}}{c_E}\left\langle\psi_i\left(\vec{\tau}_m\right)\middle|\psi_j\left(\vec{\tau}_m\right)\right\rangle=0$$

$$\Rightarrow\delta E_{HF}\left[\psi^*\left(\vec{\tau}\right)\right]-\delta\sum_{i=1}^{I}\sum_{j=1}^{J}\varepsilon_{ij}\left\langle\psi_i\left(\vec{\tau}_m\right)\middle|\psi_j\left(\vec{\tau}_m\right)\right\rangle=0, \tag{2.2.4.6}$$

where $\varepsilon_{ij}=-\frac{c_{ij}}{c_E}$ are the Lagrange multipliers. More details on the derivation of Lagrange multipliers are described by Szabo and Ostlund[114].

By subjecting the expression for the HF functional energy (Eq. 2.2.4.1) to linear variation of first order, we obtain:

$$\delta E_{HF} = \delta\sum_{i=1}^{I} H_{ii} + \frac{1}{2}\delta\sum_{i=1}^{I}\sum_{j=1}^{J} J_{ij} - \frac{1}{2}\delta\sum_{i=1}^{I}\sum_{j=1}^{J} K_{ij} \tag{2.2.4.7}$$

The first term of the above equation can be written as:

$$\begin{aligned}\delta\sum_{i=1}^{I} H_{ii} &= \delta\sum_{i=1}^{I}\left\langle \psi_i(\vec{\tau}_m)\middle|\hat{h}_m\middle|\psi_i(\vec{\tau}_m)\right\rangle \\ &= \left\langle \psi_k(\vec{\tau}_m) + \delta\psi_k(\vec{\tau}_m)\middle|\hat{h}_k\middle|\psi_k(\vec{\tau}_m)\right\rangle - \left\langle \psi_k(\vec{\tau}_m)\middle|\hat{h}_k\middle|\psi_k(\vec{\tau}_m)\right\rangle \\ &= \left\langle \psi_k(\vec{\tau}_m)\middle|\hat{h}_k\middle|\psi_k(\vec{\tau}_m)\right\rangle + \left\langle \delta\psi_k(\vec{\tau}_m)\middle|\hat{h}_k\middle|\psi_k(\vec{\tau}_m)\right\rangle - \left\langle \psi_k(\vec{\tau}_m)\middle|\hat{h}_k\middle|\psi_k(\vec{\tau}_m)\right\rangle\end{aligned}$$

$$\Leftrightarrow \delta\sum_{i=1}^{I} H_{ii} = \left\langle \delta\psi_k^*(\vec{\tau}_m)\middle|\hat{h}_k\middle|\psi_k(\vec{\tau}_m)\right\rangle \tag{2.2.4.8}$$

Regarding the disappearance of the sum over the *i* terms, we have dropped all the *i* terms except the one for which its functional derivative yields a non-zero value (denoted with the subscript *k*). For instance, if one is interested in calculating a specific derivative of the summation $\sum_{i=1}^{I} x_i$, the expression can be written as:

$$\begin{aligned}\frac{d}{dx_3}\sum_{i=1}^{I} x_i &= \cancelto{0}{\frac{d}{dx_3}x_1} + \cancelto{0}{\frac{d}{dx_3}x_2} + \frac{d}{dx_3}x_3 + \cdots \cancelto{0}{\frac{d}{dx_3}x_I} \\ &\Rightarrow \sum_{i=1}^{I}\frac{d}{dx_3}x_i = 1\end{aligned}$$

By subjecting the Coulombic term from the HF energy functional to a first order linear variation, we obtain:

$$\delta\sum_{j=1}^{J}\sum_{i=1}^{I}J_{ij}=\delta\sum_{i=1}^{I}\sum_{j=1}^{J}\left\langle\psi_i(\vec{\tau}_m)\psi_j(\vec{\tau}_n)\left|\frac{1}{r_{mn}}\right|\psi_i(\vec{\tau}_m)\psi_j(\vec{\tau}_n)\right\rangle$$
$$=\sum_{j=1}^{J}\left\{\left\langle\left[\psi_k(\vec{\tau}_m)+\delta\psi_k(\vec{\tau}_m)\right]\psi_j(\vec{\tau}_n)\left|\frac{1}{r_{mn}}\right|\psi_k(\vec{\tau}_m)\psi_j(\vec{\tau}_n)\right\rangle\right.$$
$$\left.-\left\langle\psi_k(\vec{\tau}_m)\psi_j(\vec{\tau}_n)\left|\frac{1}{r_{mn}}\right|\psi_k(\vec{\tau}_m)\psi_j(\vec{\tau}_n)\right\rangle\right\}$$
$$+\sum_{i=1}^{I}\left\{\left\langle\psi_i(\vec{\tau}_m)\left[\psi_k(\vec{\tau}_n)+\delta\psi_k(\vec{\tau}_n)\right]\left|\frac{1}{r_{mn}}\right|\psi_i(\vec{\tau}_m)\psi_k(\vec{\tau}_n)\right\rangle\right.$$
$$\left.-\left\langle\psi_i(\vec{\tau}_m)\psi_k(\vec{\tau}_n)\left|\frac{1}{r_{mn}}\right|\psi_i(\vec{\tau}_m)\psi_k(\vec{\tau}_n)\right\rangle\right\}$$

After cancelation of some of the terms included in the curly bracket, the above equation is simplified as:

$$\delta\sum_{j=1}^{J}\sum_{i=1}^{I}J_{ij}=\sum_{j=1}^{I}\left\langle\delta\psi_k(\vec{\tau}_m)\psi_j(\vec{\tau}_n)\left|\frac{1}{r_{mn}}\right|\psi_k(\vec{\tau}_m)\psi_j(\vec{\tau}_n)\right\rangle+\sum_{i=1}^{J}\left\langle\psi_i(\vec{\tau}_m)\delta\psi_k(\vec{\tau}_n)\left|\frac{1}{r_{mn}}\right|\psi_i(\vec{\tau}_m)\psi_k(\vec{\tau}_n)\right\rangle$$

Since *i* and *j* are dummy variables, the above equation can be re-written as:

$$\delta\sum_{j=1}^{J}\sum_{i=1}^{I}J_{ij}=\sum_{j=1}^{J}\left\langle\delta\psi_k(\vec{\tau}_m)\psi_j(\vec{\tau}_n)\left|\frac{1}{r_{mn}}\right|\psi_k(\vec{\tau}_m)\psi_j(\vec{\tau}_n)\right\rangle+\sum_{j=1}^{J}\left\langle\psi_j(\vec{\tau}_m)\delta\psi_k(\vec{\tau}_n)\left|\frac{1}{r_{mn}}\right|\psi_j(\vec{\tau}_m)\psi_k(\vec{\tau}_n)\right\rangle \quad (2.2.4.9)$$

Due to the hermiticity of the $\frac{1}{r_{mn}}$ operator, the second term of the above equation becomes:

$$\sum_{j}^{J}\left\langle\psi_j(\vec{\tau}_m)\delta\psi_k(\vec{\tau}_n)\left|\frac{1}{r_{mn}}\right|\psi_j(\vec{\tau}_m)\psi_k(\vec{\tau}_n)\right\rangle=\sum_{j}^{J}\left\langle\delta\psi_k(\vec{\tau}_n)\psi_j(\vec{\tau}_m)\left|\frac{1}{r_{mn}}\right|\psi_k(\vec{\tau}_n)\psi_j(\vec{\tau}_m)\right\rangle$$

By substituting the above equation into Eq. 2.2.4.9, we obtain:

$$\delta\sum_{i=1}^{I}\sum_{j=1}^{J}J_{ij}=2\sum_{j=1}^{J}\left\langle\delta\psi_k(\vec{\tau}_m)\psi_j(\vec{\tau}_n)\left|\frac{1}{r_{mn}}\right|\psi_k(\vec{\tau}_m)\psi_j(\vec{\tau}_n)\right\rangle \quad (2.2.4.10)$$

By applying the same first linear variation procedure to the exchange term, we have:

$$\delta\sum_{i=1}^{I}\sum_{j=1}^{J}K_{ij}=2\sum_{j=1}^{J}\left\langle\delta\psi_k\left(\vec{\tau}_m\right)\psi_j\left(\vec{\tau}_n\right)\left|\frac{1}{r_{mn}}\right|\psi_j\left(\vec{\tau}_m\right)\psi_k\left(\vec{\tau}_n\right)\right\rangle \tag{2.2.4.11}$$

Now that we have obtained the expression of each component of $\delta E\left[\psi^*\left(\vec{\tau}\right)\right]$, we also subject the orthonormality functional $\Omega_{ij}\left[\psi^*\left(\vec{\tau}_m\right)\right]$ to a first order linear variation where:

$$\delta\Omega_{ij}\left[\psi^*\left(\vec{\tau}_m\right)\right]=\delta\sum_{i=1}^{I}\sum_{j=1}^{J}\varepsilon_{ij}\left\langle\psi_i\left(\vec{\tau}_m\right)\middle|\psi_j\left(\vec{\tau}_m\right)\right\rangle$$

$$\Leftrightarrow\delta\Omega_{ij}\left[\psi^*\left(\vec{\tau}_m\right)\right]=\sum_{j=1}^{J}\varepsilon_{kj}\left[\left\langle\psi_k\left(\vec{\tau}_m\right)+\delta\psi_k\left(\vec{\tau}_m\right)\middle|\psi_j\left(\vec{\tau}_m\right)\right\rangle-\left\langle\psi_k\left(\vec{\tau}_m\right)\middle|\psi_j\left(\vec{\tau}_m\right)\right\rangle\right]$$

$$\Leftrightarrow\delta\Omega_{ij}\left[\psi^*\left(\vec{\tau}_m\right)\right]=\sum_{j=1}^{J}\varepsilon_{kj}\left\langle\delta\psi_k\left(\vec{\tau}_m\right)\middle|\psi_j\left(\vec{\tau}_m\right)\right\rangle \tag{2.2.4.12}$$

By inserting Eqs. 2.2.4.8, 2.2.4.10, 2.2.4.11 and 2.2.4.12 into Eq. 2.2.4.6, we obtain:

$$\left\langle\delta\psi_k\left(\vec{\tau}_m\right)\middle|\hat{h}_k\middle|\psi_k\left(\vec{\tau}_m\right)\right\rangle+\sum_{j=1}^{J}\left\langle\delta\psi_k\left(\vec{\tau}_m\right)\psi_j\left(\vec{\tau}_n\right)\left|\frac{1}{r_{mn}}\right|\psi_k\left(\vec{\tau}_m\right)\psi_j\left(\vec{\tau}_n\right)\right\rangle+$$

$$-\sum_{j=1}^{J}\left\langle\delta\psi_k\left(\vec{\tau}_m\right)\psi_j\left(\vec{\tau}_n\right)\left|\frac{1}{r_{mn}}\right|\psi_j\left(\vec{\tau}_m\right)\psi_k\left(\vec{\tau}_n\right)\right\rangle-\sum_{j=1}^{J}\varepsilon_{kj}\left[\left\langle\delta\psi_k\left(\vec{\tau}_m\right)\middle|\psi_j\left(\vec{\tau}_m\right)\right\rangle\right]=0$$

$$\Rightarrow\int d\vec{\tau}_m\delta\psi_k^*\left(\vec{\tau}_m\right)\left[\hat{h}_k\psi_k\left(\vec{\tau}_m\right)+\sum_{j=1}^{J}\int d\vec{\tau}_n\psi_j^*\left(\vec{\tau}_n\right)\frac{1}{r_{mn}}\psi_k\left(\vec{\tau}_m\right)\psi_j\left(\vec{\tau}_n\right)+\right.$$

$$\left.-\sum_{j=1}^{J}\int d\vec{\tau}_n\psi_j^*\left(\vec{\tau}_n\right)\frac{1}{r_{mn}}\psi_j\left(\vec{\tau}_m\right)\psi_k\left(\vec{\tau}_n\right)-\sum_{j=1}^{J}\varepsilon_{kj}\psi_j\left(\vec{\tau}_m\right)\right]=0$$

$$\Rightarrow\int d\vec{\tau}_m\delta\psi_k^*\left(\vec{\tau}_m\right)\left[\hat{h}_k\psi_k\left(\vec{\tau}_m\right)+\hat{J}\psi_k\left(\vec{\tau}_m\right)-\hat{K}\psi_k\left(\vec{\tau}_m\right)-\sum_{j=1}^{J}\varepsilon_{kj}\psi_j\left(\vec{\tau}_m\right)\right]=0 \tag{2.2.4.13}$$

where $\hat{J}\psi_k\left(\vec{\tau}_m\right)=\sum_{j=1}^{J}\int d\vec{\tau}_n\psi_j^*\left(\vec{\tau}_n\right)\frac{1}{r_{mn}}\psi_k\left(\vec{\tau}_m\right)\psi_j\left(\vec{\tau}_n\right)$ is the Coulomb sum, (2.2.4.13.a)

and $\hat{K}\psi_k\left(\vec{\tau}_m\right)=\sum_{j=1}^{J}\int d\vec{\tau}_n\psi_j^*\left(\vec{\tau}_n\right)\frac{1}{r_{mn}}\psi_j\left(\vec{\tau}_m\right)\psi_k\left(\vec{\tau}_n\right)$ is the exchange sum. (2.2.4.13.b)

The Coulomb sum can be written as a product of the Coulomb operator acting upon the $\psi_k(\vec{\tau}_m)$ spin orbital:

$$\hat{J}\left|\psi_k(\vec{\tau}_m)\right\rangle = \left[\sum_{j=1}^{J}\int d\vec{\tau}_n \psi_j^*(\vec{\tau}_n)\frac{1}{r_{mn}}\psi_j(\vec{\tau}_n)\right]\left|\psi_k(\vec{\tau}_m)\right\rangle \tag{2.2.4.14}$$

Here the Coulomb operator $\hat{J}$ describes the electrostatic potential felt by an electron located at position $\vec{\tau}_m$ from the average electronic charge distribution of the *j*-th electron at $\vec{\tau}_n$.

The exchange sum $\hat{K}\psi_k(\vec{\tau}_m)$ involves a lot more complexity when compared to the Coulombic term as one cannot factor out the $\psi_j(\vec{\tau}_m)$ term out of the summation. Since the $\psi_k(\vec{\tau}_m)$ spin orbital does not appear in the right hand side of the exchange sum equation, the $\hat{K}$ operator must involve some type of permutation operation that interchanges the indices of the two spin orbitals $\psi_k(\vec{\tau}_n)$ and $\psi_i(\vec{\tau}_m)$.[115] Let us define a permutation operator $\hat{P}_{jk}$ that acts upon the ket-vector $\left|\frac{1}{r_{mn}}\psi_k(\vec{\tau}_n)\psi_j(\vec{\tau}_m)\right\rangle$ in such way that:

$$\hat{P}_{jk}\left|\frac{1}{r_{mn}}\psi_j(\vec{\tau}_n)\psi_k(\vec{\tau}_m)\right\rangle = \left|\frac{1}{r_{mn}}\psi_k(\vec{\tau}_n)\psi_j(\vec{\tau}_m)\right\rangle \tag{2.2.4.15}$$

Let us multiply both sides of the above equation with the bra-vector $\left\langle\psi_j(\vec{\tau}_n)\right|$:

$$\left\langle\psi_j(\vec{\tau}_n)\right|\hat{P}_{jk}\left|\frac{1}{r_{mn}}\psi_j(\vec{\tau}_n)\psi_k(\vec{\tau}_m)\right\rangle = \left\langle\psi_j(\vec{\tau}_n)\right|\frac{1}{r_{mn}}\psi_k(\vec{\tau}_n)\psi_j(\vec{\tau}_m)\Big\rangle$$

Then we sum each side of the previous equation over the *j* terms:

$$\sum_{j=1}^{J}\left\langle \psi_j(\vec{\tau}_n)\left|\hat{P}_{jk}\right|\frac{1}{r_{mn}}\psi_j(\vec{\tau}_n)\psi_k(\vec{\tau}_m)\right\rangle = \sum_{j=1}^{J}\left\langle \psi_j(\vec{\tau}_n)\left|\frac{1}{r_{mn}}\psi_k(\vec{\tau}_n)\psi_j(\vec{\tau}_m)\right\rangle\right.$$

$$= \sum_{j=1}^{J}\left[\left\langle \psi_j(\vec{\tau}_n)\left|\frac{1}{r_{mn}}\psi_k(\vec{\tau}_n)\right\rangle\cdot\psi_j(\vec{\tau}_m)\right]\right.$$

$$= \sum_{j=1}^{J}\int d\vec{\tau}_n\psi_j^*(\vec{\tau}_n)\frac{1}{r_{mn}}\psi_k(\vec{\tau}_n)\psi_j(\vec{\tau}_m)$$

$$= \hat{K}\psi_k(\vec{\tau}_m)$$

Although we have proven that the exchange operator involves some permutation function, we have not been able to completely isolate the $\hat{K}$ operator and derive its corresponding expression. Following the derivations of the exchange operator from Kryachko and Ludena[116] and David B. Cook,[115] we obtain an explicit expression for the exchange operator by multiplying the right-hand side of Eq. 2.2.4.13.b with a unity expression of the form $\frac{\psi_k(\vec{\tau}_m)}{\psi_k(\vec{\tau}_m)}=1$, which yields:

$$\hat{K}\psi_k(\vec{\tau}_m) = \left[\sum_{j=1}^{J}\int d\vec{\tau}_n\psi_j^*(\vec{\tau}_n)\frac{1}{r_{mn}}\psi_j(\vec{\tau}_m)\psi_k(\vec{\tau}_n)\right]\frac{\psi_k(\vec{\tau}_m)}{\psi_k(\vec{\tau}_m)}$$

$$= \left[\sum_{j=1}^{J}\int d\vec{\tau}_n\psi_j^*(\vec{\tau}_n)\frac{1}{r_{mn}}\psi_j(\vec{\tau}_m)\frac{\psi_k(\vec{\tau}_n)}{\psi_k(\vec{\tau}_m)}\right]\psi_k(\vec{\tau}_m)$$

$$\Rightarrow \hat{K} = \sum_{j=1}^{J}\int d\vec{\tau}_n\psi_j^*(\vec{\tau}_n)\frac{1}{r_{mn}}\psi_j(\vec{\tau}_m)\frac{\psi_k(\vec{\tau}_n)}{\psi_k(\vec{\tau}_m)} \qquad (2.2.4.16)$$

The peculiar looking exchange operator $\hat{K}$ describes the non-local nature of the exchange potential since it operates in both ψ_j and ψ_k spin orbitals which are functions of different coordinates $(\vec{\tau}_m$ and $\vec{\tau}_n)$.

Now that we have provided a mathematical and physical definition of both Coulombic and exchange operator in the HF formalism, let us further simplify Eq. 2.2.4.13.

Equation 2.2.4.13 is equal to zero if the variation of the wave function $\delta\psi_k^*(\vec{\tau}_m) = 0$ or the factor in the integrand which is multiplied by $\delta\psi_k^*(\vec{\tau}_m)$ vanishes. Since $\delta\psi_k^*(\vec{\tau}_m)$ was initially defined as an arbitrary variation different than 0, the factor in the integrand $\left[\hat{h}_k\psi_k(\vec{\tau}_m) + \hat{J}\psi_k(\vec{\tau}_m) - \hat{K}\psi_k(\vec{\tau}_m) - \sum_{j=1}^{J}\varepsilon_{kj}\psi_j(\vec{\tau}_m)\right]$ must be equal to zero which results in:

$$\hat{h}_k\psi_k(\vec{\tau}_m) + \hat{J}\psi_k(\vec{\tau}_m) - \hat{K}\psi_k(\vec{\tau}_m) - \sum_{j=1}^{J}\varepsilon_{kj}\psi_j(\vec{\tau}_m) = 0$$

$$\Leftrightarrow \left[\hat{h}_k\psi_k(\vec{\tau}_m) + \hat{J}\psi_k(\vec{\tau}_m) - \hat{K}\psi_k(\vec{\tau}_m)\right] = \sum_{j=1}^{J}\varepsilon_{kj}\psi_j(\vec{\tau}_m)$$

$$\Leftrightarrow \left[\hat{h}_k + \hat{J} - \hat{K}\right]\psi_k(\vec{\tau}_m) = \sum_{j=1}^{J}\varepsilon_{kj}\psi_j(\vec{\tau}_m) \text{ which yields:}$$

$$\hat{f}\psi_k(\vec{\tau}_m) = \sum_{j=1}^{J}\varepsilon_{kj}\psi_j(\vec{\tau}_m) \qquad (2.2.4.17)$$

where $\hat{f} = \hat{h}_k + \hat{J} - \hat{K}$ is the Fock operator which operates on a single spin orbitals and yields a linear combination of spin orbitals multiplied with the Lagrange multipliers. Unfortunately, the above equation yields multiple different solutions of the HF equations with each corresponding to a different set of Lagrange multipliers. In order to decrease the complexity of the above equation, we will be using a unitary transformation matrix **U** which operates on the HF eigenfunctions without modifying the orthonormality conditions.[114,115,117] By rewriting Eq. 2.2.4.17 in matrix form, we obtain:

$$\hat{\mathbf{f}}\boldsymbol{\psi} = \boldsymbol{\psi}\boldsymbol{\varepsilon} \qquad (2.2.4.18)$$

where $\boldsymbol{\psi} = (\psi_1, \psi_2, \psi_3 \cdots \psi_M)$ is the row matrix of the *M* solutions of Eq. 2.2.4.17. We define the unitary $M \times M$ matrix **U** in such way that $\mathbf{U}^\dagger\mathbf{U} = \mathbf{U}\mathbf{U}^\dagger = 1$.

Let us multiply each side of the equation with the unitary matrix **U** which gives:

$$\hat{\mathbf{f}}\boldsymbol{\psi}\mathbf{U} = \boldsymbol{\psi}\boldsymbol{\varepsilon}\mathbf{U}$$

Then, we will insert the unit matrix of the form $\mathbf{U}\mathbf{U}^{\dagger}$ between $\boldsymbol{\psi}$ and $\boldsymbol{\varepsilon}$ yielding:

$$\hat{\mathbf{f}}\boldsymbol{\psi}\mathbf{U} = \boldsymbol{\psi}\mathbf{U}\mathbf{U}^{\dagger}\boldsymbol{\varepsilon}\mathbf{U}$$

$$\Leftrightarrow \hat{\mathbf{f}}\boldsymbol{\psi}' = \boldsymbol{\psi}'\boldsymbol{\varepsilon}' \text{ where } \boldsymbol{\psi}' = \boldsymbol{\psi}\mathbf{U} \text{ and } \boldsymbol{\varepsilon}' = \mathbf{U}^{\dagger}\boldsymbol{\varepsilon}\mathbf{U} = \begin{pmatrix} \varepsilon_1 & & \mathbf{0} \\ & \varepsilon_2 & \\ \mathbf{0} & & \ddots \\ & & & \varepsilon_M \end{pmatrix} \qquad (2.2.4.19)$$

Here $\boldsymbol{\varepsilon}'$ becomes a unique diagonal matrix which reduces the once complex coupled HF equations into individual eigenvalue equations. The above equation can therefore be written in the canonical form:

$$\hat{f}\psi'_k(\vec{\tau}_m) = \varepsilon'_i\ \psi'_k(\vec{\tau}_m) \qquad (2.2.4.20)$$

Although we have decreased the complexity of the HF equations, we are still left with the calculation of the $\psi'_k(\vec{\tau}_m)$ spin orbitals in multiple regions in space. By using the HF-Roothan procedure[111], we first express the HF spin orbitals as linear combinations of basis functions $\Phi_j(\vec{\tau}_m)$:

$$\psi'_k(\vec{\tau}_m) = \sum_{j=1}^{J} a_{ij}\Phi_j(\vec{\tau}_m) \qquad (2.2.4.21)$$

where a_{ij} are the expansion coefficients and *J* is the size of the basis set. This procedure is very similar to the one described in section 2.2.3 but here, we are truncating our basis set to *J* in order to decrease the mathematical complexity.

Then, we use an iterative process to solve the one-electron HF equation, in which:

- we initially generate a Slater determinant from the trial basis set.
- We later subject the Slater determinant to the Fock operator which encompasses the one-electron operator, a local Coulomb operator and a non-local exchange operator.
- We diagonalize the Fock Hamiltonian and obtain its corresponding eigenvalues and eigenvectors.
- In the final step, we compare the obtained spin orbitals with the ones we initially estimated. In case that the obtained spin orbitals correspond to our initial guess, the total energy and other characteristics of the system can therefore be calculated. On the other hand, if the calculated spin orbitals are inconsistent with the initial ones (within some acceptable margin error), the trial spin orbitals must be updated by using the same iterative process. Such procedure is called the HF self-consistent-field method since orbitals are obtained from their own effective HF potential and keep being updated trough an iterative process until convergence is reached.

More details regarding the derivation of the HF-Roothan[111] equations and the description of each step in the SCF calculation equations are given in Ref. [114]. Although at first, solving the canonical HF equations appears achievable, the accuracy of the HF formalism relies heavily on the specification of the expansion coefficients. Furthermore, the Fock operator includes an average local Coulombic term and a complicated non-local exchange operator which depend on a set of spatial orbitals in both positions $\vec{r}_n$ and $\vec{r}_m$. Accurate calculations of either term require drastic computational time and effort for systems with more than few electrons.

In the derivation of the Fock operator, we have so far restricted the wave function of the system to a single Slater determinant in which electrons of both spin quantum numbers occupy

the same spatial orbital. Such mathematical construct is called the restricted HF (RHF) wave function. Although the HF guarantees a minimum in the total energy associated with the basis set, it does not necessarily guarantee a physical meaning of the resulting spin orbitals. This point is described in the simple example of the H_2 molecule in which a RHF wave function yields ionic terms and hence completely wrong dissociation energies.[118] Circumventing such problem could be obtained by approximating the wave function of a many-body system with a linear combination of Slater determinants (unrestricted Hartree-Fock). In other words, the calculation in the unrestricted HF (UHF) approach allows either paired or unpaired electrons to occupy different spatial orbitals. As a result, the total HF energy from the UHF methods yields lower energy than the one obtained from the RHF method. However, for the case where both nuclei in the H_2 molecule are widely separated, the UHF methods yields unphysical mixture between the singlet and triplet state of the H_2 molecule and also still slightly overestimates the total energy of the system.[118] This overestimation of the ground state energy whether it is in the RHF or UHF, emerges from the many-body correlation effects which correspond to the energy of the correlated motion of electrons.

Within the HF approach, although the exchange energy is treated exactly, the Coulombic term is computed in an average manner. Consequently, the electrostatic repulsion between electrons is overestimated yielding the value of the total HF energy functional greater than the true ground-state energy. In other words, within the HF scheme, electrons tend to be too close to one another regardless of their spin quantum number. This correlation error in the HF method $\left(E_C^{HF}\right)$ is defined as the difference in energy between the total HF energy (E_{HF}) and the true ground state energy $\left(E_0\right)$ of a many-body system:

$$E_C^{HF} = E_{HF} - E_0 .$$

Since the HF energy is derived from the variational principle in which $E_{HF} \geq E_0$, the correlation energy must be positive.

If one is interested in circumventing the correlation effects and going beyond "simple" HF-SCF procedures, the use of expensive post-HF methods such as the configuration interaction, coupled cluster, Moller-Plesset perturbation theory and the quadratic configuration interaction is required. Unfortunately, several of these post-HF methods are only possible for relatively small number of atoms due to the exponential increase of computational expense associated with the size of the system and the number of basis functions.

Although we have discussed the intrinsic errors associated with the HF formalism, we have not discussed the obtained orbital energies ε_i' derived from the canonical HF equations in Eq. 2.2.4.19. Further physical insight regarding ε_i' is done in the next section.

2.2.5. Koopman's Theorem

From the compact form of the canonical HF equations (Eq. 2.2.4.20) that resemble the Schrödinger equation, one would be tempted to deduce that the orbital energy corresponds to the eigenvalue of an electron occupying the spin orbital $\psi'_k(\vec{\tau}_m)$. However, by multiplying both sides of the canonical HF equation by $\psi'^*_k(\vec{\tau}_m)$ and integrate over the $\vec{\tau}_m$ coordinates, we obtain:

$$\int \psi'^*_k(\vec{\tau}_m)\,\hat{f}\,\psi'_k(\vec{\tau}_m)\,d\vec{\tau}_m = \int \psi'^*_k(\vec{\tau}_m)\,\varepsilon'_k\psi'_k(\vec{\tau}_m)\,d\vec{\tau}_m$$
$$\Rightarrow \int \psi'^*_k(\vec{\tau}_m)\left[\hat{h}_k + \hat{J} - \hat{K}\right]\psi'_k(\vec{\tau}_m)\,d\vec{\tau}_m = \varepsilon'_k$$
$$\Leftrightarrow \int \psi'^*_k(\vec{\tau}_m)\left[\hat{h}_k\right]\psi'_k(\vec{\tau}_m)\,d\vec{\tau}_m + \int \psi'^*_k(\vec{\tau}_m)\left[\hat{J} - \hat{K}\right]\psi'_k(\vec{\tau}_m)\,d\vec{\tau}_m = \varepsilon'_k$$

$$\Leftrightarrow \varepsilon'_k = \left\langle \psi'_k(\vec{\tau}_m)\middle|\hat{h}_k\middle|\psi'_k(\vec{\tau}_m)\right\rangle + \sum_{j=1}^{J}\left\langle \psi'_k(\vec{\tau}_m)\psi_j(\vec{\tau}_n)\middle|\frac{1}{r_{mn}}\middle|\psi'_k(\vec{\tau}_m)\psi_j(\vec{\tau}_n)\right\rangle + - \sum_{j=1}^{J}\left\langle \psi'_k(\vec{\tau}_m)\psi_j(\vec{\tau}_n)\middle|\frac{1}{r_{mn}}\middle|\psi_j(\vec{\tau}_m)\psi'_k(\vec{\tau}_n)\right\rangle \quad (2.2.5.1)$$

Let $k = i$, lift the prime indexes and sum both sides of the above equation over the *i* terms:

$$\sum_{i}^{I}\varepsilon_i = \sum_{i=1}^{I}\left\langle \psi_i(\vec{\tau}_m)\middle|\hat{h}_k\middle|\psi_i(\vec{\tau}_m)\right\rangle + \sum_{i=1}^{I}\sum_{j=1}^{J}\left\langle \psi_i(\vec{\tau}_m)\psi_j(\vec{\tau}_n)\middle|\frac{1}{r_{mn}}\middle|\psi_i(\vec{\tau}_m)\psi_j(\vec{\tau}_n)\right\rangle + - \sum_{i=1}^{I}\sum_{j=1}^{J}\left\langle \psi_i(\vec{\tau}_m)\psi_j(\vec{\tau}_n)\middle|\frac{1}{r_{mn}}\middle|\psi_j(\vec{\tau}_m)\psi_i(\vec{\tau}_n)\right\rangle \quad (2.2.5.2)$$

By comparing the expression of the sum of the orbital energies ε_i to the expression of the total HF functional (E_{HF}) derived in Eq. 2.2.4.1, we notice that:

$$E_{HF} = \sum_{i}^{I} \varepsilon_i - \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J}\left\langle \psi_i(\vec{\tau}_m)\psi_j(\vec{\tau}_n)\left|\frac{1}{r_{mn}}\right|\psi_i(\vec{\tau}_m)\psi_j(\vec{\tau}_n)\right\rangle + \\ + \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J}\left\langle \psi_i(\vec{\tau}_m)\psi_j(\vec{\tau}_n)\left|\frac{1}{r_{mn}}\right|\psi_j(\vec{\tau}_m)\psi_i(\vec{\tau}_n)\right\rangle \tag{2.2.5.3}$$

Therefore within the HF approximation, the total energy is not equal to the sum of the orbital energies since such a sum double counts the electron-electron interactions. This difference in energy arises from the approximation of a true many-body system in terms of single-particle spin orbitals.

In order to obtain a physical interpretation of the orbital energy, let us compare the difference in energy between the total HF energy of an *I*-orbital-electrons system and an ionized system in which an electron has been removed from the highest occupied level of the *I*-th spin orbital, i.e, the $(I-1)$ orbital configuration.

Assuming that the remaining $I-1$ electrons do not re-arrange their distribution (electronic relaxation) once the electron is ejected, the difference in energy is written as:

$$\begin{aligned}\Delta E_{HF} &= E_{HF}\big|_I - E_{HF}\big|_{I-1}^{J-1} \\ &= \left[\sum_{i=1}^{I}\left\langle \psi_i(\vec{\tau}_m)\left|\hat{h}_i\right|\psi_i(\vec{\tau}_m)\right\rangle + \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J}\left\langle \psi_i(\vec{\tau}_m)\psi_j(\vec{\tau}_n)\left|\frac{1}{r_{mn}}\right|\psi_i(\vec{\tau}_m)\psi_j(\vec{\tau}_n)\right\rangle\right. \\ &\quad \left. - \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J}\left\langle \psi_i(\vec{\tau}_m)\psi_j(\vec{\tau}_n)\left|\frac{1}{r_{mn}}\right|\psi_j(\vec{\tau}_m)\psi_i(\vec{\tau}_n)\right\rangle\right] + \\ &\quad + \left[\sum_{i=1}^{I-1}\left\langle \psi_i(\vec{\tau}_m)\left|\hat{h}_i\right|\psi_i(\vec{\tau}_m)\right\rangle + \frac{1}{2}\sum_{i=1}^{I-1}\sum_{j=1}^{J-1}\left\langle \psi_i(\vec{\tau}_m)\psi_j(\vec{\tau}_n)\left|\frac{1}{r_{mn}}\right|\psi_i(\vec{\tau}_m)\psi_j(\vec{\tau}_n)\right\rangle\right. \\ &\quad \left. - \frac{1}{2}\sum_{i=1}^{I-1}\sum_{j=1}^{J-1}\left\langle \psi_i(\vec{\tau}_m)\psi_j(\vec{\tau}_n)\left|\frac{1}{r_{mn}}\right|\psi_j(\vec{\tau}_m)\psi_i(\vec{\tau}_n)\right\rangle\right]\end{aligned}$$

(2.2.5.4)

In order to simplify the above expression, let us first work on a simple example that involves the difference between two summations that involve inner products of spin orbitals. Define a function $F\big|_I$ where:

$$F\big|_I = \sum_{i=1}^{3}\sum_{j=1}^{3}\left\langle \psi_i(\vec{\tau})\psi_j(\vec{\tau}')\middle|\psi_i(\vec{\tau})\psi_j(\vec{\tau}')\right\rangle$$

$$\Leftrightarrow F\big|_I = \int\int d\vec{\tau}\,d\vec{\tau}'\left(\psi_1^*+\psi_2^*+\psi_3^*\right)\left(\psi_1^*+\psi_2^*+\psi_3^*\right)\cdot\left(\psi_1+\psi_2+\psi_3\right)\left(\psi_1+\psi_2+\psi_3\right),$$

$$\text{and } F\big|_{I-1}^{J-1} = \sum_{i=1}^{2}\sum_{j=1}^{2}\left\langle \psi_i(\vec{\tau})\psi_j(\vec{\tau}')\middle|\psi_i(\vec{\tau})\psi_j(\vec{\tau}')\right\rangle$$

$$\Leftrightarrow F\big|_{I-1}^{J-1} = \int\int d\vec{\tau}\,d\vec{\tau}'\left(\psi_1^*+\psi_2^*\right)\left(\psi_1^*+\psi_2^*\right)\cdot\left(\psi_1+\psi_2\right)\left(\psi_1+\psi_2\right).$$

$$\text{Then: } \Delta F = F\big|_I - F\big|_{I-1}^{J-1} = \int\int d\vec{\tau}\,d\vec{\tau}'\cdot 2\left[\psi_3^*\left(\psi_1^*+\psi_2^*+\psi_3^*\right)\cdot\psi_3\left(\psi_1+\psi_2+\psi_3\right)\right]$$

$$\Leftrightarrow \Delta F = F\big|_I - F\big|_{I-1}^{J-1} = \sum_{j=1}^{3} 2\left\langle \psi_3(\vec{\tau})\psi_j(\vec{\tau}')\middle|\psi_3(\vec{\tau})\psi_j(\vec{\tau}')\right\rangle$$

By applying the same logic into Eq. 2.2.5.4, ΔE_{HF} is simplified as:

$$\begin{aligned}\Delta E_{HF} &= E_{HF}\big|_I - E_{HF}\big|_{I-1}^{J-1}\\ &= \left\langle \psi_I(\vec{\tau}_m)\middle|\hat{h}_I\middle|\psi_I(\vec{\tau}_m)\right\rangle + \\ &\quad + \frac{1}{2}\sum_{j=1}^{J}\left[2\left\langle \psi_I(\vec{\tau}_m)\psi_j(\vec{\tau}_n)\middle|\frac{1}{r_{mn}}\middle|\psi_I(\vec{\tau}_m)\psi_j(\vec{\tau}_n)\right\rangle\right]\\ &\quad - \frac{1}{2}\sum_{j=1}^{J}\left[2\left\langle \psi_I(\vec{\tau}_m)\psi_j(\vec{\tau}_n)\middle|\frac{1}{r_{mn}}\middle|\psi_j(\vec{\tau}_m)\psi_I(\vec{\tau}_n)\right\rangle\right]\end{aligned} \qquad (2.2.5.5)$$

Since $\frac{1}{r_{mn}}$ is a Hermitian operator, we can exchange the indices of the orbitals if performed simultaneously in either side of the operator:

$$\begin{aligned}\Delta E_{HF} &= E_{HF}\big|_I^J - E_{HF}\big|_{I-1}^{J-1} \\ &= \left\langle \psi_I(\vec{\tau}_m)\middle|\hat{h}_I\middle|\psi_I(\vec{\tau}_m)\right\rangle + \\ &\quad + \sum_{j=1}^{J}\left[\left\langle \psi_I(\vec{\tau}_m)\psi_j(\vec{\tau}_n)\middle|\frac{1}{r_{mn}}\middle|\psi_I(\vec{\tau}_m)\psi_j(\vec{\tau}_n)\right\rangle\right] \\ &\quad - \sum_{j=1}^{J}\left[\left\langle \psi_I(\vec{\tau}_m)\psi_j(\vec{\tau}_n)\middle|\frac{1}{r_{mn}}\middle|\psi_j(\vec{\tau}_m)\psi_I(\vec{\tau}_n)\right\rangle\right]\end{aligned} \tag{2.2.5.5}$$

where $\psi_I(\vec{\tau}_m)$ is the highest occupied spin orbital. By comparing the above equation with Eq. 2.2.5.1, we notice that:

$$\Delta E_{HF} = E_{HF}\big|_I^J - E_{HF}\big|_{I-1}^{J-1} = \varepsilon_I,$$

where ε_I is the energy of the highest occupied molecular orbital. From the definition of the ionization potential *IP* from which:

$$IP = E_{HF}\big|_{I-1}^{J-1} - E_{HF}\big|_I^J = -\Delta E_{HF} = -\varepsilon_I, \tag{2.2.5.6}$$

we have proven Koopman's theorem which states that for "frozen" orbitals, the negative of the orbital energy of the highest occupied molecular orbital corresponds to the first ionization energy.

Likewise, the HF orbital energy can also be related to the electron affinity (*A*), i.e, the energy required to add an electron into an empty HF orbital. By definition, the electron affinity is expressed as:

$$A = E_{HF}\big|_{I}^{J} - E_{HF}\big|_{I+1}^{J+1} = -\varepsilon_{I+1}, \tag{2.2.5.7}$$

where ε_{I+1} corresponds to the energy of the lowest unoccupied orbital. Nevertheless, we expect Koopman's predictions for the first ionization potential and electron affinity to be only an approximation. The use of Eqs. 2.2.5.6 and 2.2.5.7 views the ionization potential (electron affinity) as a simple removal (addition) of an electron without reorganization of the remaining electronic charge. This complete lack of relaxation overestimates both the values of $E_{HF}\big|_{I-1}^{J-1}$ for the *IP* and $E_{HF}\big|_{I+1}^{J+1}$ in the calculation of *A*. Consequently, one would deduce that Koopman's theorem would yield ionization potentials that are too positive when compared to experimental data while it would underestimate the electron affinities of molecules. In addition to relaxation errors, Koopman's approximation does not take into account correlation effects (intrinsic inaccuracy in HF theory) of a many-body system. As we have previously discussed, the HF correlation energy is calculated to be positive and we hence expect the energy of the neutral molecule $E_{HF}\big|_{I}^{J}$ to yield greater correlation energy than the energy of the cation $E_{HF}\big|_{I-1}^{J-1}$ (since the neutral molecule possesses more electrons than its corresponding cation). Given that the relaxation energy overestimates $E_{HF}\big|_{I-1}^{J-1}$ and the correlation energy overestimates the value of $E_{HF}\big|_{I}^{J}$, we should presume a partial cancellation of errors in using Koopman's approximation of the first *IP* of molecules. Nonetheless, regarding the calculation of the electron affinity, the

correlation error of the anion adds to the relaxation error when computing the value of $E_{HF}\big|_{I+1}^{J+1}$ which consequently yields inaccurate predictions of the first electron affinities in molecules. Overall, Koopman's predictions of first ionization potentials are only reasonable first approximation to experimental *IP* in molecules. However, in solids, Koopman's approximation of the *IP* tends to produce worse results since the correlation and screening effects become more pronounced due to periodic boundary conditions.

In conclusion, we have seen that in the HF method, the many-body wave function is expressed a Slater determinant that is constructed from a set of *I* one-electron spin orbitals. However, from the constrained form of a Slater determinant and the treatment of the electrostatic interaction between electrons in an average manner, we either obtain optimized spin orbitals that are physically unreasonable or an overestimation of the ground state energy of a multi-electron system.

Instead of approaching the multi-electron problem from a rather complex wave function standpoint, one might reconsider approximating the solution of the many-body Schrodinger equation from an electron density perspective. By using an electron density type of approach, one can tentatively decrease the complexity of the non-local exchange interaction term and possibly improve inter-electronic correlation effects. To appreciate the full extent of the use of electron density in the approximation of the Schrödinger equation of a many-body system, a brief analysis of the concept of functionals is given in the following section.

2.3. Functional and Functional derivatives

The mathematical and physical explanations of functional and functional derivatives are crucial if one needs to understand the quantum mechanical treatment of interacting electrons in terms of DFT or hybrid functional theories.

A functional is closely related to the more familiar concept of a function and one recalls that a function *f* is defined to be a mapping of a variable *y* to a number *f(y)*. A functional, on the other hand, i.e $F[f(y)]$, assigns a unique number to an entire function and is therefore a mapping of a function *f* onto a value or number.[110] In other words, a functional can be roughly defined as a function of a function. A fairly simple example of a functional can be conceptualized by looking at the total charge enclosed Q in a closed surface *S* bounded by a given volume Ω:

$$Q = \int_{\Omega} n(\vec{r})d\vec{r} = Q[n(\vec{r})] \tag{2.3.1}$$

where $n(\vec{r})$ is the total charge density of the system. From this functional, we notice that $Q[n(\vec{r})]$ is a rule for going from a function $n(\vec{r})$ to a number Q. The square bracket notation $Q[n(\vec{r})]$ indicates that Q depends on $n(\vec{r})$ everywhere in the volume Ω. Moreover, $Q[n(\vec{r})]$ is a *local-functional* since the functional does not depend on its gradient, Laplacian or other higher-order derivatives. More detailed analysis of functionals is given by Volterra[119] (1959), and Parr and Yang[110] (1989).

Differentiation of functionals is an extension of the concept of partial differentiation for multi-variable functions[120]. Functional derivatives basically allow us to study how a functional changes with respect to variation of a function *f* at the point *y*. Let a function *f(y)* be defined over a specific interval $[y_{\min}, y_{\max}]$, and be a subject to an arbitrary small perturbation $\delta f(y)$ that is

of the form $f(y) \rightarrow f(y) + \delta f(y)$. Then the perturbation $\delta f(y)$ is also defined over the same interval $[y_{\min}, y_{\max}]$. For some functional $F[f(y)]$, the value of $F[f(y) + \delta f(y)]$ can be therefore approximated by using a Taylor's[121] expansion in powers of the perturbation $\delta f(y)$:

$$F[f(y) + \delta f(y)] = F[f(y)] + \int_{y_{\min}}^{y_{\max}} \Lambda_1(y) \delta f(y) dy +$$

$$+ \frac{1}{2!} \int_{y_{\min}}^{y_{\max}} \Lambda_2(y, y_1) \delta f(y) \delta f(y_1) dy_1 dy_2 + \ldots \qquad (2.3.2)$$

where $\Lambda_1(y) = \dfrac{\delta F[f(y)]}{\delta f(y)}$ is the first Taylor's expansion coefficient and describes the rate of change of the functional or the functional slope for a small variation of *f* at *y*. Correspondingly, $\Lambda_2(y, y_1) = \dfrac{\delta^2 F[f(y)]}{\delta f(y) \delta f(y_1)}$ is the second Taylor's expansion coefficient and described the rate of change of the functional when *f* is simultaneously subjected to small perturbations at *y* and y_1. Since the functional derivative only measures the first order change in a functional, the second integral of Eq. 2.3.2 is neglected; hence the quantity $\dfrac{\delta F[f(y)]}{\delta f(y)}$ becomes the *functional derivative of F with respect to f* at the point *y* and Eq. 2.3.2 can be approximated as: [110]

$$F[f(y) + \delta f(y)] \approx F[f(y)] + \int_{y_{\min}}^{y_{\max}} \frac{\delta F[f(y)]}{\delta f(y)} \delta f(y) dy$$

$$\Leftrightarrow F[f(y) + \delta f(y)] - F[f(y)] \approx \int_{y_{\min}}^{y_{\max}} \frac{\delta F[f(y)]}{\delta f(y)} \delta f(y) dy$$

$$\Leftrightarrow \delta F[f(y)] \approx \int_{y_{\min}}^{y_{\max}} \frac{\delta F[f(y)]}{\delta f(y)} \delta f(y) dy \qquad (2.3.3)$$

If the functional is a function of another function and a constant *k*, i.e. $F\left[f\left(y,k\right)\right]$, then according to Ref. [110], the partial derivative of *F* with respect to the constant *k* is expressed as:

$$\frac{\partial F\left[f\left(y,k\right)\right]}{\partial k}=\int \frac{\delta F\left[f\left(y,k\right)\right]}{\delta f\left(y\right)}\frac{\partial f\left(y\right)}{\partial k}dy \tag{2.3.4}$$

In case that a function of a functional such as $\Gamma\left\{F\left[f\left(y\right)\right]\right\}$ is differentiable, we obtain from the chain rule:[110]

$$\frac{\delta\Gamma\left\{F\left[f\left(y\right)\right]\right\}}{\delta f\left(y\right)}=\frac{d\Gamma\left\{F\left[f\left(y\right)\right]\right\}}{dF\left[f\left(y\right)\right]}\frac{\delta F\left[f\left(y\right)\right]}{\delta f\left(y\right)} \tag{2.3.5}$$

Now that we have briefly described what a functional is and how to differentiate functionals, we can introduce the Hohenberg-Kohn theorems in the following section.

2.4. The Hohenberg-Kohn (HK) theorems

One may argue that the very core of the entire field of DFT rests on the two HK theorems.[7] Before introducing the HK theorems, let us introduce the concept of electron density and the components of the HK Hamiltonian.

Associated with each individual electron, there is a spin orbital $\Psi_m(\vec{r},s)$ whose intensity at position $\vec{r}$ denotes the probability *P* of finding the *m*-th electron in a volume element $d\vec{r}$ at a point $\vec{r}$:

$$P = \Psi_m^*(\vec{r},s)\Psi_m(\vec{r},s)d\vec{r}$$
$$\Leftrightarrow P = \left|\Psi_m(\vec{r},s)\right|^2 d\vec{r}$$

By dividing both sides of the above equation by the volume element $d\vec{r}$, we obtain the probability density or the charge density of the *m*-th electron $n_m(\vec{r})$ which is expressed as:

$$n_m(\vec{r},s) = \left|\Psi_m(\vec{r},s)\right|^2$$

The total charge density is then expressed as:

$$n(\vec{r},s) = n_\alpha(\vec{r}) + n_\beta(\vec{r}) = \sum_{s=\alpha,\beta}\left[\sum_{m=1}^{M} n_m(\vec{r},s)\right] = \sum_{s,m=1}^{M}\left|\Psi_m(\vec{r},s)\right|^2$$

By taking the integral of each side of the above equation over the volume element $d\vec{r}$:

$$\int n(\vec{r},s)d\vec{r} = \int \sum_{s,m=1}^{M} \left|\Psi_m(\vec{r},s)\right|^2 d\vec{r}$$

$$\Leftrightarrow \int n(\vec{r},s)d\vec{r} = \sum_{s,m=1}^{M} \int \left|\Psi_m(\vec{r},s)\right|^2 d\vec{r}$$

$$\Leftrightarrow \int n(\vec{r},s)d\vec{r} = \sum_{s,m=1}^{M} 1$$

$$\Leftrightarrow \int n(\vec{r},s)d\vec{r} = M \qquad (2.4.1)$$

where M is the total number of electrons. The inclusion of spin is mathematically intricate and not necessary for the discussion of the HK theorems. Further discussion related to spin character of the charge density is given in section 2.6. In non-spin polarized systems, the total number of electrons M is expressed as:

$$M = \int n(\vec{r})d\vec{r} \qquad (2.4.2)$$

Now let's consider M interacting electrons subjected to an external potential energy $\hat{U}_{ext}$, which operates on the coordinates of each electron as:

$$\hat{U}_{ext} = \hat{U}_{mp} = \sum_{m=1}^{M}\sum_{p=1}^{P} \frac{Z_p}{\left|\vec{r}_m - \vec{r}_p\right|} = \sum_{m=1}^{M} u_{ext}(\vec{r}_m) \qquad (2.4.3)$$

where $u_{ext}(\vec{r}_m) = \sum_{p=1}^{P} \frac{Z_p}{\left|\vec{r}_m - \vec{r}_p\right|}$.

Thus far, we have defined $\hat{U}_{ext}$ as an operator function of the position of each electron and nuclei while M denotes the total number of electrons in the system.

Therefore, $\hat{U}_{ext}$ and M establish all properties for the construction of the electronic HK Hamiltonian:

$$\hat{H}_{HK} = \hat{T}_m + \hat{U}_{mn} + \hat{U}_{ext} \tag{2.4.4}$$

Here $\hat{T}_m$ and $\hat{U}_{mn}$ are the kinetic and electron-electron potential operator describes in Section 2.1. By applying the variational principle (see section 2.2.3) to the HK Hamiltonian, one can "technically" obtain the ground state wave function which consequently allows the calculation of the ground state energy and any other properties of the system. Therefore, the external potential and M indirectly determine all the properties required for the calculation of the ground-state density $n(\vec{r})$ of the system.

2.4.1. The first HK Theorem

The first theorem established by Hohenberg and Kohn substitutes $\hat{U}_{ext}$ and *M* for the use of the electronic density $n(\vec{r})$ as a variable. The correlation between the external potential and the electronic density can be regarded as follows:

For any system of *M* interacting electrons subjected to an external potential, *the external potential is a unique functional (within a trivial additive constant) of the electronic density* $n(\vec{r})$.[110,107] Schematically, this can be expressed as:

$$n(\vec{r}) \rightarrow \left\{M, \hat{U}_{ext}\right\} \rightarrow \hat{H}_{HK} \rightarrow \Psi \rightarrow E\left[n(\vec{r})\right] \tag{2.4.1.1}$$

In other words, the electron density uniquely defines all electronic properties of the system. The proof of the first theorem is quite straight-forward and can be obtained via the indirect contradiction method (*reduction ad absurdum*) described in any DFT book. Since the total energy is a unique functional of the electron density, the expectation values of each component of the previously derived HK Hamiltonian (Eq. 2.4.4) can be expressed as:

$$E_{HK}\left[n(\vec{r})\right] = \left\{T_m\left[n(\vec{r})\right] + U_{mn}\left[n(\vec{r})\right]\right\} + \int u_{ext}\left(\vec{r}\right) n(\vec{r}) d\vec{r} \tag{2.4.1.2}$$

$$\Rightarrow E_{HK}\left[n(\vec{r})\right] = F_{HK}\left[n(\vec{r})\right] + \int u_{ext}\left(\vec{r}\right) n(\vec{r}) d\vec{r}\,, \tag{2.4.1.3}$$

Here $F_{HK}\left[n(\vec{r})\right]$ is the HK functional that describes the expectation value of the kinetic and electron-electron potential energy operators. On the contrary to the $\int u_{ext}\left(\vec{r}\right) n(\vec{r}) d\vec{r}$ term, the HK functional $F_{HK}\left[n(\vec{r})\right]$ is completely independent of the total number electrons *M*, the nuclei configurations r_p and atomic number Z_p of the nuclei. Hence $F_{HK}\left[n(\vec{r})\right]$ is a universal functional that can in principle be used to solve exactly the many-body Schrödinger equation. Although the

first HK theorem rigorously proves the existence of $E\left[n(\vec{r})\right]$ that can be used to solve the many-electron system, the theorem says nothing on how to minimize it and how to obtain a correct electronic density.

2.4.2. The second HK Theorem

Fortunately, the second HK theorem describes a central property of the functional. It states that for any positive trial electron density $n_{Trial}(\vec{r})$ such that $M = \int d\vec{r} n_{Trial}(\vec{r})$, its corresponding HK energy functional $E\left[n_{Trial}(\vec{r})\right]$ is always greater than or equal to the true ground state energy. In other words:

$$E\left[n_{Trial}(\vec{r})\right] = F_{HK}\left[n_{Trial}(\vec{r})\right] + \int u_{ext}(\vec{r}) n_{Trial}(\vec{r}) d\vec{r} > E_0 \text{ if } n_{Trial}(\vec{r}) \neq n_0(\vec{r})$$
$$\text{or } E\left[n_0(\vec{r})\right] = F_{HK}\left[n_0(\vec{r})\right] + \int u_{ext}(\vec{r}) n_0(\vec{r}) d\vec{r} = E_0 \text{ if } n_{Trial}(\vec{r}) = n_0(\vec{r})$$

From the above expressions, one can therefore vary the electron density until the energy of the functional is minimized, hence giving a prescription for the calculation of the appropriate electron density.[122] Nevertheless, the domain over which $E\left[n_0(\vec{r})\right]$ is defined is only for the set of electronic densities for which one must find a ground state (belonging to some external potential $\hat{U}_{ext}$) that minimizes the total energy. More details regarding the *M*- and $\hat{U}_{ext}$-representability is given in Engel and Dreizler (2009)[96] [pps. 29-31]. Further simplicity over the domain search of $E\left[n_0(\vec{r})\right]$ is obtained via the Levy constrained-search-formalism[123,124,125,126] in which an extra step is added to the HK search method. Here the search for the true ground state energy is performed in two stages: [118,127]

- Stage 1:

  We initially search over a subset Γ_1 of all antisymmetric *M*-electron wave functions that upon integration generate a specific density $n(\vec{r})_{\Gamma_1}$ (subjected to the constraint $M = \int d\vec{r} n(\vec{r})$) that gives the overall lowest energy for this particular density $n(\vec{r})_{\Gamma}$. This can be pictorially expressed as: $\min_{\Psi \to n(\vec{r})} \left\langle \Psi \left| \hat{T}_m + \hat{U}_{mn} + \hat{U}_{ext} \right| \Psi \right\rangle \Big|_{\Gamma_1}$

- Stage 2:

  We then eliminate the constraint of a specific density and spread the search over the set of all densities $(\Gamma_1, \Gamma_2, \cdots \Gamma_M)$ until we find the density that yields the overall lowest total energy. This is schematically represented as:

$$E\left[n_0(\vec{r})\right] = \min_{n(\vec{r})} \left\{ \min_{\Psi \to n(\vec{r})} \left\langle \Psi \left| \hat{T}_m + \hat{U}_{mn} + \hat{U}_{ext} \right| \Psi \right\rangle \Big| \right\}_{(\Gamma_1, \Gamma_2, \cdots \Gamma_M)} \tag{2.4.2.1}$$

In other words, we are searching over all allowed antisymmetric *M*-particle wave functions until finding an electronic density which, among the set of densities, generates the minimum value of $\left\langle \hat{T}_m + \hat{U}_{mn} + \hat{U}_{ext} \right\rangle$. By applying the Levy-constrained[123-126] search formalism to the HK universal functional F_{HK}, we have:

$$F_{HK}\left[n(\vec{r})\right] = \min_{\Psi \to n(\vec{r})} \left\langle \Psi \left| \hat{T}_m + \hat{U}_{mn} \right| \Psi \right\rangle$$

Here, F_{HK} searches over all the ensemble of statistical mixtures which yield the best density $n(\vec{r})$, and provides the minimum expectation value $\left\langle \hat{T}_m + \hat{U}_{mn} \right\rangle$. This approach eliminates the *M*- and $\hat{U}_{ext}$-representability problems from the second HK theorem and provides an easier way to

carry out the domain search. Such method is analogous to the variational principle discussed in section 2.2.3 and the method of Lagrange multipliers described in section 2.2.4. However, in the HK case, instead of imposing the constraints of orthonormality conditions (HF formalism), we shall impose the constraint $M = \int d\vec{r} n(\vec{r})$, which can be written as:

$$\delta E_{HK}\left[n(\vec{r})\right] = 0,$$

subjected to $\Theta\left[n(\vec{r})\right] = M - \int d\vec{r} n(\vec{r}) = 0.$ (2.4.2.2)

By using the same procedure used in the derivation of the Lagrange multipliers for the HF functional, we obtain:

$$\delta E_{HK}\left[n(\vec{r})\right] - \mu \cdot \delta\Theta\left[n(\vec{r})\right] = 0, \quad (2.4.2.3)$$

where μ is the Lagrange multipliers that ensures the correct value of *M*. For any small variations of the electron density of the form $n(\vec{r}) \to n(\vec{r}) + \delta n(\vec{r})$, the functional derivative of $\Theta\left[n(\vec{r})\right]$ can be expressed as:

$$\begin{aligned}
&\Theta\left[n(\vec{r}) + \delta n(\vec{r})\right] = M - \int d\vec{r} \cdot \left[n(\vec{r}) + \delta n(\vec{r})\right] \\
&\Leftrightarrow \Theta\left[n(\vec{r}) + \delta n(\vec{r})\right] = \underbrace{M - \int d\vec{r} n(\vec{r})}_{\Theta\left[n(\vec{r})\right]} + \int d\vec{r} \delta n(\vec{r}) \\
&\Leftrightarrow \Theta\left[n(\vec{r}) + \delta n(\vec{r})\right] - \Theta\left[n(\vec{r})\right] = \int d\vec{r} \delta n(\vec{r}) \\
&\Leftrightarrow \delta\Theta\left[n(\vec{r})\right] = \int d\vec{r} \delta n(\vec{r}) \quad (2.4.2.4)
\end{aligned}$$

By inserting the above equation into Eq. 2.4.2.3, we obtain the following:

$$\begin{aligned}
&\delta E_{HK}\left[n(\vec{r})\right] - \mu \cdot \int d\vec{r} \delta n(\vec{r}) = 0 \\
&\Leftrightarrow \mu = \frac{\delta E_{HK}\left[n(\vec{r})\right]}{\delta n(\vec{r})} \quad (2.4.2.5)
\end{aligned}$$

From the above equation, we are allowing any infinitesimal variations of the electron density, which indirectly suggest the existence of the total HK energy functional $E\left[n(\vec{r})\right]$ for fractional particle numbers. Nonetheless, all energy functionals that were described up to now were only defined for the integer number *M*. Consequently, one needs to find a way to extend the expression of the total HK energy functional to non-integer numbers.

2.4.2.1. Non-Integer Particle Number and Derivative Discontinuity

Very thorough mathematical and physical explanations of the derivative discontinuity in DFT is given in Ref. [141] and by Engel and Dreizler [96] (pps. 37-39). In this section, we will be using the very intuitive description of C. Ullrich[127] (pp. 24-25) to introduce the concept of derivative discontinuity. Let us define a trial density $n_{Trial}\left(\vec{r}\right)$ that integrates up to $M+q$ as:

$$\int d\vec{r} n_{Trial}\left(\vec{r}\right) = M + q$$

where *M = 1, 2, 3,...* and $0 \le q \le 1$. At the extremum values of q $\left(q = 0 \text{ or } q = 1\right)$, the trial density integrates to the integer number *M* and Eq. 2.4.2.5 can be expressed as:

$$\mu = \frac{\delta E_{HK}\left[M\left[n\left(\vec{r}\right)\right]\right]}{\delta n\left(\vec{r}\right)} = \frac{dE_{HK}\left(M\right)}{dM}\frac{\delta M\left[n\left(\vec{r}\right)\right]}{\delta n\left(\vec{r}\right)}, \qquad (2.4.2.1.1)$$

which follows from the differentiable functional rule of Eq. 2.2.5.

For any small variations of the form $n(\vec{r}) \to n(\vec{r}) + \delta n(\vec{r})$, the functional derivative of M from Eq. 2.4.2.2 can be written as:

$$M\left[n(\vec{r}) + \delta n(\vec{r})\right] = \underbrace{\int d\vec{r} n(\vec{r})}_{M\left[n(\vec{r})\right]} + \int d\vec{r}\,\delta n(\vec{r})$$

$$\Leftrightarrow M\left[n(\vec{r}) + \delta n(\vec{r})\right] - M\left[n(\vec{r})\right] = \int d\vec{r}\,\delta n(\vec{r})$$

$$\Leftrightarrow \delta M\left[n(\vec{r})\right] = \int d\vec{r}\,\delta n(\vec{r})$$

$$\Rightarrow \frac{\delta M\left[n(\vec{r})\right]}{\delta n(\vec{r})} = 1$$

By substituting the above equation into Eq. 2.4.2.1.1, the Lagrange multipliers can be expressed as:

$$\mu = \frac{dE_{HK}(M)}{dM}, \tag{2.4.2.1.2}$$

In other words, the Lagrange multipliers μ denote the chemical potential of an M –integer electrons system.

Now, let us create an open M-electrons system containing a gap that is attached to some type of electron reservoir. In such case, the chemical potential μ behaves very much like the Fermi level in semiconductors at low temperatures; where all levels below μ become occupied and the levels above μ are completely empty. Once the levels' occupational number in the open system changes, i.e. the lowest unoccupied level becomes occupied, an additional electron is therefore allowed to be added into the system and the total number of electrons will immediately change from M to M+1. In order to reflect the new change in occupational number, the chemical potential will increase by a value consistent with M+1. Therefore, μ can be represented as a mathematical step function of the total particle number. Now that we have briefly described how

an *M*-electrons system changes once we add an electron, let us relate it to the trial density $n_{Trial}(\vec{r})$ that integrates to non-integer particle number $M+q$.

From statistical mechanics, the probability density $\left(P_0^{M+1}\right)$ for an *M*+1-electrons *pure state* $\left(\Psi_0^{M+1}\right)$ containing fractional occupation number *q* is given by:[141]

$$P_0^{M+1} = qn_0^{M+1}(\vec{r}) = q\left\langle \Psi_0^{M+1*}(\vec{r}) \middle| \Psi_0^{M+1}(\vec{r}) \right\rangle, \tag{2.4.2.1.3}$$

while the electronic probability density $\left(P_0^M\right)$ of the *M*-electrons *pure state* is expressed as:[141]

$$P_0^M = (1-q)n_0^M(\vec{r}) = (1-q)\left\langle \Psi_0^{M*}(\vec{r}) \middle| \Psi_0^M(\vec{r}) \right\rangle \tag{2.4.2.1.4}$$

Therefore, the total probability density that integrates to a non-integer particle number is achieved via the sum of the probability ground state densities of the *M*- and *M*+1-electrons system:

$$P_0 = P_0^{M+1} + P_0^M$$

$$\Leftrightarrow P_0 = q\left\langle \Psi_0^{M+1*}(\vec{r}) \middle| \Psi_0^{M+1}(\vec{r}) \right\rangle + (1-q)\left\langle \Psi_0^{M*}(\vec{r}) \middle| \Psi_0^M(\vec{r}) \right\rangle \tag{2.4.2.1.5}$$

The derivation of P_0 is based on quite loose mathematical arguments, but more accurate formulation is described in Ref. [141]. By using the same statistical argument from the previous equation on the total HK energy for fractional number, we obtain:

$$E[M+q] = qE[M+q] + (1-q)E[M] \tag{2.4.2.1.6}$$

From the above expression, we therefore conclude that the energy for fractional numbers is a piecewise linear function that displays kinks or derivative discontinuities at integer values (schematically shown in Fig. 1).

Now let us relate the bandgap of our open system to the HK derivative discontinuities that we just discussed. By definition, the bandgap (E_g) is given by:

$$E_g = IP - A \tag{2.4.2.1.7}$$

where *IP* and *A* are the ionization potential and electron affinity discussed in section 2.2.5, respectively. For mathematical simplicity, let us define the fractional particle number $M+q$ as:

$$N = M + q\,. \tag{2.4.2.1.8}$$

For any infinitesimal deviations of N of the form $N \pm \eta$, the limits of the derivatives of the piecewise function $E\left[M+q\right] = E\left[N\right]$ (shown in Fig. 1) can be expressed as:

$$\lim_{\eta\to 0}\mu\big|_{N+\eta} = \lim_{\eta\to 0}\left.\frac{\partial E\left[N\right]}{\partial N}\right|_{N+\eta} = \left.\frac{\delta E_{HK}\left[n(\vec{r})\right]}{\delta n(\vec{r})}\right|_{N+\eta} = E\left[M\right] - E\left[M-1\right] = -IP, \tag{2.4.2.1.9}$$

$$\lim_{\eta\to 0}\mu\big|_{N-\eta} = \lim_{\eta\to 0}\left.\frac{\partial E\left[N\right]}{\partial N}\right|_{N-\eta} = \left.\frac{\delta E_{HK}\left[n(\vec{r})\right]}{\delta n(\vec{r})}\right|_{N-\eta} = E\left[M+1\right] - E\left[M\right] = -A, \tag{2.4.2.1.10}$$

where $\mu\big|_{N\pm\eta}$ is the chemical potential described in Eq. 2.4.2.5 which can either correspond to the ionization potential (*IP*) or the electron affinity (*A*). Therefore, the HK bandgap is given by:

$$E_g^{HK} = IP - A = \left.\frac{\delta E_{HK}\left[n(\vec{r})\right]}{\delta n(\vec{r})}\right|_{N+\eta} - \left.\frac{\delta E_{HK}\left[n(\vec{r})\right]}{\delta n(\vec{r})}\right|_{N-\eta}, \tag{2.4.2.1.11}$$

which means that $E_g^{HK} \neq E_g$ since $\dfrac{\delta E_{HK}\left[n(\vec{r})\right]}{\delta n(\vec{r})}$ is undefined at the extremum values of *q*. In other words, the band gap of an interacting system differs from the HK bandgap by the magnitude of the derivative discontinuity. More details on the physical meaning of the derivative discontinuity are given in page 27 of Ref. [127].

Although the two HK theorems discussed in the above sections provide an exact approach to solve the many-body Schrödinger equation via the electronic density, nothing about the explicit form of the universal function F_{HK} is discussed. Hence, we are still left with the initial problem of a multi-electrons system subjected to an external potential with no acceptable solution in sight. In the following chapters, we will discuss various methods that allow possible construction or approximation of the universal functional.

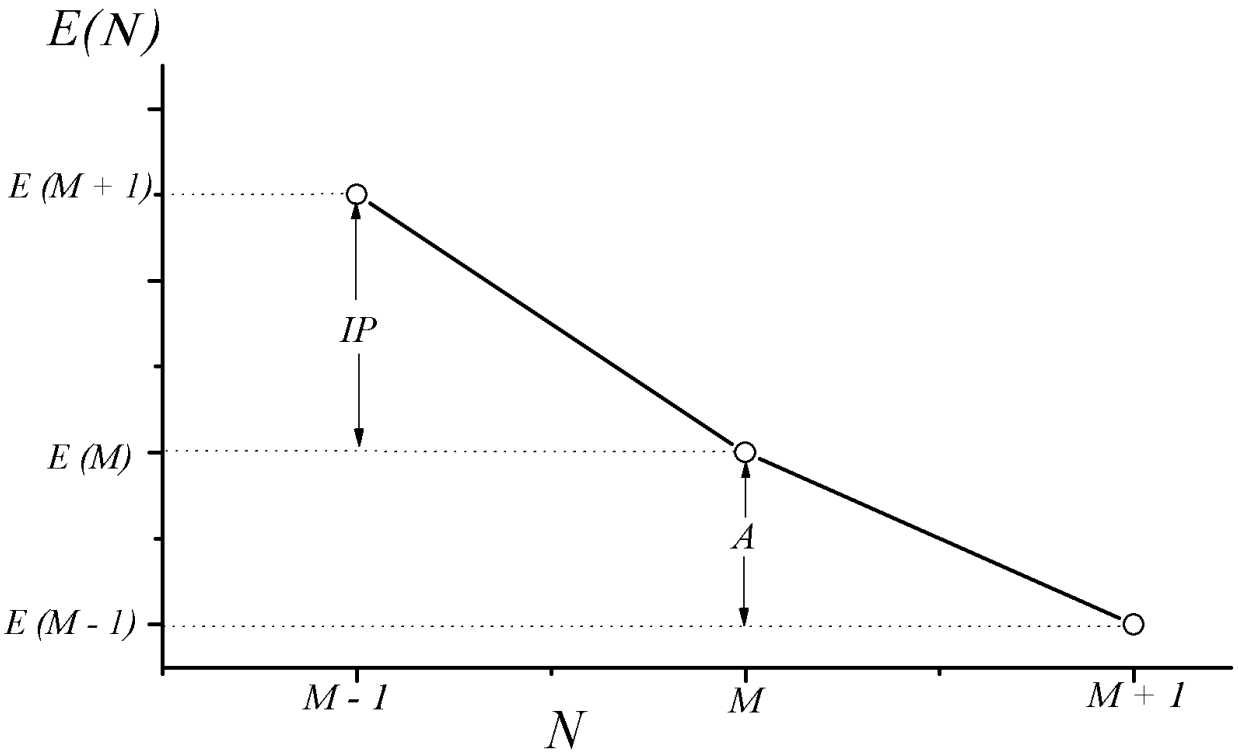

Figure 1: Schematic illustration of the Hohenberg-Kohn total energy $E(N)$ as a function of non-integer number *N*. At extremum values of *q*, there are kinks for which $\frac{dE(N)}{dN}$ does not exist. Here *IP* and *A* denote the ionization potential and electron affinity, respectively.

2.5. Thomas-Fermi-Dirac approximation

Even though the Thomas-Fermi-Dirac (TFD)[101,102,103] approximation historically precedes the HK theorems, we decided to introduce the TFD method right after the brief description of the HK theorems. Such odd structured plan can be justified since one requires the HK theorems to validate the use of electron density functional as a method to possibly solve a multi-electrons system.

A simple, although crude method to solve the many-body Schrödinger equation was provided in the early days of quantum mechanics by expressing the total energy of the system as a functional of a one-electron density:

$$\hat{H}_{TF}\left[n(\vec{r})\right] = \hat{T}_{TF}\left[n(\vec{r})\right] + \frac{1}{2}\int\int \frac{n(\vec{r})n(\vec{r}')}{|\vec{r}-\vec{r}'|} d\vec{r}' d\vec{r} + \int u_{ext}(\vec{r})\, n(\vec{r}) d\vec{r}. \tag{2.5.1}$$

In such formalism, the electron-electron interaction potential is approximated with the Hartree term that was introduced in section 2.1 while the electron-ion term is represented by the external potential (cf. section 2.1). However, we are still left with the construction of the complex kinetic energy functional.

The starting point for the development of the kinetic energy functional in the TF method was to express the kinetic operator as the sum of one-electron kinetic operator:

$$\hat{T}_{TF}\left[n(\vec{r})\right] = \sum_{m=1}^{M} \hat{t}_{TF}\left[n(\vec{r}_m)\right] \tag{2.5.2}$$

The expectation value of the above expression is given by:

$$\left\langle \Psi \left| \hat{T}_{TF} \right| \Psi \right\rangle = T_{TF}\left[n(\vec{r})\right] = \int \hat{t}_{TF}\left[n(\vec{r})\right] \cdot n(\vec{r})\, d\vec{r} \tag{2.5.3}$$

The TF approach to the description of the kinetic operator is based on the motion of non-interacting electrons in the uniform gas model. In such model, electrons are completely free, except for an attractive potential of a smeared out positive background. The confinement of electrons in the cube is imposed by subjecting the system to periodic boundary conditions. Then, we define a Fermi sphere of radius k_F (Fermi momentum) for which all states outside of the sphere are empty and the ones inside the sphere are occupied by two electrons subjected to the Pauli spin exclusion principle. As a result, the total number of electrons M_k in allowed k-states, is given by the ratio of the total volume of the sphere $V = \left(\frac{4\pi}{3} k_F^3 \right)$ divided by the volume of k-space per k-point $\delta k^3 = \left(\frac{2\pi}{L} \right)^3$:

$$M_k = 2 \cdot \frac{V}{\delta k^3} = 2 \cdot \left(\frac{4\pi}{3} k_F^3 \right) \left(\frac{L}{2\pi} \right)^3 = \frac{L^3}{3\pi^2} k_F^3 \tag{2.5.4}$$

Here, L is the length of the side of the cube and the factor "2" corresponds to the number of electrons with opposite quantum numbers occupying the same k-state. Therefore, the corresponding uniform electron density $n_k(\vec{r})$ is given by:

$$n_k(\vec{r}) = \frac{M_k}{V} = \frac{2\left(\frac{4\pi}{3} k_F^3 \right) \left(\frac{2\pi}{L} \right)^{-3}}{L^3}$$
$$= \frac{2\left(\frac{4\pi}{3} k_F^3 \right) \left(\frac{L}{2\pi} \right)^3}{L^3}$$

$$\Rightarrow n_k(\vec{r}) = \left(\frac{k_F^3}{3\pi^2} \right) \text{ or } k_F^2 = \left[3\pi^2 n_k(\vec{r}) \right]^{\frac{2}{3}} \tag{2.5.5}$$

By solving the Schrödinger equation of our simple system subjected to periodic boundary conditions, the available kinetic energy of each region in *k*-space is given by:

$$\varepsilon_k = \frac{\hbar^2 k^2}{2m} \tag{2.5.6}$$

Then the total kinetic energy (T_k) of the electrons within the volume V is the sum of the occupied k values in the Fermi sphere:

$$T_k = 2\sum_{k}^{k_F} \varepsilon_k \, .$$

By multiplying the right-hand side of the above equation by unity of the form $\frac{1}{\delta k^3} \cdot \delta k^3 = 1$, we obtain:

$$T_k = 2\sum_{k}^{k_F} \varepsilon_k \frac{1}{\delta k^3} \cdot \delta k^3 = 2\sum_{k}^{k_F} \varepsilon_k \frac{L^3}{(2\pi)^3} \cdot \delta k^3$$

$$\Leftrightarrow T_k = 2\frac{L^3}{(2\pi)^3} \sum_{k}^{k_F} \varepsilon_k \cdot \delta k^3 \tag{2.5.7}$$

One can transform the above summation term into an integral by taking the limit of the sum as $\delta k \to 0$ or L goes to infinity[161], yielding:

$$T_k = 2\lim_{\delta k \to 0} \frac{L^3}{(2\pi)^3} \sum_{k}^{k_F} \varepsilon_k \cdot \delta k^3$$

$$\Rightarrow T_k = 2\frac{L^3}{(2\pi)^3} \int_0^{\pi} \sin\theta d\theta \int_0^{2\pi} d\phi \int_0^{k_F} k^2 \frac{\hbar^2 k^2}{2m} dk \text{ (spherical coordinates)}$$

$$\varepsilon_k$$

$$\Leftrightarrow T_k = \frac{L^3}{4\pi^3} \int_0^{k_F} 4\pi k^2 \frac{\hbar^2 k^2}{2m} dk = \frac{\hbar^2 L^3}{10\pi^2 m} k_F^5 \tag{2.5.8}$$

The kinetic energy per electron $\left(\hat{t}_k\right)$, in the ground state, is obtained by dividing the previous expression by the total number of electrons M_k within the Fermi sphere:

$$\hat{t}_k = \hat{t}_{TF} = \frac{T_k}{M_k} = \frac{\hbar^2 L^3}{10\pi^2 m} k_F^5 \cdot \frac{3\pi^2}{L^3 k_F^3}$$

$$\Leftrightarrow \hat{t}_{TF} = \frac{3\hbar^2 k_F^2}{10m}$$

One can express $\hat{t}_{TF}$ as a functional of the electron density by substituting the value of k_F from Eq. 2.5.5 into the above equation yielding:

$$\hat{t}_{TF}\left[n_k\left(\vec{r}\right)\right] = \frac{3\hbar^2\left[3\pi^2 n_k\left(\vec{r}\right)\right]^{\frac{2}{3}}}{10m}$$

$$\Leftrightarrow \hat{t}_{TF}\left[n_k\left(\vec{r}\right)\right] = \frac{3\hbar^2(3\pi^2)^{\frac{2}{3}}}{10m}\left[n_k\left(\vec{r}\right)\right]^{\frac{2}{3}} \tag{2.5.9}$$

By inserting the above equation into Eq. 2.5.3, we obtain the TF kinetic energy functional:

$$\hat{T}_{TF}[n(\vec{r})] = C_k \int n(\vec{r})^{\frac{5}{3}} d^3 r \text{, where } C_k = \frac{3\hbar^2(3\pi^2)^{\frac{2}{3}}}{10m} \tag{2.5.10}$$

Such approximation is quite decent for slow variations of the electron density in an homogeneous medium. However, it includes self-interaction and it violates the Pauli principle. In 1930, Dirac[103] formulated a local approximation to the Hartree exchange energy:

$$\hat{K}_D[n] = -C_D \int n(\vec{r})^{4/3} d\vec{r} \text{ with } C_D = \frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3} \tag{2.5.11}$$

By inserting the previously derived expressions for the kinetic energy and the expression of exchange energy in the original TF equation (Eq. 2.5.1), we obtain the TFD formulation for electrons subjected to an external potential,

$$E_{TFD}\left[n(\vec{r})\right] = C_k \int n(\vec{r})^{5/3} d\vec{r} - C_D \int n(\vec{r})^{4/3} d\vec{r} + \int u_{ext}\left(r\right) n(\vec{r}) d\vec{r} + \frac{1}{2}\iint d\vec{r} d\vec{r}' \frac{n(\vec{r}) n(\vec{r}')}{\left|\vec{r} - \vec{r}'\right|} \tag{2.5.12}$$

The ground state energy and density can be found by minimizing the functional $E_{TFD}\left[n(\vec{r})\right]$ under the condition that the total number of electrons is given by $M = \int n(\vec{r}) d\vec{r}$.

By incorporating this constraint by the method of Lagrange multipliers (β_{TFD}), the ground state density must therefore satisfy the variational principle:

$$\delta\left\{E_{TFD}\left[n\left(\vec{r}\right)\right] - \beta_{TFD}\left(\int n(\vec{r}) d\vec{r} - M\right)\right\} = 0. \tag{2.5.13}$$

Even though the TF approach provides reasonably good predictions for atoms, it leads to severe inaccuracies for the description of more complex systems since it does not encompass the true orbital structure of electrons. In fact, one notices that as we get further away from the nucleus, the charge density is represented by a power function and as we infinitesimally approach the nucleus, the charge density blows up to infinity.[97] As a result of such crude approximation, the shell structure of atoms and binding of molecules are described incorrectly.[128,129,130] In order to overcome those deficiencies, future works on improvements and modifications of the original TFD approximation have mainly been conveyed by Weisacker (1935)[131], Gross and Dreizler (1981)[132] and Perdew (1985)[9].

Countless implementations into DFT have continued for many years but it should be realized that one cannot truly provide a way to fully understand the properties of a material by simply investigating the nature of the electronic density. This therefore leads us to the Kohn-

Sham approach, in which an accurate kinetic energy functional is computed in terms of orbitals of non-interacting electrons.

2.6. The Kohn-Sham (KS) approach

By realizing the complexity of constructing the true kinetic energy of a multi-electron system, Kohn and Sham introduced the concept of an auxiliary non-interacting electrons system so that the major part of the unknown true kinetic energy can be computed with good accuracy.[5] For mathematical simplicity, in this section we will only be discussing the KS approach in non-spin-polarized system. Detailed derivation of the spin-polarized KS ansatz is given in most advanced DFT books. The KS Hamiltonian of non-interacting electrons in an unpolarized system is represented by:

$$\hat{H}_{KS} = \hat{T}_{KS} + \hat{U}_{ext} = \sum_{m=1}^{M}\left(-\frac{\nabla_m^2}{2} + u_{ext}\left(\vec{r}_m\right)\right), \tag{2.6.1}$$

where $\hat{T}_{KS}$ and $\hat{U}_{ext}$ denote the kinetic and potential operators. By subjecting the KS equation to the first HK theorem, we obtain:

$$E_{KS}\left[n(\vec{r})\right] = T_{KS}\left[n(\vec{r})\right] + \int u_{ext}(\vec{r})\, n(\vec{r})\, d\vec{r} \tag{2.6.2}$$

As discussed in section 2.2.2, the exact wave-function of non-interacting electrons is represented by a Slater determinant composed of one-electron orbitals:

$$\Psi_{KS}(\vec{r}_1, \vec{r}_2, ..., \vec{r}_M) = \frac{1}{\sqrt{M!}} \begin{vmatrix} \chi_1(\vec{r}_1) & \chi_2(\vec{r}_1) & \cdots & \chi_I(\vec{r}_1) \\ \chi_1(\vec{r}_2) & \chi_2(\vec{r}_2) & \cdots & \chi_I(\vec{r}_2) \\ \vdots & \vdots & & \vdots \\ \chi_1(\vec{r}_M) & \chi_2(\vec{r}_M) & & \chi_I(\vec{r}_M) \end{vmatrix},$$

where $\chi(\vec{r})$ is the KS one-electron orbital.

The connection between the artificial KS system and the real many-body system is the choice of an effective potential $U_{eff}(\vec{r})$ such that the ground-state density of our collection of non-interacting electrons can represent the exact ground state density of the system:

$$U_{eff}(\vec{r}) \to n_{KS}(\vec{r}) = \sum_{i=1}^{I} \left|\chi_i(\vec{r}_m)\right|^2 = n_0(\vec{r}). \tag{2.6.3}$$

By constructing non-interacting electrons with the same density as the real system, the KS orbitals would yield the exact non-interacting kinetic energy, which includes most of the true kinetic energy.[105] The residual difference between the exact independent-particle kinetic energy and the true kinetic energy can be obtained by adding to the right-hand-side of the HK functional (see Eq. 2.4.1.2) the expression: $\left\{T_{KS}\left[n_0(\vec{r})\right] - T_{KS}\left[n_0(\vec{r})\right]\right\} + \left\{U_{mn}^{Har}\left[n_0(\vec{r})\right] - U_{mn}^{Har}\left[n_0(\vec{r})\right]\right\}$ yielding:

$$\begin{aligned} E_{KS}\left[n_0(\vec{r})\right] &= T_m\left[n_0(\vec{r})\right] + U_{mn}\left[n_0(\vec{r})\right] + \int u_{ext}(\vec{r})\, n_0(\vec{r}) d\vec{r} + \\ &\quad + T_{KS}\left[n_0(\vec{r})\right] - T_{KS}\left[n_0(\vec{r})\right] + U_{mn}^{Har}\left[n_0(\vec{r})\right] - U_{mn}^{Har}\left[n_0(\vec{r})\right] \end{aligned}$$

$$\begin{aligned} \Leftrightarrow E_{KS}\left[n_0(\vec{r})\right] &= T_{KS}\left[n_0(\vec{r})\right] + \int u_{ext}(\vec{r})\, n_0(\vec{r}) d\vec{r} + U_{mn}^{Har}\left[n_0(\vec{r})\right] + \\ &\quad + \underbrace{\left\{T_m\left[n_0(\vec{r})\right] - T_{KS}\left[n_0(\vec{r})\right]\right\} + \left\{U_{mn}\left[n_0(\vec{r})\right] - U_{mn}^{Har}\left[n_0(\vec{r})\right]\right\}}_{U_{XC}}, \end{aligned} \tag{2.6.4}$$

where the exchange-correlation parameter U_{XC} describes the residual kinetic part, the self-interaction terms, the exchange components and the Coulomb correlation energy. In other words, the exchange-correlation energy corresponds to anything that is not explicitly known in the system.

In analogy to the derivation of the HF integro-differentiable equation (section 2.2.4), one can obtain the best KS orbitals by subjecting the KS Hamiltonian to linear variations of first order constrained to the orthonormality conditions of the KS orbitals:

$$\delta E_{KS}\left[\chi^*(\vec{r})\right]-\delta\sum_{j=1}^{J}\sum_{i=1}^{I}\varepsilon_{ij}\left[\left\langle\chi_i(\vec{r})\middle|\chi_j(\vec{r})\right\rangle\right]=0$$

$$\Leftrightarrow\frac{\delta E_{KS}\left[\chi^*(\vec{r})\right]}{\delta\chi_k^*(\vec{r})}-\frac{\sum_{j=1}^{J}\varepsilon_{kj}\left[\left\langle\delta\chi_k(\vec{r})\middle|\chi_j(\vec{r})\right\rangle\right]}{\delta\chi_k^*(\vec{r})}=0 \tag{2.6.5}$$

$$\Leftrightarrow\frac{\delta E_{KS}\left[n(\vec{r})\right]}{\delta n(\vec{r})}\frac{\delta n(\vec{r})}{\delta\chi_k^*(\vec{r})}-\sum_{j=1}^{J}\varepsilon_{kj}\chi_j(\vec{r})=0 \tag{2.6.6}$$

$$\text{Here }\frac{\delta n(\vec{r})}{\delta\chi_k^*(\vec{r})}=\frac{\delta\sum_{j=1}^{J}\chi_j^*(\vec{r})\chi_j(\vec{r})}{\delta\chi_k^*(\vec{r})}=\frac{\delta\chi_k^*(\vec{r})\chi_k(\vec{r})}{\delta\chi_k^*(\vec{r})}$$

$$\Rightarrow\frac{\delta n(\vec{r})}{\delta\chi_k^*(\vec{r})}=\chi_k(\vec{r})$$

By inserting the above equation into Eq. 2.6.6, we obtain:

$$\Leftrightarrow\frac{\delta E_{KS}\left[n(\vec{r})\right]}{\delta n(\vec{r})}\chi_k(\vec{r})-\sum_{j=1}^{J}\varepsilon_{kj}\chi_j(\vec{r})=0$$

$$\Leftrightarrow\frac{\delta}{\delta n(\vec{r})}\left[T_{KS}\left[n(\vec{r})\right]+\int u_{ext}(\vec{r})n(\vec{r})d\vec{r}+\frac{1}{2}\int\int\frac{n(\vec{r}')n(\vec{r})}{\left|\vec{r}-\vec{r}'\right|}d\vec{r}d\vec{r}'+U_{XC}\left[n(\vec{r})\right]\right]\chi_k(\vec{r})=\sum_{j=1}^{J}\varepsilon_{kj}\chi_j(\vec{r})$$

$$\Leftrightarrow\left\{\frac{\delta}{\delta n(\vec{r})}\left[\int\frac{-\nabla^2}{2}n(\vec{r})d\vec{r}\right]+u_{ext}(\vec{r})+\int\frac{n(\vec{r}')}{\left|\vec{r}-\vec{r}'\right|}d\vec{r}'+\frac{\delta U_{XC}\left[n(\vec{r})\right]}{\delta n(\vec{r})}\right\}\chi_k(\vec{r})=\sum_{j=1}^{J}\varepsilon_{kj}\chi_j(\vec{r})$$

$$\Leftrightarrow\left[-\frac{\nabla^2}{2}+u_{ext}(\vec{r})+\int\frac{n(\vec{r}')}{\left|\vec{r}-\vec{r}'\right|}d\vec{r}'+u_{XC}\left[n(\vec{r})\right]\right]\chi_k(\vec{r})=\sum_{j=1}^{J}\varepsilon_{kj}\chi_j(\vec{r}) \tag{2.6.7}$$

where the functional derivative of U_{XC} with respect to $n(\vec{r})$ is given by:

$$u_{XC}\left[n(\vec{r})\right]=\frac{\delta U_{XC}\left[n(\vec{r})\right]}{\delta n(\vec{r})} \tag{2.6.8}$$

Now, Eq. 2.6.7 can be written in a more compact form as:

$$\left[-\frac{\nabla^2}{2}+u_{eff}(\vec{r})\right]\chi_k(\vec{r})=\sum_{j=1}^{J}\varepsilon_{kj}\chi_j(\vec{r}) \tag{2.6.9}$$

where $u_{eff}(\vec{r})=u_{ext}(\vec{r})+\int\frac{n(\vec{r}')}{|\vec{r}-\vec{r}'|}d\vec{r}'+\frac{\delta U_{XC}[n(\vec{r})]}{\delta n(\vec{r})}$ is the KS effective potential and $u_{ext}(\vec{r})=\sum_{p=1}^{P}\frac{Z_p}{|\vec{r}_m-\vec{r}_p|}$ is the one-electron external potential.

Although the KS equations resemble the ones of HF, one needs to realize that if the explicit form of U_{XC} was to be known, the KS method would yield the exact solution of multi-electrons systems. Unlike in the HF approach, which from the beginning assumes the wave function as an antisymmetric Slater determinant, the KS method is exact in principle. The approximation in the KS formalism only comes into play when describing the exchange-correlation parameter. Unfortunately, unlike in the HF approach, the relationship between the exchange and correlation is quite intimate and it is fairly complicated to explicitly separate exchange from correlation. Now that we have briefly discussed the KS approach to multi-electrons system and the complexity into approximating the exchange-correlation parameter, we shall be investigating the physical meaning (if any) of the KS Lagrange multipliers derived in Eq. 2.6.9 in the following subsection.

2.6.1. Kohn-Sham eigenvalues and Janak's theorem

In analogy to the method used for the derivation of the HF canonical equations, by subjecting a unitary transformation to the KS one-electron Hamiltonian, we obtain the KS canonical equation:

$$\hat{f}_{KS}\chi_k(\vec{r}) = \left[-\frac{\nabla^2}{2} + u_{ext}(\vec{r}) + \int \frac{n(\vec{r}')}{|\vec{r}-\vec{r}'|} d\vec{r}' + u_{XC}[n(\vec{r})] \right] \chi_k(\vec{r}) = \varepsilon_k \chi_k(\vec{r}), \tag{2.6.1.1}$$

which must be solved iteratively as in the HF formalism. A schematic representation of the self-consistent-field KS loop is represented in Figure 2 which is based on the *non-interacting-U_{ext}-representability* assumption. Although the constructions of the HF and KS formalisms differ in principle, their similitude is quite intriguing. Analogous to the HF total energy, the KS total energy is not a simple sum of its corresponding eigenvalues since:

$$E_{KS}[n(\vec{r})] = \sum_{i=1}^{I} \varepsilon_i + \left[-\frac{1}{2} \iint d\vec{r} d\vec{r}' \frac{n(\vec{r})n(\vec{r}')}{|\vec{r}-\vec{r}'|} + U_{XC}[n(\vec{r})] - \int u_{XC}(\vec{r}) n(\vec{r}) d\vec{r} \right] \tag{2.6.1.2}$$

Here the expression of ε_i is obtained by re-writing Eq. 2.6.1.1 as:

$$\varepsilon_i = \langle \chi_i(\vec{r}) | \left[-\frac{\nabla^2}{2} + u_{ext}(\vec{r}) + \int \frac{n(\vec{r}')}{|\vec{r}-\vec{r}'|} d\vec{r}' + u_{XC}[n(\vec{r})] \right] | \chi_i(\vec{r}) \rangle \tag{2.6.1.3}$$

$$\Rightarrow \sum_{i=1}^{I} \varepsilon_i = \sum_{i=1}^{I} \Bigg[\langle \chi_i(\vec{r}) | -\frac{1}{2}\nabla^2 | \chi_i(\vec{r}) \rangle + \langle \chi_i(\vec{r}) | u_{ext}(\vec{r}) | \chi_i(\vec{r}) \rangle + \\ + \langle \chi_i(\vec{r}) | \int \frac{n(\vec{r}')}{|\vec{r}-\vec{r}'|} d\vec{r}' | \chi_i(\vec{r}) \rangle + \langle \chi_i(\vec{r}) | u_{XC}(\vec{r}) | \chi_i(\vec{r}) \rangle \Bigg]$$

$$\Leftrightarrow \sum_{i=1}^{I} \varepsilon_i = T_{KS}[n(\vec{r})] + \int u_{ext}(\vec{r}) n(\vec{r}) d\vec{r} + \int\int \frac{n(\vec{r}')n(\vec{r})}{|\vec{r}-\vec{r}'|} d\vec{r} d\vec{r}' + \int u_{XC}(\vec{r}) n(\vec{r}) d\vec{r}, \tag{2.6.1.4}$$

where $n(\vec{r}) = \sum_{i=1}^{I} \chi_i^*(\vec{r}) \chi_i(\vec{r})$

By using the same notation described in Parr and Yang[110] (pps. 164-165), let us extend the above definition of the charge density to non-integral occupation numbers *q* (see section 2.4.1.1) of the form:

$$n_J(\vec{r}) = \sum_{i=1}^{I} q_i \chi_i^*(\vec{r})\chi_i(\vec{r}) \qquad (2.6.1.5)$$

where $0 \le q_i \le 1$ and the total number of electrons *M* is defined as: $M = \sum_i^I q_i$. The fractional occupation of a KS eigenstate is physically implausible since the total number of electrons cannot vary by non-integers number but from a mathematical perspective, it can be investigated.[133] Now that we have extended our system to fractional occupational numbers, the constrained search formulation discussed in section 2.4.2 searches over all the ensemble of statistical mixtures (with non-integer electron values) yielding the best density $n(\vec{r})$, and ultimately providing the minimum expectation value of the KS total energy $\left\langle E_{KS}\left[n(\vec{r})\right]\right\rangle$. Such generalization is introduced by Janak[134] (*J*) in which the KS total energy from Eq. 2.6.4 is re-written as $E_J\left[n(\vec{r})\right]$:

$$E_{HK}\left[n(\vec{r})\right]\Big|_J = E_J\left[n_J(\vec{r})\right] = T_{KS}\left[n_J(\vec{r})\right] + \int u_{ext}(\vec{r})\, n_J(\vec{r})\, d\vec{r} + U_{mn}^{Har}\left[n_J(\vec{r})\right] + \underbrace{+\left\{T_{HK}\left[n_J(\vec{r})\right] - T_{KS}\left[n_J(\vec{r})\right]\right\} + \left\{U_{mn}\left[n_J(\vec{r})\right] - U_{mn}^{Har}\left[n_J(\vec{r})\right]\right\}}_{U_{XC}\left[n_J(\vec{r})\right]}$$

Minimization of the Janak's functional $E_J\left[n_J(\vec{r})\right]$ with respect to undetermined orbitals and under the normalization condition $\int \chi_i^*(\vec{r})\chi_i(\vec{r})d\vec{r} = 1$ (*orthogonality is not required due to hermiticity of* $E_J\left[n_J(\vec{r})\right]$)[96] yields:

$$\Leftrightarrow \frac{\delta E_J\left[\chi^*(\vec{r})\right]}{\delta\chi_k^*(\vec{r})} - \frac{\delta}{\delta\chi_k^*(\vec{r})}\sum_{j=1}^{J}\varepsilon_j\left[\left\langle\chi_j(\vec{r})\middle|\chi_j(\vec{r})\right\rangle - 1\right] = 0 \qquad (2.6.1.6)$$

$$\Leftrightarrow \frac{\delta E_J\left[n_J(\vec{r})\right]}{\delta n_J(\vec{r})}\frac{\delta n_J(\vec{r})}{\delta\chi_k^*(\vec{r})} - \frac{\varepsilon_k\left[\left\langle\delta\chi_k(\vec{r})\middle|\chi_k(\vec{r})\right\rangle\right]}{\delta\chi_k^*(\vec{r})} = 0$$

$$\Leftrightarrow \frac{\delta E_J\left[n_J(\vec{r})\right]}{\delta n_J(\vec{r})}\frac{\delta n_J(\vec{r})}{\delta\chi_k^*(\vec{r})} - \varepsilon_k\chi_k(\vec{r}) = 0 \qquad (2.6.1.7)$$

Here $\frac{\delta n_J(\vec{r})}{\delta\chi_k^*(\vec{r})} = \frac{\delta\sum_{j=1}^{J} q_i\chi_j^*(\vec{r})\chi_j(\vec{r})}{\delta\chi_k^*(\vec{r})} = \frac{q_k\delta\chi_k^*(\vec{r})\chi_k(\vec{r})}{\delta\chi_k^*(\vec{r})}$

$$\Rightarrow \frac{\delta n_J(\vec{r})}{\delta\chi_k^*(\vec{r})} = q_k\chi_k(\vec{r}) \qquad (2.6.1.8)$$

By substituting the above expression into Eq. 2.6.1.7, we have:

$$\frac{\delta E_J\left[n_J(\vec{r})\right]}{\delta n_J(\vec{r})}q_k\chi_k(\vec{r}) - \varepsilon_k\chi_k(\vec{r}) = 0,$$

which means that:

$$\frac{\delta}{\delta n_J(\vec{r})}\left[T_{KS}\left[n_J(\vec{r})\right] + \int u_{ext}(\vec{r})n_J(\vec{r})d\vec{r} + \frac{1}{2}\int\int\frac{n_J(\vec{r}')n_J(\vec{r})}{|\vec{r}-\vec{r}'|}d\vec{r}d\vec{r}' + U_{XC}\left[n_J(\vec{r})\right]\right]q_k\chi_k(\vec{r}) = \varepsilon_k\chi_k(\vec{r})$$

$$\Leftrightarrow \left\{\frac{\delta}{\delta n_J(\vec{r})}\left[\int\frac{-\nabla^2}{2}n_J(\vec{r})d\vec{r}\right] + u_{ext}(\vec{r}) + \int\frac{n(\vec{r}')}{|\vec{r}-\vec{r}'|}d\vec{r}' + \frac{\delta U_{XC}\left[n(\vec{r})\right]}{\delta n(\vec{r})}\right\}q_k\chi_k(\vec{r}) = \varepsilon_k\chi_k(\vec{r})$$

$$\Leftrightarrow \left[\frac{-\nabla^2}{2} + u_{ext}(\vec{r}) + \int\frac{n_J(\vec{r}')}{|\vec{r}-\vec{r}'|}d\vec{r}' + \frac{\delta U_{XC}\left[n_J(\vec{r})\right]}{\delta n_J(\vec{r})}\right]q_k\chi_k(\vec{r}) = \varepsilon_k\chi_k(\vec{r})$$

$$\Leftrightarrow \left[\frac{-\nabla^2}{2} + u_{ext}(\vec{r}) + \int\frac{n_J(\vec{r}')}{|\vec{r}-\vec{r}'|}d\vec{r}' + \frac{\delta U_{XC}\left[n_J(\vec{r})\right]}{\delta n_J(\vec{r})}\right]\chi_k(\vec{r}) = \frac{\varepsilon_k}{q_k}\chi_k(\vec{r})$$

$$\Leftrightarrow \left[\frac{-\nabla^2}{2} + u_{ext}(\vec{r}) + \int\frac{n_J(\vec{r}')}{|\vec{r}-\vec{r}'|}d\vec{r}' + \frac{\delta U_{XC}\left[n_J(\vec{r})\right]}{\delta n_J(\vec{r})}\right]\chi_k(\vec{r}) = \varepsilon_J\chi_k(\vec{r}) \qquad (2.6.1.9)$$

where $\varepsilon_J = \frac{\varepsilon_k}{q_k}$. The above equation is identical to the KS canonical equations given that $n_J(\vec{r}) = \sum_{i=1}^{I} q_i \chi_i^*(\vec{r})\chi_i(\vec{r})$. Let us evaluate the partial derivative $\left(\frac{\partial E_J\left[n_J(\vec{r},q_k)\right]}{\partial q_k}\right)$ of Janak's total energy functional $E_J\left[n_J(\vec{r},q_k)\right]$ with respect to occupation number q_k. Since q_k is a constant, we follow the chain rule described in Eq. 2.2.4 and the expression $\frac{\partial E_J\left[n_J(\vec{r},q_k)\right]}{\partial q_k}$ can hence be evaluated as:

$$\frac{\partial E_J\left[n_J(\vec{r},q_k)\right]}{\partial q_k} = \int \frac{\delta E_J\left[n_J(\vec{r})\right]}{\delta n_J(\vec{r})} \frac{\partial n_J(\vec{r})}{\partial q_k} d\vec{r}$$

Then, $$\frac{\partial E_J\left[n_J(\vec{r},q_k)\right]}{\partial q_k} = \int \left[\frac{-\nabla^2}{2} + u_J^{eff}(\vec{r})\right] \frac{\partial n_J(\vec{r})}{\partial q_k} d\vec{r} \qquad (2.6.1.10)$$

where $u_J^{eff}(\vec{r}) = \left[u_{ext}(\vec{r}) + \int \frac{n_J(\vec{r}')}{|\vec{r}-\vec{r}'|} d\vec{r}' + \frac{\delta U_{XC}\left[n_J(\vec{r})\right]}{\delta n_J(\vec{r})}\right]$.

Now let us simplify $\frac{\partial n_J(\vec{r})}{\partial q_k}$ from Eq. 2.6.1.10:

$$\frac{\partial n_J(\vec{r})}{\partial q_k} = \frac{\partial \sum_{i=1}^{I} q_i \chi_i^*(\vec{r})\chi_i(\vec{r})}{\partial q_k} = \chi_k^*(\vec{r})\chi_k(\vec{r}).$$

By inserting the above expression into Eq. 2.6.1.10, we obtain the following expression:

$$\frac{\partial E_J\left[n_J(\vec{r},q_k)\right]}{\partial q_k} = \left\langle \chi_k(\vec{r}) \middle| \frac{-\nabla^2}{2} + u_J^{eff}(\vec{r}) \middle| \chi_k(\vec{r}) \right\rangle,$$

The right-hand side of the above equation was already defined in Eq. 2.6.1.3 as the KS eigenvalue. Then:

$$\frac{\partial E_J\left[n_J\left(\vec{r},q_k\right)\right]}{\partial q_k}=\varepsilon_k \tag{2.6.1.11}$$

The above equation is Janak's theorem[134] which states that the partial derivative of the total energy with respect to fractional occupational numbers corresponds to the *k*-th KS eigenvalue. From the Fundamental Theorem of Calculus, the above equation can be rewritten in an integral form:

$$\int_{M-1}^{M}\frac{\partial E}{\partial q_k}dq\Bigg|_{n_J(\vec{r})}=E\left[M\right]-E\left[M-1\right]=\int_{M-1}^{M}\varepsilon_k dq$$

$$\Leftrightarrow E\left[M\right]-E\left[M-1\right]=\int_0^1\varepsilon_I dq$$

$$\Leftrightarrow -IP=\int_0^1\varepsilon_I dq, \tag{2.6.1.12}$$

where *IP* is the ionization potential. Here ε_I denotes the eigenvalue of the highest occupied molecular orbital (HOMO) and for that reason Janak's theorem can only be applied to the addition or subtraction of an electron to the highest occupied orbital.[135] This is in contrast to Koopmans' theorem in the Hartree-Fock formalism, where electrons can technically be added/subtracted to any state (although the Koopman's approximation worsens as one further adds/removes electrons from occupied orbitals). Furthermore, correlation and wave function relaxation (used within the optimized effective potential[136]) are taken into consideration into the derivation of the *IP* in the KS ansatz, given the exact form of the exchange-correlation parameter is known.

However, the integral derived in Eq. 2.6.1.12 is complex in a multi-electron system but can be approximated using the trapezoid formula:

$$\Leftrightarrow -IP = \int_0^1 \varepsilon_I dq \approx \frac{\varepsilon_I[M] + \varepsilon_I[M+1]}{2}, \tag{2.6.1.13}$$

which corresponds to the average value between the HOMOs of the neutral system and the ionized system. The same procedure can be applied into the calculation of the electron affinity yielding:

$$-A = \int_0^1 \varepsilon_{I+1} dq, \tag{2.6.1.14}$$

where ε_{I+1} corresponds to the lowest unoccupied molecular orbital (LUMO). Recall from section 2.4.2.1 that the chemical potential in the HK equations is subjected to a derivative discontinuity of the form:

$$\lim_{\eta\to 0}\mu\big|_{N+\eta} = \lim_{\eta\to 0}\frac{\partial E[N]}{\partial N}\bigg|_{N+\eta} = \frac{\delta E_{HK}[n(\vec{r})]}{\delta n(\vec{r})}\bigg|_{N+\eta} = E[M] - E[M-1] = -IP,$$

$$\lim_{\eta\to 0}\mu\big|_{N-\eta} = \lim_{\eta\to 0}\frac{\partial E[N]}{\partial N}\bigg|_{N-\eta} = \frac{\delta E_{HK}[n(\vec{r})]}{\delta n(\vec{r})}\bigg|_{N-\eta} = E[M+1] - E[M] = -A,$$

where N is an average total number of electrons that includes non-integers values. By comparing the above equations to the Janak's definition of the KS eigenvalues, we notice that the highest occupied KS eigenvalue and the chemical potential are related:

$$\begin{cases} Vacuum - \varepsilon_I = \lim_{\eta\to 0}\mu\big|_{N+\eta} = E[M] - E[M-1] = -IP & (M-1 < N < M) \quad (2.6.1.15) \\ Vacuum - \varepsilon_{I+1} = \lim_{\eta\to 0}\mu\big|_{N-\eta} = E[M+1] - E[M] = -A & (M < N < M+1) \quad (2.6.1.16) \end{cases}$$

where M corresponds to the integer value of the total number of electrons.

Let us further discuss the physical meaning of the above equation by first recalling the definition of the HK bandgap:

$$E_g^{HK} = IP - A = \left.\frac{\delta E_{HK}\left[n(\vec{r})\right]}{\delta n(\vec{r})}\right|_{N+\eta} - \left.\frac{\delta E_{HK}\left[n(\vec{r})\right]}{\delta n(\vec{r})}\right|_{N-\eta},$$

and the expression of the KS canonical equations:

$$\left[\frac{\delta T_{KS}\left[n_J(\vec{r})\right]}{\delta n(\vec{r})} + u_{ext}(\vec{r}) + \int \frac{n(\vec{r}')}{\left|\vec{r}-\vec{r}'\right|} d\vec{r}' + \frac{\delta U_{XC}\left[n(\vec{r})\right]}{\delta n(\vec{r})}\right] \chi_k(\vec{r}) = \varepsilon_k \chi_k(\vec{r}).$$

We notice from the above equation that the only components that involve variations of the charge density are the KS non-interacting kinetic energy and the exchange-correlation parameter. By inserting the variations of T_{KS} and the XC-parameter into the definition of the HK gap, we obtain the KS gap:

$$E_g^{KS} = IP - A = \left[\left.\frac{\delta T_{KS}\left[n_J(\vec{r})\right]}{\delta n_J(\vec{r})}\right|_{N+\eta} - \left.\frac{\delta T_{KS}\left[n_J(\vec{r})\right]}{\delta n_J(\vec{r})}\right|_{N-\eta}\right] + \left[\left.\frac{\delta U_{XC}\left[n_J(\vec{r})\right]}{\delta n_J(\vec{r})}\right|_{N+\eta} - \left.\frac{\delta U_{XC}\left[n_J(\vec{r})\right]}{\delta n_J(\vec{r})}\right|_{N-\eta}\right],$$

(2.6.1.15)

which means that the band structure computed from the Kohn Sham ansatz underestimates the actual gap width by an amount equal to the derivative discontinuity of the KS non-interacting kinetic energy and the XC-energy.[137] In other words, the difference in energy between the KS HOMO and LUMO eigenvalues is unphysical and does not describe the actual band gap of a given system.

Nevertheless, although the KS HOMO-LUMO gap is incorrect, the correlation between the *IP/A* and the KS eigenvalues is of paramount importance. For instance, in intrinsic semiconductors, the previously discussed HOMO is physically equivalent to the valence band maximum (VBM) while the LUMO denotes the conduction band minimum (CBM). If one is

interested in calculating the thermodynamic transition levels (cf. section 4.4.1) or optics (see section 4.4.2) of defects in the gap, the exact positions of both CBM or VBM are not required. In fact, from Eq. 2.6.1.15 or Eq. 2.6.1.17, we can define the CBM/VBM positions with respect to vacuum level and hence circumvent the problem of derivative discontinuity (intrinsic to the DFT approach) or any other problems related to the calculation of the exact bandgap of a many-body system.

In summary, we have seen that in the KS method, the multi-electrons interacting system is mapped into an auxiliary non-interacting system and their difference in energy is encompassed in the exchange-correlation functional. An explicit form of the exchange-correlation parameter is required if one needs to solve the canonical KS equations and ultimately obtain the exact solution of complex multi-electrons systems. The quest into obtaining accurate exchange-correlation parameter is at the heart of modern DFT and remains the greatest task for the application of DFT to many-body systems. In the next section, we described a simple, yet not too crude approximation of the exchange-correlation term that was initially proposed by Kohn and Sham.

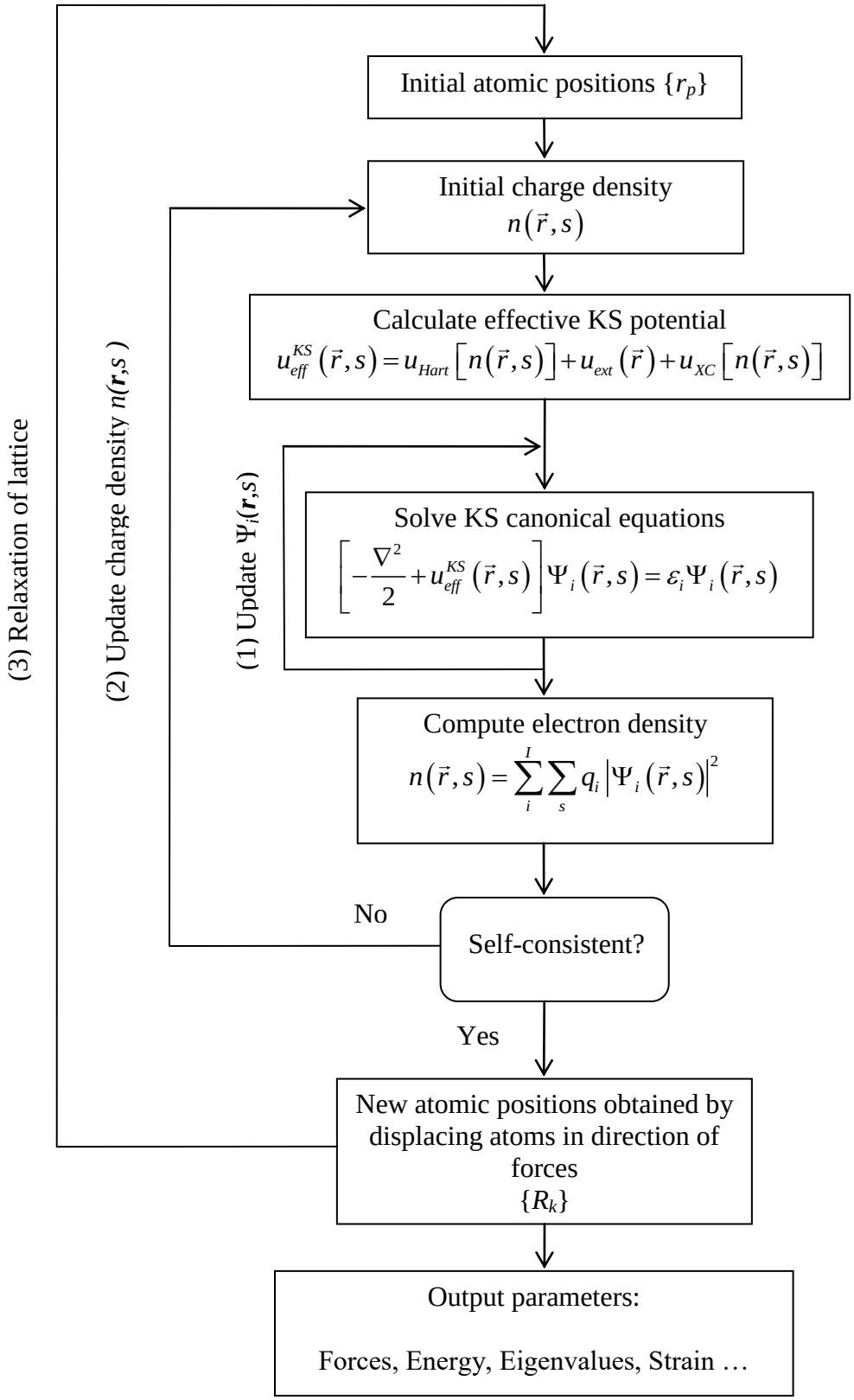

Figure 2: Schematic representation of the self-consistent loop in real space where the charge density $n(\vec{r},s)$ and wave function $\Psi_i(\vec{r},s)$ are spin-dependent. The first (1) and second (2) loop *must be iterated simultaneously for both spins*.[7]

2.6.2. Local Density Approximation (LDA) of the Exchange and Correlation Energy

In the LDA method, the exchange-correlation energy can be approximated by treating the exchange-correlation term as a local functional in the uniform-electron gas model:

$$E_{XC}^{LDA} = \int u_{XC}^{LDA}\left[n(\vec{r})\right] \cdot n(\vec{r}) d\vec{r}, \tag{2.6.2.1}$$

where $u_{XC}^{LDA}\left[n(\vec{r})\right]$ designates the exchange-correlation energy per electron at $\vec{r}$ in a uniform electron gas of density $n(\vec{r})$. For small perturbations of the electronic density in the form of $n(\vec{r}) \to n(\vec{r}) + \delta n(\vec{r})$, linear variation of the corresponding exchange-correlation potential is written as:

$$\begin{aligned}
\delta E_{XC}^{LDA} &= \delta \int u_{XC}^{LDA}\left[n(\vec{r})\right] \cdot n(\vec{r}) d\vec{r} \\
&= \int \delta u_{XC}^{LDA}\left[n(\vec{r})\right] \cdot n(\vec{r}) d\vec{r} + \int u_{XC}^{LDA}\left[n(\vec{r})\right] \cdot \delta n(\vec{r}) d\vec{r} \\
&= \int \delta u_{XC}^{LDA}\left[n(\vec{r})\right] \cdot n(\vec{r}) \frac{\delta n(\vec{r})}{\delta n(\vec{r})} d\vec{r} + \int u_{XC}^{LDA}\left[n(\vec{r})\right] \cdot \delta n(\vec{r}) d\vec{r} \\
&= \int \frac{\delta u_{XC}^{LDA}\left[n(\vec{r})\right]}{\delta n(\vec{r})} \cdot n(\vec{r}) \delta n(\vec{r}) d\vec{r} + \int u_{XC}^{LDA}\left[n(\vec{r})\right] \cdot \delta n(\vec{r}) d\vec{r}
\end{aligned}$$

From the definition of functional derivatives described in Eq. 2.3.3, the above equation becomes:

$$\Leftrightarrow \frac{\delta E_{XC}^{LDA}}{\delta n(\vec{r})} = \frac{\delta u_{XC}^{LDA}\left[n(\vec{r})\right]}{\delta n(\vec{r})} \cdot n(\vec{r}) + u_{XC}^{LDA}\left[n(\vec{r})\right] \tag{2.6.2.2}$$

In the homogeneous gas model, the second term $u_{XC}^{LDA}\left[n(\vec{r})\right]$ in the right-hand side of the above equation can be divided into exchange and correlation components:

$$u_{XC}^{LDA}\left[n(\vec{r})\right] = u_{X}^{LDA}\left[n(\vec{r})\right] + u_{C}^{LDA}\left[n(\vec{r})\right] \tag{2.6.2.3}$$

Hence Eq. 2.6.2.1 becomes:

$$E_{XC}^{LDA} = E_{X}^{LDA} + E_{C}^{LDA} = \int u_{X}^{LDA}\left[n(\vec{r})\right] \cdot n(\vec{r}) d\vec{r} + \int u_{C}^{LDA}\left[n(\vec{r})\right] \cdot n(\vec{r}) d\vec{r}, \tag{2.6.2.4}$$

where $E_X^{LDA} = \int u_X^{LDA}\left[n(\vec{r})\right] \cdot n(\vec{r}) d\vec{r}$, (2.6.2.4.a)

and $E_C^{LDA} = \int u_C^{LDA}\left[n(\vec{r})\right] \cdot n(\vec{r}) d\vec{r}$. (2.6.2.4.b)

By comparing the Dirac exchange energy expressed in Eq. 2.6.12 with the previous equation, the exchange LDA potential becomes:

$$u_X^{LDA} = -C_D n(\vec{r})^{1/3} \text{ where } C_D = \frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3} \quad (2.6.2.5)$$

Unfortunately, the correlation parameter is far more complicated and can only be calculated analytically in the high density and the low density limits.[133] The correlation energy can be vaguely conceptualized as the energy that is created from the correlated motion of the electrons due to electrostatic repulsion and attraction to the compensated positively charged background. Over the years, various approximations of the correlation energy have been made and one of the most accurate calculations was the quantum Monte Carlo computations for uniform gas.[138]

By looking back at the derivation of the exchange and correlation potential described in Eq. (2.6.2.2), we define the first term $\frac{\delta u_{XC}^{LDA}\left[n(\vec{r})\right]}{\delta n(\vec{r})} \cdot n(\vec{r})$ as the "response function". This "response function" is explicitly defined by Gritsenko et al. (1994)[139] as the response potential of the exchange correlation hole subjected to perturbation of the density. In other words, the $u_{XC}(\vec{r})$ term can be pictured as if the electron *creates a hole around itself* and interacts with the exchange-correlation hole via electrostatic Coulombic repulsion.[140] More details regarding the physical meaning of the response function can be found in Refs. [141,142].

After locally approximating the exchange-correlation potential, the KS orbital equations can be rewritten as:

$$\left[-\frac{\nabla^2}{2}+u_{ext}\left(\vec{r}\right)+\int\frac{n(\vec{r}')}{\left|\vec{r}-\vec{r}'\right|}d\vec{r}'+u_{XC}^{LDA}\left[n(\vec{r})\right]\right]\chi_k\left(\vec{r}\right)=\varepsilon_k\chi_k(\vec{r}), \tag{2.6.2.6}$$

which are solved in a self-consistent manner.

Before further introducing any approximation to the exchange-correlation parameter, we first recall that the KS approach has been restricted so far to non-spin-polarized systems. By adding a magnetic field to the usual scalar external potential acting on the many-electrons systems, we are building more physical insight into the approximation of the exchange correlation functional. The next two sections are devoted to describing relevant functionals that approximate the spin dependent exchange correlation parameter.

2.6.3. Local Spin Density Functional (LSDA)

By inserting spin dependence into the homogenous exchange correlation parameter from the LDA formalism, we obtain:

$$E_{XC}^{LSDA} = E_X^{LSDA} + E_C^{LSDA} = \int u_X \left[n_\alpha(\vec{r}), n_\beta(\vec{r}) \right] \cdot n(\vec{r}, s) d\vec{r} + \int u_C \left[n_\alpha(\vec{r}), n_\beta(\vec{r}) \right] \cdot n(\vec{r}, s) d\vec{r} \quad (2.6.3.1)$$

The LSDA can be described in terms of the up and down spin densities, but it is usually expressed in terms of the total charge density $n_\alpha(\vec{r}) + n_\beta(\vec{r})$ and the local relative spin polarization:

$$\zeta(\vec{r}) = \frac{n_\alpha(\vec{r}) - n_\beta(\vec{r})}{n_\alpha(\vec{r}) + n_\beta(\vec{r})}. \quad (2.6.3.2)$$

For the far more sophisticated case of spin dependent correlation energy, there have been some serious attempts to parameterize the uniform gas correlation energy as a function of spin polarization, $u_C(n, \zeta)$.[143,144] Even though the LSDA principles were funded on the basis of a homogenous gas approximation, its success for very inhomogeneous cases are not to be overlooked. One reason for such success might be the cancellation of errors between exchange and correlation. In 1966, Tong and Sham[145] noticed that in LSDA calculations, the total exchange energy is typically underestimated by about 10% while the correlation energy is overestimated by a factor of two or more. Since for many physical systems, the exchange energy is about four times greater than correlation,[105] the overestimation of the correlation energy greatly cancels the underestimation of the exchange energy. Due to the partial cancellation of errors, the LDSA gives excellent approximations of bond lengths, ionization and binding and dissociation energies. Unfortunately, LSDA tends to fail into describing weak bondings, systems with slowly varying densities, correct band gaps and magnetism of transition metals. To improve

upon the local spin density formalism, one should take into account the problem of the unphysical self-interaction term in the approximation of the exchange and correlation parameter. In order to remove those spurious self-interaction terms, Perdew and Zunger (1981)[146] suggested a self- interaction correction (SIC) to LSDA. In such corrections, the newly obtained Kohn Sham orbitals vary for different potentials hence causing non-orthogonality of the orbitals. Further physical insight into the non-orthogonality of the orbitals is given in the original review.[146]

Even after improving the LSDA, one observes that the exchange correlation potential has a quite short range and it only depends on the local density. Consequently, the LSDA potential has the wrong asymptotic dependence as r tends to infinity and one might consider a functional that depends not only on the electronic density at a specific point but also depends on the magnitude of the gradient of the density that would describe the inhomogeneity of the true electron density.

2.6.4. Generalized Gradient Approximations (GGA)

The idea of constructing a functional that is not as computationally demanding as non-local functionals (due to HF exchange) and yet possesses a potential that does not diverge for exponentially decaying densities would be groundbreaking. One of the first suggestions into expanding the exchange correlation parameter in function of the magnitude of the gradient of the density was given in the original paper of Hohenberg and Kohn[107] and it was referred to as a gradient expansion approximation (GEA). The insertion of gradient-dependent functionals into the local approximation of the exchange correlation parameter was first thought to be very attractive for applications but unfortunately most gradient expansion approximations did not lead to consistent improvements over the LSDA. In fact, the major drawbacks of the GEA were that its corresponding exchange correlation hole was not physical, nor did it satisfy the normalization conditions of the exchange and correlation holes and the *negativity state of the exchange hole.*[9,147] By eliminating the spurious long-range term of the second order expansion of the GEA exchange-correlation hole, generalized gradient approximations[148,149,150] (GGAs) were created and their corresponding exchange-correlation energy were expressed as:

$$E_{XC}^{GGA}\left[n_\alpha(\vec{r}),n_\beta(\vec{r})\right]=\int d\vec{r}n(\vec{r})\cdot u_{XC}^{LSDA}\left[n(\vec{r})\right]\cdot F_{XC}\left[n_\alpha(\vec{r}),n_\beta(\vec{r}),|\nabla n_\alpha|,|\nabla n_\beta|\right] \qquad (2.6.4.1)$$

where F_{XC} is called the enhancement factor which adjusts the XC-parameter from the homogeneous gas model (LSDA) based on variations of the electronic density in the neighborhood of the considered region in space.[151]

Unlike its GEA predecessors, GGA functionals arise from the second order gradient expansion and can be derived with two different approaches:

- Method 1: empirical fitting.

Here the second order expansion is obtained by fitting as much experimental data as possible.[149] As a general trend, such GGAs yield more accurate atomization energies[152,153] and barriers to chemical reactions,[154] compared to LSDA when applied to molecular systems, and provides mixed results when applied to solids. As a result, an entire zoo of GGA functionals emerged in the early eighties which forced most computational physicists/chemists to use large sets of benchmark systems in order to find the optimum GGA that would yield the most accurate data.

- Method 2: mathematical construct of the enhancement factor F_{XC}.

In order to distinguish a GGA that would offer a consistent improvement over the previous LSDA, several authors[155] decided to plot the GGA dimensionless enhancement factor $F_{XC}\left[s(\vec{r}), r_s(\vec{r})\right]$ in function of reduced density gradient:

$$s(\vec{r}) = \frac{|\nabla n(\vec{r})|}{2k_F n(\vec{r})},$$

where $k_F = \left(3\pi^2 n(\vec{r})\right)^{1/3}$ is the local Fermi wavevector, for various values of Wigner-Seitz radius $r_s(\vec{r}) = \left(\frac{3}{4\pi n(\vec{r})}\right)^{1/3}$. Here the Wigner-Seitz radius denotes the radius of a sphere for which its volume *is computed as the expected value of the volume per atom in solids.*[156]

By using this method, the GGA functional is built from a mathematical construct of the exchange-correlation parameter that satisfies as many of the known properties of the exact XC-hole.

Such GGA construction was met with success since it improved upon the predictions of: [157]

a) electron affinities and first ionization potentials of several atoms

b) atomization energies of various hydrocarbon molecules

c) lattice constants and bulk moduli of Li and Na that used to be slightly underestimated by the LSDA

d) the ferromagnetic configuration of bcc ground state of metallic iron.

The culmination of the complex derivation of GGA functionals is arguably reached with the Perdew-Burke-Enzherof (PBE)-GGA functional[14] since it succeeded into including a relatively simple mathematical construct of a gradient correction without introducing experimentally fitted parameters. Here the exchange functional was rewritten as:

$$E_X^{PBE}\left[n(\vec{r})\right] = \int d\vec{r} u_X^{LDA}\left[n(\vec{r})\right] \cdot F_X\left[s(\vec{r}), r_s(\vec{r})\right] \qquad (2.6.4.2)$$

where $F_X\left[s(\vec{r}), r_s(\vec{r})\right] = \kappa - \frac{3\kappa^2}{3\kappa + \beta\pi^2\left[s(\vec{r})\right]^2} + 1, \kappa = 0.804$ and $\beta = 0.066725$. Here the values of κ and β are derived in such way that the Lieb-Oxford inequality[158] is satisfied and *response function of the LDA exchange is restored*.[7] A detailed description of the correlation term and the physical meaning of every single component the PBE exchange term are given in the original review.[14] The PBE is commonly known for its applicability to both homogenous and inhomogenous systems and tends to accurately predict various electronic properties in a wide range of complex systems. Although the PBE-GGA predicts several correct physical properties, it highly underestimates the values of bandgaps in semiconductors/insulators and tends to fail to accurately describe weak Van der Waals interactions. A way forward into improving the shortcomings of the PBE functional might be a density functional theory in which there would be a combination of the HF exact exchange functional which would partially cancel fictitious self-

interaction with a GGA semi-local functional that would accurately describe the correlation component. In the next section, we will focus on giving a brief overview of this new class of mixed functionals called “hybrid functionals”.

2.7. Hybrid Functionals

In the last decades, wave-functions-based methods and DFT have proven to be very powerful tools in describing various properties in a wide range of materials. Nevertheless, the explicit derivation of the exchange correlation parameter in DFT and the computational cost of post-HF methods are still methodological barriers that have not been overcome. First-principles method based on DFT such as LSDA has been quite successful, especially for those where the electronic density is quite uniform. In order to address the main limitations of LSDA, an expansion of the density in terms of the gradient and higher order derivatives has been carried out. Unfortunately, the improvement of GGAs over LSDA is not substantial since GGAs still underestimate the band gap in many systems and do not satisfy known asymptotic behaviors for isolated atoms. The main reasons of such limitations is probably due to the fact that self-interaction are still present in the Hartree term, the non-locality of the exchange component is not fully taken into account and the complexity of computing the correlation term. The next step beyond first-principles methods based on DFT might be the introduction of so-called hybrid functionals which are obtained by an admixture of a non-local fixed amount of Fock exchange to GGA-type functionals. The theoretical justification behind this approach is explained via the adiabatic connection, which is briefly discussed in the next section.

2.7.1. Adiabatic approach

Let us first recall the expression of the HK universal functional of a many-body system:

$$\hat{F}_{HK} = \hat{T}_m\left[n(\vec{r})\right] + \hat{U}_{mn}\left[n(\vec{r})\right],$$

where each component was described in section 2.4. Let us define a switching parameter λ which describes the strength of the electron-electron interactions and can vary from 0 to 1. One can connect the above fully interacting Hamiltonian to the KS non-interacting system by steadily increasing the coupling parameter λ via the *adiabatic* equation:[110]

$$\hat{F}_{HK}^{\lambda} = \hat{T}_m\left[n^{\lambda}(\vec{r})\right] + \lambda\hat{U}_{mn}\left[n^{\lambda}(\vec{r})\right] \tag{2.7.1.1}$$

Here for $\lambda = 0$, the above system becomes the Hamiltonian of the KS non-interacting system:

$$\hat{F}_{HK}^{\lambda=0} = \hat{T}_m\left[n^{\lambda=0}(\vec{r})\right] = \hat{T}_{KS}\left[n(\vec{r})\right], \tag{2.7.1.2}$$

while for $\lambda = 1$, the full electron-electron interaction is recovered and we obtain the $\hat{F}_{HK}$ universal functional. Although the parameter switches gradually from 0 to 1, the electronic density remains constant throughout the process. By subtracting the universal functional energy of the fully interacting system from the energy of the KS auxiliary system we obtain:

$$\begin{aligned} F^{\lambda=1} - F^{\lambda=0} &= \hat{T}_m\left[n(\vec{r})\right] + \hat{U}_{mn}\left[n(\vec{r})\right] - \hat{T}_{KS}\left[n(\vec{r})\right], \\ &= \hat{T}_m\left[n(\vec{r})\right] + \hat{U}_{mn}\left[n(\vec{r})\right] - \hat{T}_{KS}\left[n(\vec{r})\right] + \left\{\hat{U}_{mn}^{Har}\left[n(\vec{r})\right] - \hat{U}_{mn}^{Har}\left[n(\vec{r})\right]\right\} \\ &= \underbrace{\hat{T}_m\left[n(\vec{r})\right] - \hat{T}_{KS}\left[n(\vec{r})\right] + \hat{U}_{mn}\left[n(\vec{r})\right] - \hat{U}_{mn}^{Har}\left[n(\vec{r})\right]}_{E_{XC}\left[n(\vec{r})\right]} + \hat{U}_{mn}^{Har}\left[n(\vec{r})\right] \end{aligned}$$

$$\Leftrightarrow \int_0^1 dF^{\lambda} = E_{XC}\left[n(\vec{r})\right] + \hat{U}_{mn}^{Har}\left[n(\vec{r})\right]. \tag{2.7.1.3}$$

To exactly solve the above equation, one requires the value of the exchange correlation parameter at intermediate values of λ. However such information is unobtainable and one can

approximate the above integral by mixing a certain fraction of the pure HF exact exchange with local GGA/LDA XC-hole.

Starting from this approach, several hybrid formalisms have been constructed but the pioneering work in hybrid schemes construction was proposed by Becke[22] where he suggested a single coefficient a_1 to be shared between the HF and GGA exchange energy:

$$E_{XC}^{Becke} = a_1\left(E_X^{GGA} - E_X^{HF}\right) + E_{XC}^{GGA} \tag{2.7.1.4}$$

Recent work done in Ref. [23] has shown that by using perturbation theory, one can obtain the optimum amount of exact exchange $a_1 \approx 0.25$ to be admixed with a density functional approximation. As a result, the parameter free hybrid scheme Perdew-Burke-Ernzherof-zero-parameter (PBE0)[159,160] was constructed and so far seems to correctly predict band gaps, bulk moduli, lattice constants and thermochemical[22] properties of many complex systems. Nevertheless, the computational cost of calculating the exact Fock exchange in systems subjected to periodic boundary conditions (solids), made the use of PBE0 problematic. In addition to the intractability of the Fock exchange parameter, one should notice in metallic systems, the *divergence*[96] *of the partial derivative* of *the single particle Hartree-Fock eigenvalue* $\varepsilon\left(\vec{k}\right)$, *with respect to the crystal momentum k at the Fermi level* k_F, where Aschcroft and Mermin (1976, Chap. 17)[161] defined the single particle HF energy $\varepsilon\left(\vec{k}\right)$ as:

$$\varepsilon\left(\vec{k}\right) = \frac{\hbar^2 k^2}{2m} - 2\frac{e^2}{\pi}k_F F\left(\frac{k}{k_F}\right) \tag{2.7.1.5}$$

Here $F\left(\frac{k}{k_F}\right) = \frac{1}{2} + \frac{{k_F}^2 - k^2}{4kk_F}\ln\left|\frac{k_F + k}{k_F - k}\right|$ (2.7.1.6)

Such logarithmic divergence can be explained by *the divergence of the Fourier transform* $\frac{4\pi e^2}{k^2}$ *of the Coulomb interaction* $\frac{e^2}{r}$ *at* $k = 0$.[161]

In order to make the Fock exchange parameter tractable in periodic systems and circumvent possible divergence of the exact exchange term in metallic systems, Heyd (2006)[25] argued that the spatial decay of the exchange interactions must be either be accelerated or one might artificially cut off or truncate part of the exchange interactions. The first alternative, which is the acceleration of spatial decay, might well predict the total energy of the system but would still neglect long range exchange-correlation. The second approach, truncating the exchange interactions are known to work quite well in localized systems. Nevertheless, in case of delocalized charge distribution where the HF exchange does not rapidly decay over distance, such truncation methods appear completely unphysical and create some severe convergence problems in the self-consistent-field process. Hence, in order to accelerate the HF decay without neglecting long range interactions, the 1/*r* part of the exchange interaction can be replaced with a Coulomb screened potential.

By screening the 1/*r* part of the exchange interaction, we obtain a potential that has a shorter range than 1/*r* and we can therefore diminish the computational complexity of the Fock exchange and eliminate the unphysical singularity of the anomalous divergence of the derivative of the one-electron HF energy.

2.7.2. Heyd-Scusseria-Ernzherof (HSE) hybrid functionals

The starting point of this new class of hybrid screened functionals or the so-called Heyd-Scusseria-Ernzherof screened hybrid functional (HSE03)[26] is the separation of the Coulomb interaction operator into short-range (*sr*) and long-range (*lr*) components[162], respectively:

$$\frac{1}{r} = \underbrace{\frac{erfc(wr)}{r}}_{short-range} + \underbrace{\frac{erf(wr)}{r}}_{long-range} \qquad (2.7.2.1)$$

where $w = \frac{2}{R_{sr}}$ is an adjustable parameter that describes the range of short-range interactions $\left(R_{sr}\right)$. Choosing the error function *erf(wr*) and its complementary error function $erfc(wr) = 1 - erf(wr)$ to achieve the Coulomb partition, ensures that the short-range component $\frac{erfc(wr)}{r}$ is singular and rapidly decays as a Gaussian, while the long-range part $\frac{erf(wr)}{r}$ is nonsingular and smoothly decays as *r* tends to infinity. Figure 3 illustrates the behavior of 1/*r* as *r* increases, the fast decay of $\frac{erfc(wr)}{r}$ and the smooth decay of $\frac{erf(wr)}{r}$ for $w = 1$.

The construction of HSE screened hybrid functionals is based on the admixture of both HF and PBE-GGA exchange in the short range parts while the long range term is solely represented by the PBE-GGA functional.[8]

Before further describing the HSE hybrid functional, we initially start by writing the exchange correlation parameter PBE0 hybrid functional as:

$$E_{XC}^{PBE0} = a_1 E_X^{HF} + (1-a_1)E_X^{PBE} + E_C^{PBE}$$
$$\Leftrightarrow E_{XC}^{PBE0} = a_1 E_X^{HF} + E_X^{PBE} - a_1 E_X^{PBE} + E_C^{PBE} \quad (2.7.2.2)$$

where a_1 = 1/4 is the exchange coefficient that is determined by an adiabatic expansion calculated by Perdew et al. (2006).[23]

By partitioning all terms of the above PBE0 exchange energy into short range (*sr*) and long-range (*lr*) components, we obtain the HSE functional:

$$E_{XC}^{HSE} = a_1\left[E_X^{HF,lr}(w) + E_X^{HF,sr}(w)\right] + \left[E_X^{PBE,lr}(w) + E_X^{PBE,sr}(w)\right] + - a_1\left[E_X^{PBE,lr}(w) + E_X^{PBE,sr}(w)\right] + E_C^{PBE}$$

$$\Leftrightarrow E_{XC}^{HSE} = a_1\left[E_X^{HF,lr}(w) - E_X^{PBE,lr}(w)\right] + a_1\left[E_X^{HF,sr}(w) - E_X^{PBE,lr}(w)\right] + \left[E_X^{PBE,lr}(w) + E_X^{PBE,sr}(w)\right] + E_C^{PBE} \quad (2.7.2.3)$$

By performing benchmark numerical tests on physically acceptable values of *w*, Ref. [26] noticed that the *lr* HF and PBE exchange terms tend to cancel each other out, yielding:

$$E_{XC}^{HSE} = a_1\left[E_X^{HF,sr}(w) - E_X^{PBE,lr}(w)\right] + \left[E_X^{PBE,lr}(w) + E_X^{PBE,sr}(w)\right] + E_C^{PBE}$$
$$\Leftrightarrow E_{XC}^{HSE} = a_1 E_X^{HF,sr}(w) + (1-a_1)E_X^{PBE,sr}(w) + E_X^{PBE,lr}(w) + E_C^{PBE}$$

By substituting the value of the coefficient a_1 into the above expression, we obtain:

$$E_{XC}^{HSE} = \frac{1}{4}E_X^{HF,sr}(w) + \frac{3}{4}E_X^{PBE,sr}(w) + E_X^{PBE,lr}(w) + E_C^{PBE} \quad (2.7.2.4)$$

In order to achieve balance between physically accurate results and computational effort, the screening parameter for the HF and PBE was chosen to be $w_{HF} = \frac{0.15}{\sqrt{2}}$ Bohrs^{-1} (or $R_{sr}^{HF} \approx 9.97$ Å) and $w_{PBE} = 0.15 \times 2^{1/3}$ Bohrs^{-1} (or $R_{sr}^{PBE} \approx 5.6$ Å). As the adjustable parameter *w* goes to 0, the HSE functional becomes very similar to the original hybrid PBE0 method

(discrepancies are caused by the mathematical construct of the XC-hole)[26] and is equivalent to pure PBE-GGA as *w* approaches infinity. The elimination of the long-range component of the Fock exchange drastically reduces the range over which the integral of the exact exchange is computed.[8] Subsequently, in real space, the HSE03 is less computationally demanding than most hybrid functionals and can therefore be applied to solids and large molecules. Furthermore, in reciprocal space, due to the increased locality of the Fock exchange in the HSE03 formalism, the Fock exchange can be computed in a much coarser mesh of points in the Brillouin zone.[8]

The precision of the HSE03 method is illustrated in Fig. 4 where theoretical band gaps of various materials are compared with their respective experimental band gaps. The band gaps obtained by the PBE formalism (blue circles) are largely underestimated due to the approximation of the exchange and correlation energy. The hybrid functional PBE0 (green squares) yields better results than the PBE but it still tends to overestimate the band gaps of some materials because of the incorporation of unscreened Fock exchange in their formalism. The HSE03 functional computed with the standard amount of Fock exchange (25 %) provides better agreement with the experimental results. Even though HSE03 functional allows some improvements in the correct prediction of thermochemical energies, band-gaps and atomization energies, its precision is yet to be desired. Paier et al. (2006)[8] argue that the HSE03 functional is not mature enough to fully replace well-known semi-local functionals because of the severe underestimation of the cohesive energy in several systems, the overestimation of the magnetic moment of transition metals and incorrect prediction of large gaps materials. Such inaccuracy might be traced back to the use of different values of *w* in the development of the original functional.

The construction of a new screened hybrid functional (HSE06)[31] was based on the reexamination of the screening parameter w in which:

$w_{HF} = w_{PBE} \approx 0.11$ Bohr^{-1} or $R_{sr}^{HF} = R_{sr}^{PBE} \approx 9.62$ Å

With such adjustment, better thermochemical predictions are obtained while equilibrium between computational cost and accurate physical results is still maintained.

Now that we have provided a brief overview of the state-of-the-art methodology for performing first-principles calculations, we will introduce in the next section how to implement these calculations to investigate the electronics and optical properties of defects in GaN.

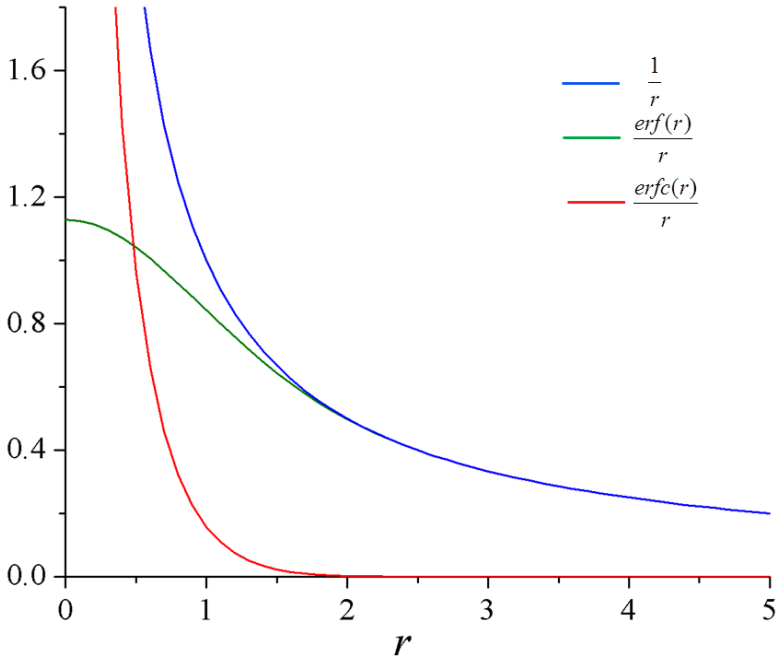

Figure 3: Graphs of $\frac{1}{r}$, $\frac{erf(wr)}{r}$ and $\frac{erfc(wr)}{r}$ in function of r from Eq. 2.7.2.1 for $w=1$. In the short range, one notices that $\frac{erfc(r)}{r}$ (red color) displays a rapid decay similar to the inverse function $\frac{1}{r}$ (blue color), while in the long range $\frac{erf(r)}{r}$ (green color) is identical to $\frac{1}{r}$

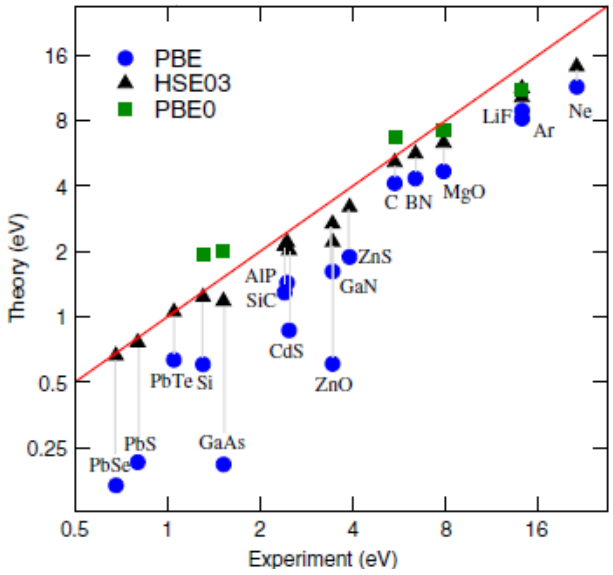

Figure 4: Illustrative comparison of band gaps done by Marsman et al.[163] where the theoretical band gaps obtained from PBE, PBE0 and HSE03 calculations are plotted against the experimental band gaps.

Section 3. Techniques for estimating supercell defects calculations

In the preceding KS chapter (section 2.6), it was discussed that several observables of the many-body system can be mapped into corresponding observables in a single-particle problem subjected to an effective potential. Nevertheless, the heavy task of computing the electronic density of a system that is subjected to periodic boundary conditions still remains. Such dilemma can be overcome by applying Bloch's theorem to the HSE06 formalism and by introducing plane waves basis set. Plane waves are aesthetically appealing for periodic systems since they possess the advantage of completely spanning the Hilbert space and they also provide mathematical simplicity for practical calculations. The following section is organized first to describe the HSE06 method in reciprocal space via the use of Bloch's theorem. Then, the remaining sections are devoted to explaining relevant concepts that enter in the computation of formation energies and optics of defects in GaN.

3.1. Plane waves (PW) basis sets in HSE06 formalism

In this chapter, the derivation of the basis sets used in our calculations is based on the description of PW used in Refs. [12,151]. We initially start by recalling Bloch's theorem[161] stating that within a perfectly periodic potential, each electronic wave function can be rewritten as a product of a wavelike part and a cell-periodic part where:

$$\Psi_{j,\vec{k}}(\vec{r}) = e^{i\vec{k}\cdot\vec{r}} u_{j,\vec{k}}(\vec{r}), \tag{3.1.1}$$

where $\vec{k}$ represents the wave vector of the PW confined within the first Brillouin zone (BZ) and j describes the band index. Here $u_{j,\vec{k}}(\vec{r})$ possesses the periodicity of the potential and obeys the

relation $u_{j,\vec{k}}(\vec{r}) = u_{j,\vec{k}}(\vec{r}+\vec{R})$ for all $\vec{R}$ in a Bravais lattice vector. Since $u_{j,\vec{k}}(\vec{r})$ is a periodic function, it can be expanded using a basis set of plane waves whose wave vectors are reciprocal lattice vectors of the crystal,

$$u_{j,\vec{k}} = \sum_{\vec{G}=0}^{\infty} C_{j,\vec{k}}(\vec{G}) e^{i\vec{G}\bullet\vec{r}} \tag{3.1.2}$$

Here, $\vec{G}$ are the reciprocal lattice vectors defined by $\vec{G}\cdot\vec{R} = 2\pi m$ for all $\vec{R}$, m is any integer and $C_{j,\vec{k}}(\vec{G})$ are the plane wave coefficients. By inserting Eq. 3.1.2 into Eq. 3.1.1, each electronic wave function in a cell of volume V can be described by:

$$\Psi_{j,\vec{k}}(\vec{r}) = \frac{1}{\sqrt{V}} \sum_{\vec{G}=0}^{\infty} C_{j,\vec{k}}(\vec{G}) e^{i(\vec{k}+\vec{G})\bullet\vec{r}} \tag{3.1.3}$$

where the plane waves basis functions are defined by

$$\psi_{\vec{G}}(\vec{r}) = \frac{1}{\sqrt{V}} e^{i\vec{G}\bullet\vec{r}} \tag{3.1.4}$$

which satisfy the orthonormality conditions:

$$\left\langle \psi_{\vec{G}'} \middle| \psi_{\vec{G}} \right\rangle = \frac{1}{V}\int_V e^{i(\vec{G}-\vec{G}')\bullet\vec{r}} d\vec{r} = \delta_{\vec{G},\vec{G}'} \tag{3.1.5}$$

where $\vec{G}'$ and $\vec{G}$ differ by a reciprocal latticevector $\vec{G}''$, or $\vec{G}'' = \vec{G} - \vec{G}'$

Now Eq. 3.1.3 can be rewritten as:

$$\Psi_{j,\vec{k}}(\vec{r}) = \sum_{\vec{G}=0}^{\infty} C_{j,\vec{k}}(\vec{G}) e^{i\vec{k}\bullet\vec{r}} \psi_{\vec{G}}(\vec{r}) \tag{3.1.6}$$

From the previous equation, one remarks that except for $\vec{G} = 0$, the reciprocal lattice vectors $\vec{G}$ that describe the plane wave expansion always lie outside the first BZ, while the wavelike part $e^{i\vec{k}\bullet\vec{r}}$ involves a wave vector $\vec{k}$ in the first BZ.[151] However, if one is interested into solving the HSE06 canonical equations in a self-consistent manner, the electronic density must be expressed

as a *BZ average*[151] where the $\vec{k}$ wave vectors must occupy the entire BZ. Thus one must incorporate the wavelike part into our previous plane wave basis functions in order to cover the entire BZ:

$$\psi_{\vec{k},\vec{G}} = \frac{1}{\sqrt{V}} e^{i(\vec{k}+\vec{G})\cdot\vec{r}} \tag{3.1.7}$$

Now the electronic wave function can be expressed as:

$$\Psi_{j,\vec{k}}(\vec{r}) = \sum_{\vec{G}=0}^{\infty} C_{j,\vec{k}}(\vec{G})\psi_{\vec{k},\vec{G}}(\vec{r}) \text{ , anywhere in the BZ.} \tag{3.1.8}$$

Here the coefficient $C_{j,\vec{k}}(\vec{G})$ is given by:

$$\psi^*_{\vec{k},\vec{G}'}(\vec{r})\Psi_{j,\vec{k}}(\vec{r}) = \sum_{\vec{G}=0}^{\infty} \psi^*_{\vec{k},\vec{G}'}(\vec{r})C_{j,\vec{k}}(\vec{G})\psi_{\vec{k},\vec{G}}(\vec{r})$$

$$\Rightarrow \int \psi^*_{\vec{k},\vec{G}'}(\vec{r})\Psi_{j,\vec{k}}(\vec{r})d\vec{r} = \int \sum_{\vec{G}=0}^{\infty} \psi^*_{\vec{k},\vec{G}'}(\vec{r})C_{j,\vec{k}}(\vec{G})\psi_{\vec{k},\vec{G}}(\vec{r})d\vec{r}$$

$$\Rightarrow \int \psi^*_{\vec{k},\vec{G}'}(\vec{r})\Psi_{j,\vec{k}}(\vec{r})d\vec{r} = \sum_{\vec{G}=0}^{\infty} C_{j,\vec{k}}(\vec{G})\delta_{\vec{G},\vec{G}'}$$

$$\Rightarrow C_{j,\vec{k}}(\vec{G}) = \int \psi^*_{\vec{k},\vec{G}}(\vec{r})\Psi_{j,\vec{k}}(\vec{r})d\vec{r}$$

The expansion of the kinetic energy term from the HSE06 canonical equations in term of plane waves gives rise to a diagonal kinetic energy operator which is given by:

$$\hat{T}_{\vec{G},\vec{G}'} = \left\langle \Psi_{j,\vec{k}}(\vec{r}) \middle| -\frac{1}{2}\nabla_j^2 \middle| \Psi_{j,\vec{k}}(\vec{r}) \right\rangle = \int \Psi^*_{j,\vec{k}}(\vec{r}) \cdot \left[-\frac{1}{2}\nabla_j^2 \right] \Psi_{j,\vec{k}}(\vec{r})\, d\vec{r}$$

$$\Leftrightarrow \hat{T}_{\vec{G},\vec{G}'} = \sum_{\vec{G}}^{\infty}\sum_{\vec{G}'}^{\infty} \int C^*_{j,\vec{k}}(\vec{G}') \frac{1}{\sqrt{V}} e^{-i(\vec{k}+\vec{G}')\cdot\vec{r}} \cdot C_{j,\vec{k}}(\vec{G}) \frac{\left|\vec{k}+\vec{G}\right|^2}{2\sqrt{V}} e^{i(\vec{k}+\vec{G})\cdot\vec{r}} d\vec{r}$$

$$\Leftrightarrow \hat{T}_{\vec{G},\vec{G}'} = \sum_{\vec{G}}^{\infty}\sum_{\vec{G}'}^{\infty} C^*_{j,\vec{k}}(\vec{G}') C_{j,\vec{k}}(\vec{G}) \frac{\left|\vec{k}+\vec{G}\right|^2}{2} \frac{1}{V} \int e^{i(\vec{G}-\vec{G}')\cdot\vec{r}} d\vec{r}$$

$$\Leftrightarrow \hat{T}_{\vec{G},\vec{G}'} = \sum_{\vec{G}}^{\infty}\sum_{\vec{G}'}^{\infty} C^*_{j,\vec{k}}(\vec{G}') C_{j,\vec{k}}(\vec{G}) \frac{\left|\vec{k}+\vec{G}\right|^2}{2} \delta_{\vec{G},\vec{G}'}$$

$$\Leftrightarrow \hat{T}_{\vec{G},\vec{G}'} = \frac{1}{2}\sum_{\vec{G}}^{\infty}\left|C_{j,\vec{k}}\left(\vec{G}\right)\right|^2\left|\vec{k}+\vec{G}\right|^2 \tag{3.1.9}$$

In addition to the kinetic energy, the Hartree potential can also be computed in reciprocal space where the Hartree term is expressed in terms of its Fourier transforms:

$$u^{Hart}\left(\vec{G}\right) = \frac{1}{V}\int u^{Hart}(\vec{r})e^{-i\vec{G}\bullet\vec{r}}d\vec{r} = \frac{1}{V}\int\int\frac{n(\vec{r}')}{\left|\vec{r}-\vec{r}'\right|}e^{-i\vec{G}\bullet\vec{r}}d\vec{r}'d\vec{r}$$

$$\Leftrightarrow u^{Hart}\left(\vec{G}\right) = \frac{1}{V}\int\int\frac{n(\vec{r}')}{\left|\vec{r}-\vec{r}'\right|}e^{-i\vec{G}\bullet(\vec{r}-\vec{r}')}e^{-i\vec{G}\bullet\vec{r}'}d\vec{r}'d\vec{r}$$

Let $\vec{v} = \vec{r} - \vec{r}'$, then: $\frac{d\vec{v}}{d\vec{r}} = 1$ or $d\vec{v} = d\vec{r}$ and the above equation becomes:

$$u^{Hart}\left(\vec{G}\right) = \frac{1}{V}\int\int\frac{n(\vec{r}')}{\left|\vec{v}\right|}e^{-i\vec{G}\bullet\vec{v}}e^{-i\vec{G}\bullet\vec{r}'}d\vec{r}'d\vec{v}$$

$$\Leftrightarrow u^{Hart}\left(\vec{G}\right) = \underbrace{\int\frac{1}{\left|\vec{v}\right|}e^{-i\vec{G}\bullet\vec{v}}d\vec{v}}_{\frac{4\pi}{G^2}}\cdot\frac{1}{V}\int n(\vec{r}')e^{-i\vec{G}\bullet\vec{r}'}d\vec{r}'$$

$$\Leftrightarrow u^{Hart}\left(\vec{G}\right) = \frac{4\pi}{G^2}\frac{1}{V}\int n(\vec{r}')e^{-i\vec{G}\bullet\vec{r}'}d\vec{r}'$$

$$\Leftrightarrow u^{Hart}\left(\vec{G}\right) = \frac{4\pi}{G^2}n\left(\vec{G}\right) \text{ where } n\left(\vec{G}\right) = \frac{1}{V}\int n(\vec{r}')e^{-i\vec{G}\bullet\vec{r}'}d\vec{r}' \tag{3.1.10}$$

From the expression of the above equation, we notice a divergence of the Hartree potential in reciprocal space at $\vec{G} = 0$. More details regarding the physical meaning of the $\vec{G} = 0$ are given in section 3.3.2.

By expanding the Hartree term in function of the electronic PWs, we obtain:

$$\left\langle u^{Hart}\left(\vec{G}\right)\right\rangle = \left\langle \Psi_{j,\vec{k}}\left(\vec{r}\right)\middle| u^{Hart}\left(\vec{G}\right)\middle|\Psi_{j,\vec{k}}\left(\vec{r}\right)\right\rangle$$

$$\Leftrightarrow \left\langle u^{Hart}\left(\vec{G}\right)\right\rangle = \frac{1}{V}\int\sum_{\vec{G}}^{\infty}\sum_{\vec{G}'}^{\infty} C^{*}_{j,\vec{k}}\left(\vec{G}'\right)C_{j,\vec{k}}\left(\vec{G}\right)e^{i(\vec{G}-\vec{G}')\bullet\vec{r}}\,\frac{4\pi}{G^2}n\left(\vec{G}\right)d\vec{r}$$

$$\Leftrightarrow \left\langle u^{Hart}\left(\vec{G}\right)\right\rangle = \sum_{\vec{G}}^{\infty} C_{j,\vec{k}}\left(\vec{G}\right)\sum_{\vec{G}'}^{\infty} C^{*}_{j,\vec{k}}\left(\vec{G}'\right)\frac{4\pi}{G^2}n\left(\vec{G}\right)\left[\frac{1}{V}\int e^{i(\vec{G}-\vec{G}')\bullet\vec{r}}d\vec{r}\right]$$

$$\Leftrightarrow \left\langle u^{Hart}\left(\vec{G}\right)\right\rangle = \sum_{\vec{G},\vec{G}'}^{\infty} C_{j,\vec{k}}\left(\vec{G}\right)C^{*}_{j,\vec{k}}\left(\vec{G}'\right)\frac{4\pi}{G^2}n\left(\vec{G}\right)\delta_{\vec{G},\vec{G}'}$$

$$\Leftrightarrow \left\langle u^{Hart}\left(\vec{G}\right)\right\rangle = \sum_{\vec{G}}^{\infty}\left|C_{j,\vec{k}}\left(\vec{G}\right)\right|^2\frac{4\pi}{G^2}n\left(\vec{G}\right) \qquad (3.1.11)$$

Contrary to the Hartree term, we do not have an explicit form for the XC-potential, hence its general form in reciprocal space is given by:

$$u_{XC}\left(\vec{G}\right) = \left\langle \Psi_{j,\vec{k}}\left(\vec{r}\right)\middle| u_{XC}\left(\vec{r}\right)\middle|\Psi_{j,\vec{k}}\left(\vec{r}\right)\right\rangle = \frac{1}{V}\int\sum_{\vec{G}}^{\infty}\sum_{\vec{G}'}^{\infty} C^{*}_{j,\vec{k}}\left(\vec{G}'\right)C_{j,\vec{k}}\left(\vec{G}\right)e^{i(\vec{G}-\vec{G}')\bullet\vec{r}}u_{XC}\left(\vec{r}\right)d\vec{r}$$

$$\Leftrightarrow u_{XC}\left(\vec{G}\right) = \sum_{\vec{G}}^{\infty} C_{j,\vec{k}}\left(\vec{G}\right)\sum_{\vec{G}'}^{\infty} C^{*}_{j,\vec{k}}\left(\vec{G}'\right)\left[\frac{1}{V}\int e^{i(\vec{G}-\vec{G}')\bullet\vec{r}}u_{XC}\left(\vec{r}\right)d\vec{r}\right]$$

$$\Leftrightarrow u_{XC}\left(\vec{G}\right) = \sum_{\vec{G},\vec{G}'}^{\infty} C_{j,\vec{k}}\left(\vec{G}\right)C^{*}_{j,\vec{k}}\left(\vec{G}'\right)u_{XC}\left(\vec{G}-\vec{G}'\right) \qquad (3.1.12)$$

In addition to the XC-potential, we can also rewrite the external potential in function of $\vec{G}$ in the same form as in the above equation:

$$u_{ext}\left(\vec{G}\right) = \sum_{\vec{G},\vec{G}'}^{\infty} C_{j,\vec{k}}\left(\vec{G}\right)C^{*}_{j,\vec{k}}\left(\vec{G}'\right)u_{ext}\left(\vec{G}-\vec{G}'\right) \qquad (3.1.13)$$

One can also show that similarly to the Hartree term, the external potential also displays a divergence for the $\vec{G}=0$ term and details regarding its derivation can be found in pps. 241-242 of Ref. [7].

Now that we have described each component of the HSE06 canonical equations in reciprocal space, Eq. 2.6.1.1 is rewritten as:

$$\frac{1}{2}\sum_{\vec{G}}^{\infty}\left|C_{j,\vec{k}}\left(\vec{G}\right)\right|^2\left|\vec{k}+\vec{G}\right|^2+\sum_{\vec{G},\vec{G}'}^{\infty}C_{j,\vec{k}}\left(\vec{G}\right)C_{j,\vec{k}}^{*}\left(\vec{G}'\right)u_{ext}\left(\vec{G}-\vec{G}'\right)+$$
$$\sum_{\vec{G}}^{\infty}\left|C_{j,\vec{k}}\left(\vec{G}\right)\right|^2\frac{4\pi}{G^2}n\left(\vec{G}\right)+\sum_{\vec{G},\vec{G}'}^{\infty}C_{j,\vec{k}}\left(\vec{G}\right)C_{j,\vec{k}}^{*}\left(\vec{G}'\right)u_{XC}\left(\vec{G}-\vec{G}'\right)=\sum_{\vec{G}'}^{\infty}\left|C_{j,\vec{k}}\left(\vec{G}'\right)\right|^2\varepsilon_{\vec{k},j} \tag{3.1.14}$$

Here the canonical equations are self-consistently solved for each $\vec{k}$ vector in the Brillouin zone (BZ).

Since we derived the expression of the electronic wave function, the electronic charge density is derived from an average of the BZ as:

$$n(\vec{r})=\sum_{\vec{k}}^{BZ}\eta_{\vec{k}}\sum_{j=1}^{J}Q_{\vec{k},j}\Psi_{\vec{k},j}^{*}\left(\vec{r}\right)\Psi_{\vec{k},j}\left(\vec{r}\right) \tag{3.1.15}$$

The first sum in the right hand side of the above equation denotes the sum over all $\vec{k}$ vectors belonging to the BZ while the second sum runs from the *j*-th band to the total number of bands *J* occupied at each $\vec{k}$ vector. Here $\eta_{\vec{k}}$ denotes the weight factors which depend on the symmetry of the system and $Q_{\vec{k},j}$ describes the occupation number of the *j*-th band at each $\vec{k}$.

In principle, provided that each electron occupies a band *j* corresponding to a specific $\vec{k}$, we require an infinite number of reciprocal lattice vectors $\vec{G}$ to represent the wave functions with absolute accuracy. However, it can be shown that the contributions of Fourier coefficients $C_{j,\vec{k}}(\vec{G})$ of the plane wave function are inversely proportional to $\left|\vec{k}+\vec{G}\right|^2$ which means that the infinite plane wave expansion can be efficiently shortened to a finite number of terms. By introducing a kinetic cut-off energy, we can choose all basis functions into the basis set that

fulfill $\frac{1}{2}\left|\vec{k}+\vec{G}\right|^2 < E_{Cutoff}$. At the Γ point, we obtain a cutoff sphere whose radius is given by:

$$\left|\vec{G}\right| < \left(2E_{Cutoff}\right)^{\frac{1}{2}}. \tag{3.1.16}$$

Nevertheless, the optimum value of the cutoff energy is not trivially defined. In a real many-body system, the electronic wave functions react to the nuclear potential. This means that as one gets closer to the nuclei, the spatial variation of wave functions becomes quite fast and one must require an infinitely large number of plane waves. To overcome these difficulties, we will be using a pseudopotential method, i.e the projector-augmented-wave (PAW) method in which the KS PW electronic wave function is substituted by an auxiliary smooth wave function within a cutoff radius of an augmentation sphere centered on the nuclei. In the PAW method, the KS core states are identical to the atomic core states (frozen core approximation) and the valence wave functions within the augmentation sphere become extremely smooth. As a result, the previously discussed energy cutoff can directly be approximated from the atomic pseudopotentials but it usually needs to be tested in sets of benchmark systems. More details on the linear transformation of the KS PW into pseudized smooth wave functions can be found in Refs. [164,165].

Although technically appealing, the application of Bloch's theorem is not valid in semiconductors containing impurities since the incorporation of external defects breaks the perfect periodicity of the system. In order to overcome such problem, one might require the creation of a large cell (supercell) that would be periodically reproduced throughout space.

3.2. Supercell method using HSE06

Before creating a supercell, one might be interested into calculating the theoretical lattice constant of the host crystal which is obtained by relaxing atoms in the primitive unit cell. Relaxation is obtained by displacing atoms in such way that energy, forces, stresses and any other output quantities of the HSE06 self-consistent loop (Fig. 2) are minimized. Once we obtain a converged lattice constant, we create a supercell in which the defect is surrounded by a region of bulk crystal that is subjected to periodic boundary conditions. The supercell method is quite convenient since it allows the use of Bloch's theorem which requires translational periodicity of the system. However, it is crucial to include enough bulk solid in the supercell such that the defects are sufficiently separated and properties of isolated defects can thus be computed. In case of a charged system, the supercell approximation introduces the concept of uniform, neutralizing jellium background charge that circumvents the divergence of the Coulomb energy between the periodic charged defect images. However, the inclusion of a jellium background creates fictitious interactions which need to be corrected (cf. section 3.5). By using appropriate corrections, the supercell method becomes quite accurate and can be used within the HSE06 formalism to investigate the total energy of a system containing defects. Before describing the nature of any type of corrections, in the next section, we shall discuss how to compute the likelihood of a defect to be formed in a perfect lattice within the supercell method.

3.3. Defect Formation Energy

The probability of realizing a particular defect configuration *D* in a host lattice containing *p*-type atoms in the charge state *q* within the supercell formalism is given by:

$$E_f[D^q] = E_{tot}[D^q] - E_{tot}[bulk] + q\cdot(E_{VBM} + \Delta E_F) - \sum_p^P n_p\mu_p + \Delta E_{PA} + \Delta E_{LZ} \quad (3.3.1)$$

where $E_{tot}[D^q]$ and $E_{tot}[bulk]$ are the total energies of the *bulk+defect* and bulk-only supercell, respectively. The third term $q\cdot(E_{VBM} + \Delta E_F) = qE_F$ is the amount of energy it cost to charge the neutral impurity assuming that the exchange of electrons occurs at the Fermi level E_F. The number of atoms of type *p* (bulk atoms or defect atoms) that have been removed from or added to the host lattice are represented by the variable n_p. Here, μ_p indicates the elemental chemical potential of its corresponding *p*-th atom.

In principle, in thermodynamic equilibrium, Gibbs free energy of formation should be used instead of Eq. 3.3.1 in order to estimate the concentration of defects in GaN.[64,204] One reason is that defect concentrations are determined by the free energy of defects, and in cases of high temperature growth such as MOCVD, entropic contributions are significant.[36,166] This means that it is difficult to expect direct correspondence between computed formation energies and measured defect concentrations. Furthermore, experimentally measured defect concentrations also tend to be very sensitive to atomic fluxes in the growth chamber, orientations of the facets being grown, and many other "uncontrolled" factors. All of them are not included in any of our calculations, and therefore defect concentrations dictated by the equilibrium formation energies are most likely very different from measured defect concentrations.

However, although absolute values of formation energy are quite uncertain, the intersections of the formation energy lines, i.e. thermodynamic transition levels (see section 3.4.1) are well defined quantities, since any uncertainty in calculation of formation energy, due to either the choice of the elemental chemical potentials or discrepancies between the computed formation energies and actual concentrations of defects in the sample, cancels out. Thus, transition levels (including thermodynamic and optical) can be directly compared to PL experimental results.

3.3.1. Chemical potentials

Chemical potentials in semiconductors are typically obtained by:

a) initially establishing the competing phases for all atomic elements involved in the semiconductor growth,
b) subsequently selecting the most energetically stable energy phases of the corresponding atomic elements when compared to bulk GaN energy phase.

Nevertheless, accurate identification of all existing competing phases of the atomic elements involved in the semiconductor formation is far from trivial due to the variety of growth methods and environmental growth conditions. Therefore, exact computation of elemental chemical potentials is often impossible and one usually estimates atomic chemical potentials in limiting or extreme growth regimes.

For instance, in a host GaN lattice, the chemical potentials of its individual constituents gallium (Ga) and nitrogen (N), can be theoretically estimated in extreme Ga-rich and N-rich environmental growths. According to Van de Walle et al. (2004)[64], these limiting environmental conditions correspond to placing *upper and lower bounds* on the chemical potentials of Ga and N, respectively. As a result, Ga-rich conditions are present when Ga chemical potential is

subjected to an upper bound and equals that of metal Ga, where $\mu_{Ga}^{up} = \mu_{Ga}^{[metal]}$. Here for Ga atom, we are assuming that the formation of metallic Ga is the next competing phase to the growth of bulk GaN. Similarly, extreme nitrogen (N)-rich conditions occur when the chemical potential of N is also subjected to an upper bound and equals that of N_2 gas, where $\mu_N^{up} = \mu_N^{[N_2]}$. Here in the case of N, we are presuming that the formation of N_2 gas is the secondary phase to the formation of GaN. The sum of the chemical potentials of Ga and N atoms denotes the energy of bulk GaN, which describes the stability condition for the growth of GaN. For a primitive GaN unit cell composed of two atoms, its energy is defined as:

$$\mu\left(E_{GaN}^{prim}\right) = \mu_{Ga} + \mu_N \tag{3.3.1.1}$$

Enforcing *an upper bound* on μ_{Ga}, obtained from the energy of metallic Ga leads to a lower bound on μ_N:

$$\mu_N^{low} = \mu\left(E_{GaN}^{prim}\right) - \mu_{Ga}^{[metal]} \tag{3.3.1.2}$$

Analogically, enforcing an *upper bound* on μ_N, given by the energy of N_2 gas yields a lower bound on μ_{Ga}:

$$\mu_{Ga}^{low} = \mu\left(E_{GaN}^{prim}\right) - \mu_N^{[N_2]}. \tag{3.3.1.3}$$

By definition, the enthalpy of formation $\Delta H\left(GaN\right)$ or the energy gain in forming the crystal bulk GaN is defined as:

$$\Delta H\left(GaN\right) = \mu\left(E_{GaN}^{prim}\right) - \left[\mu_{Ga}^{[metal]} + \mu_N^{[N_2]}\right] \tag{3.3.1.4}$$

$$\Leftrightarrow \mu\left(E_{GaN}^{prim}\right) = \Delta H\left(GaN\right) + \mu_N^{[N_2]} + \mu_{Ga}^{[metal]} \tag{3.3.1.5}$$

By substituting Eq. 3.3.1.5 into Eq. 3.3.1.3, we notice that Ga-rich environment is achieved by setting:

$$\mu_N^{low} = \Delta H\left(GaN\right) + \mu_N^{[N_2]} + \mu_{Ga}^{[metal]} - \mu_{Ga}^{[metal]}$$
$$\Leftrightarrow \mu_N^{low} = \Delta H\left(GaN\right) + \mu_N^{[N_2]} \quad (3.3.1.6)$$

given that $\mu_{Ga} = \mu_{Ga}^{[metal]}$.

Similarly, N-rich condition is also obtained by setting:

$$\mu_{Ga}^{low} = \mu_{Ga}^{[metal]} + \Delta H\left(GaN\right) \text{ given that } \mu_N = \mu_N^{[N_2]} \quad (3.3.1.7)$$

Here, $\Delta H\left(GaN\right)$ can be calculated using total energies of the primitive GaN unit cell, orthorhombic metal Ga, and N_2 molecule, with volumes and atomic coordinates fully relaxed within the HSE method. Furthermore, the formation of other solubility-limiting phases due to the incorporation of impurities should also be considered. For instance, it has been suggested by Ref. [93] that when oxygen (O) is being incorporated in GaN, the O atom can interact with N and form Ga_2O_3 which corresponds to the next competing phase to the growth of GaN, yielding:

$$2\mu_{Ga}^{[Ga_2O_3]} + 3\mu_O^{[Ga_2O_3]} = 2\mu_{Ga}^{[metal]} + 3\mu_O^{[O_2]} + \Delta H(Ga_2O_3) \quad (3.3.1.8)$$

In the left-hand side of the equation, Ga and O are taken from the Ga_2O_3 reservoir and in the right-hand side, Ga and O are taken from metallic Ga and O_2 molecule gas reservoirs, respectively. Here $\Delta H(Ga_2O_3)$ is the enthalpy of formation of Ga_2O_3, experimentally measured to be approximately -11.29 eV.[167] Following Ref. [64], we calculated that the solubility limit of Ga_2O_3 occurs under Ga-rich environmental growth, i.e Ga-rich, which means that the chemical potential of Ga is expressed as:

$$\mu_{Ga}^{[Ga_2O_3]} = \mu_{Ga}^{[metal]} \quad (3.3.1.9)$$

and in N-rich:

$$\mu_{Ga}^{[Ga_2O_3]} = \mu_{Ga}^{[metal]} + \Delta H(GaN) \quad (3.3.1.10)$$

By combining Eqs. 3.3.1.9 and 3.3.1.8, the chemical potential of O in Ga-rich conditions that is taken from the Ga_2O_3 reservoir is described as:

$$\mu_{Ga}^{[Ga_2O_3]} = \frac{1}{3}\left[3\mu_O^{[O_2]} + \Delta H(Ga_2O_3)\right], \tag{3.3.1.11}$$

while in N-rich, the chemical potential of O is expressed as:

$$\mu_O^{[Ga_2O_3]} = \frac{1}{3}\left[3\mu_O^{[O_2]} - 2\Delta H(GaN) + \Delta H(Ga_2O_3)\right] \tag{3.3.1.12}$$

In principle, for gas-phase molecules such as N_2 or O_2, it is crucial to take into account temperature and pressure dependence in the calculation of chemical potentials. In this work, neither pressure nor temperature are taken into consideration since, as previously discussed in section 3.3, the calculated thermodynamic transition levels (that can be compared to experiment), are obtained from formation energy differences; hence any ill-defined physical quantities such as chemical potentials will be subtracted out and will not affect the calculations of the thermodynamic or optical transition levels.

3.3.2. Adjustment of finite-size effects in supercell calculations

Describing defects in a supercell method from a band structure perspective seems quite intuitive. Interactions between defects in neighboring supercells usually lead to a dispersed impurity band instead of a single localized eigenstate. In case of an infinitely large supercell in which the impurity would be totally isolated, the defect-induced band would be completely flat. According to Van de Walle et al. (2004)[64], one might avoid choosing the Γ point as one of the sampling points since at such high symmetry point, defect-defect interactions reach its maximum which would thus lead to mediocre description of the band structure of a given system. One way to circumvent such difficulty would be the use of special **k**-points which provides a way of averaging over the defect band and therefore offers a better way to plot the band structure. Such method would essentially compute the impurity band's center of mass whose band level would be quite similar to the case of a completely isolated impurity embedded in a SC.

One may disagree with such method by arguing that in the case that the VBM/CBM-defect band interactions becomes so strong that the actual level of the defect would be shifted and would never correspond to the approximated center of mass that is computed from the special **k**-points method. Furthermore, for the case of deep defects, by using large supercells (up to 300 atoms), interactions between neighboring defects is almost negligible and including the Γ point would lead to acceptable description of the band structure of the system. Therefore, the incorporation of the Γ point in the BZ integration results to adequate accuracy and further numerical simplicity for first-principles computations when compared to the use of the more complex special **k**-points method.

The use of the supercell method within the HSE06 self-consistent calculations includes several physical errors that need to be corrected. While various approaches for such corrections

have been suggested in the literature[42,168,169], we will be briefly discussing two different sets of corrections methods used in our calculations.

3.3.2.1. Image-charge correction

Computations of total energies or formation energies of a periodically repeated charged system require care because of the divergence of the $\vec{G}=0$ terms in the Hartree and ionic potentials in the canonical KS equations (Eq. 3.1.11). In the neutral case, the $\vec{G}=0$ term is dropped because of the exact cancellation between the positive kinetic energy associated with the rapid fluctuations of the wave function and the negative potential energy of the electron close to the nuclei.[170] In a charged supercell, we are eliminating the $\vec{G}=0$ component in the Fourier expansion of the canonical KS equations by introducing a compensating background charge (jellium) that restores the neutrality of the system and hence avoids divergence of long-range Coulomb terms. Even though this artificial background takes care of the calculation of the $\vec{G}=0$ terms, the charge compensation only affects the total potential while the jellium's corresponding charge density is usually not included in the KS self-consistent calculations.[171] Furthermore, the use of periodically repeated finite-sized supercells in our calculations introduces fictitious long-range electrostatic and elastic interactions of charged defects with its periodic images and compensating background. These spurious electrostatic interactions have to be corrected for, which is the ΔE_{LZ} correction term of Eq. 3.3.1.

The incorporation of image-charge corrections in the treatment of charged supercell has been a debatable issue in the last decade.[168,169,172] For the purpose of our calculations, we will be using a modified version of Makov and Payne (MP)[173] corrections described in the Lany and

Zunger review.[174] In the original MP review, the fictitious interactions are corrected by using a multipole expansion of the electrostatic potential energy, due to a charge density $\tilde{n}(\vec{r})$ of the defect contained in the supercell. [173] In such expansion, the first order term (monopole) and the third order term (quadrupole) dominate the interaction energy as:

$$\Delta E_{MP} = \underbrace{\frac{q^2\alpha}{2\varepsilon\left(V_{SC}\right)^{1/3}}}_{Monopole} + \underbrace{\frac{2\pi q}{3\varepsilon V_{SC}}\int r^2 \cdot \Delta\tilde{n}_h(\vec{r})d\vec{r}}_{Quadrupole}, \qquad (3.3.2.2.1)$$

where $\Delta\tilde{n}_h(\vec{r}) = \tilde{n}_{h+d}(\vec{r}) - \tilde{n}_h(\vec{r})$

The first component (monopole interactions) is the Madelung energy describing the electrostatic interaction due to periodically repeated point charges interacting with a uniform background. In this case, α is the crystal structure-dependent Madelung constant and V_{SC} is the volume of the supercell. In the derivation of Eq. 3.3.2.2.1, Makov and Payne took into account the simplest type of screening effects of the host lattice by scaling the interaction terms with the macroscopic dielectric constant ε.

Following Oba's procedure, one can also use Madelung's corrections for the case of neutral shallow defects.[175] This is because a supercell cannot encompass a shallow defect wave function, and in a neutral charge state, a carrier (electron or hole) occupies the CBM which becomes a delocalized perturbed host state. This leads to artificial interactions, similar to those for a charged defect in a compensating charge density, pushing shallow transition levels deeper into the bandgap. For this reason, in the literature, somewhat deep transition levels (several hundred meV) are sometimes reported for cases of shallow defect states.

The second term (quadrupole interactions) from the above equation describes the interaction between the uniform compensating background and the electron density difference between the host with defect and pure-crystal lattice. However, for realistic cases of defects in

solids, MP corrections are found not to always improve the convergence of the formation energy in function of the size of the supercell.[176,177,178] Furthermore, the exact computation of the quadrupole term is quite complex because of complications encountered in defining $\Delta\tilde{n}(\vec{r})$. Following the Lany and Zunger approach, $\Delta\tilde{n}(\vec{r})$ is computed from direct DFT calculations of the differences in total charge densities between the charged and neutral system.[174] From benchmark numerical tests performed on selenium substituting arsenide in GaAs bulk, Ref. [174] found that beyond the immediate neighborhood of the defect, the charge density difference is dominated by the delocalized part of the defect density expressed as:

$$\Delta\tilde{n}\left(\vec{r}\right)_{screened} = \frac{1}{V_{SC}}\left[q(\vec{r}) - \frac{q(\vec{r})}{\varepsilon} \right] \qquad (3.3.2.2.2)$$

The logic behind using the above expression is explained in Ref. [204] as such:

- Upon incorporation of a defect with total charge $q(\vec{r})$ in the host lattice, the *localized* charge $q(\vec{r})$ electrostatically attracts a screening charge $-\frac{q(\vec{r})}{\varepsilon}$ yielding a potential of $u(\vec{r}) = -\frac{q(\vec{r})}{\varepsilon\left|\vec{r}-\vec{r}'\right|}$ where $\left|\vec{r}-\vec{r}'\right|$ is the distance between the screening charge and the localized charge *q*. As a result, the total amount of the screening charges is computed as:

  $$\sum q_{screen}(\vec{r}) = q(\vec{r}) - \frac{q(\vec{r})}{\varepsilon}$$

  For defects in solids subjected to PBC, the screening charges must originate from the host lattice. This *consequently changes the average charge density far from the defect from its bulk value by an amount* of $\Delta\tilde{n}\left(\vec{r}\right)_{screened} = \frac{1}{V_{SC}}\left(q - \frac{q}{\varepsilon} \right)$. More details regarding the derivation of the average density difference applied to the case of diamond vacancy in the 2- charge state is discussed in Ref. [204].

Now that $\Delta\tilde{n}(\vec{r})$ is represented by a simple expression, the once complex quadrupole term of the MP corrections (Eq. 3.3.2.2.1) can be expressed in function of the Madelung corrections as:

$$\frac{2\pi q}{3\varepsilon V_{SC}}\int r^2\cdot\Delta\tilde{n}_h(\vec{r})d\vec{r} = f\cdot\frac{q^2\alpha}{2\varepsilon\left(V_{SC}\right)^{1/3}}$$

where *f* is a proportionality constant that depends on the shape of the supercell. Hence, following the Lany-Zunger[174] approach, the previously derived MP correction is re-written as:

$$\Delta E_{LZ} = \frac{q^2\alpha}{2\varepsilon\left(V_{SC}\right)^{1/3}} + f\cdot\frac{q^2\alpha}{2\varepsilon\left(V_{SC}\right)^{1/3}} \qquad (3.3.2.2.3)$$

Although the LZ scheme has been successfully applied to various practical systems, one of its potential drawbacks is that it does not yield the exact energy corrections for the special cases of point charges as it was initially derived in the MP review.[179]

3.3.2.2. Potential alignment (PA) correction for neutral and charged supercells

In the previous section, for the neutral case, we have discussed that the otherwise divergent electrostatic potential ($\vec{G}=0$) is set to zero in reciprocal space. Consequently, the *spectrum* of HSE06 eigenvalues is defined up to an unknown constant which depends on the average potential of the supercell and the choice of the pseudopotentials.[42] In the case of charged systems, the eigenvalues will also shift by the same unknown constant since the KS eigenvalue is related to the change in total energy with respect to occupation numbers (Janak's theorem[134], see section 2.7.1). In order to obtain consistency in the potentials, we decide to examine the potential in the supercell far from the impurity and align it with the average electrostatic potential of the pure host crystal. Such alignment gives rise to a shift or potential alignment ΔE_{PA} which is expressed as:

$$\Delta E_{PA} = q\left[V(D^q) - V(0)\right]_R \qquad (3.3.2.2.1)$$

Here $\left[V(D^q) - V(0)\right]_R$ is the difference of potentials between the host + defect and pure-crystal lattice at a specific reference point *R*. However, it has been reported[180] that taking into account *PA* corrections may overestimate the corrections of fictitious interactions within the supercell method when applied conjointly with electrostatic interaction corrections. As a result, the *PA* correction has not been used in several defects studies and still remains a controversial subject.[181,182,183,184]

3.4. Defects Levels

3.4.1. Thermodynamic Transitions

The incorporation of neutral and charged impurities in any material plays a major role in its electrical and optical properties. Upon changing the charge state, the defect undergoes two types of transitions:

- Thermodynamic transition:

Thermodynamic transitions occur when a defect D in its q_1 charge state changes into the q_2 charge state, given that for each charge state, the atomic structure corresponds to its relaxed atomic configuration. The Fermi-level position at which two different charge states q_1 and q_2 have equal formation energy corresponds to the thermodynamic transition level $\varepsilon_T(q_1/q_2)$ which is defined as:

$$\varepsilon_T(q_1/q_2) = \left.\frac{E_f\left[D^{q_1}\right]_{\{q_1\}} - E_f\left[D^{q_2}\right]_{\{q_2\}}}{q_2 - q_1}\right|_{\Delta E_F=0}, \qquad (3.4.1.1)$$

when referenced from the VBM $(\Delta E_F = 0)$. Here the formation energies are calculated from the relaxed atomic configuration $\{q_i\}$ of its corresponding q-ith charge state. Thermodynamic transitions typically occur on the *phonons* time scale[185] and also correspond to thermal ionization energies of defects that can be observed in PL experiments. More details regarding the correlation between experimentally observed ionization energies and the theoretically predicted $\varepsilon_T(q_1/q_2)$ are given in sections 3.4.2.2. Here, $\varepsilon_T(q_1/q_2)$ are not to be confused with the KS eigenvalues derived from Janak's theorem discussed in section 2.6.1.

By re-writing $\varepsilon_T(q_1/q_2)$ in function of total energies instead of formation energies, we obtain:

$$\varepsilon_T(q_1/q_2) = \left.\frac{E_f\left[D^{q_1}\right]_{\{q_1\}} - E_f\left[D^{q_2}\right]_{\{q_2\}}}{q_2 - q_1}\right|_{\Delta E_F = 0}$$

$$\Leftrightarrow \varepsilon_T(q_1/q_2) = \frac{1}{q_2 - q_1}\left[E_{tot}\left[D^{q_1}\right]_{\{q_1\}} - E_{tot}[bulk] + q_1\cdot(E_{VBM}) - \sum_p^P n_p\mu_p + \left[\Delta E_{PA} + \Delta E_{LZ}\right]_{\{q_1\}}\right.$$
$$\left. - E_{tot}\left[D^{q_2}\right]_{\{q_2\}} + E_{tot}[bulk] - q_2\cdot(E_{VBM}) + \sum_p^P n_p\mu_p - \left[\Delta E_{PA} + \Delta E_{LZ}\right]_{\{q_2\}}\right]$$

$$\Leftrightarrow \varepsilon_T(q_1/q_2) = \left[E_{tot}\left[D^{q_1}\right]_{\{q_1\}} - E_{tot}\left[D^{q_2}\right]_{\{q_2\}}\right]\frac{1}{q_2 - q_1} - (E_{VBM}) + \frac{\left[\Delta E_{PA} + \Delta E_{LZ}\right]_{\{q_1\}} - \left[\Delta E_{PA} + \Delta E_{LZ}\right]_{\{q_2\}}}{q_2 - q_1}$$

Given that $q_2 = q_1 + 1e^-$, the above equation is written as:

$$\varepsilon_T\left[q_1/\left(q_1 + 1e^-\right)\right] = \left[E_{tot}\left[D^{q_1}\right]_{\{q_1\}} - E_{tot}\left[D^{q_1+1e^-}\right]_{\{q_1+1e^-\}}\right]$$
$$- E_{VBM} + \left[\Delta E_{PA} + \Delta E_{LZ}\right]_{\{q_1\}} - \left[\Delta E_{PA} + \Delta E_{LZ}\right]_{\{q_1+1e^-\}} \qquad (3.4.1.2)$$

Here the first component of the above equation corresponds to the ionization potential (*IP*) previously derived in section 3.4.2.2. Therefore the thermodynamic transition level is related to the KS eigenvalues via the equation:

$$\varepsilon_T\left[q_1/\left(q_1 + 1e^-\right)\right] \approx \frac{\varepsilon_I[M] + \varepsilon_I[M+1]}{2} - E_{VBM} + \left[\Delta E_{PA} + \Delta E_{LZ}\right]_{\{q_1\}} - \left[\Delta E_{PA} + \Delta E_{LZ}\right]_{\{q_1+1e^-\}}$$

(3.4.1.3)

In section 1.3, we have discussed that deep defects have localized wave functions which give rise to defect levels far from band edges. Such spatially confined states can assimilate multiple electrons depending on the electronic structure of the defect state. The occupation of these localized states by several electrons gives rise to positive Coulombic repulsion between the electrons.[185] A direct correlation between the electronic repulsion in deep localized states and the formation energies of the defect is schematically shown in Fig. 6. Following Ref. [185], in order

to differentiate between the lattice and the electronic contributions to the changes of the formation energies of the defect, we fix the (–) and (+) charge states of the defect into the atomic configuration of the defect's neutral charge state. By not allowing atomic distortion of either the (+) and (–) charge state of the defect, we can deduce that any resulting variations in the formation energy (or thermodynamic transition level) would be a direct consequence of the variations in the electronic structure of the defect's state. As shown in Fig. 6, dropping/removing electrons into the defect state while the system is kept in the atomic structure of the neutral charge state would raise the formation energy of the (+) and (–) charge states, which consequently shifts the position of the thermodynamic transition level in the bandgap. Such shift is an immediate result of the formerly discussed positive electronic Coulombic repulsion between electrons in the defect state and is called the $+U$ electronic parameter ($U_{Coulomb}$).

If one allows atomic distortions to take place in the (+) and (–) charge states, the formation energies would become lower due to the negative value of the relaxation energy and the thermodynamic transition level would subsequently shift to new Fermi-level positions in the bandgap.

Nevertheless, if the addition/removal of electrons in the deep defect state is accompanied with substantial lattice relaxation, a new situation arises. In such case, the lattice relaxation energy becomes larger than the electronic U repulsion which gives rise to a negative-U *potential energy* calculated as:

$$-U = U_{Coulomb} - U_{relaxation} \text{ given that } U_{relaxation} > U_{Coulomb} \quad (3.4.1.4)$$

In other words, the relaxation around the defect is significant enough that it overcomes the positive electrostatic repulsion and creates "attraction" between electrons located in the defect state. This consequently means that the system can transition from charge state q_1 (represented

by (+) charge state in Fig. 6) to another charge state $q_3 = q_1 \pm 2e^-$ (denoted by (–) charge state in Fig. 6), implying that the in-between charge state q_2 (displayed as the 0 charge state in Fig. 6) becomes *thermodynamically unstable*.[190] As a result the $-U$ potential energy can be computed as:

$$-U = \varepsilon_T\left(q_3 / q_2\right) - \varepsilon_T\left(q_1 / q_2\right) \tag{3.4.1.5}$$

More details regarding the negative-U parameter is given in Ref. [185].

- Optical transition:

Here the transition occurs rapidly enough so that the defect in the initial charge state q_1 is frozen and does not have time to relax into the relaxed configuration of the q_2 charge state. Such fast transition corresponds to optical transitions and is described in more thorough details in the following section.

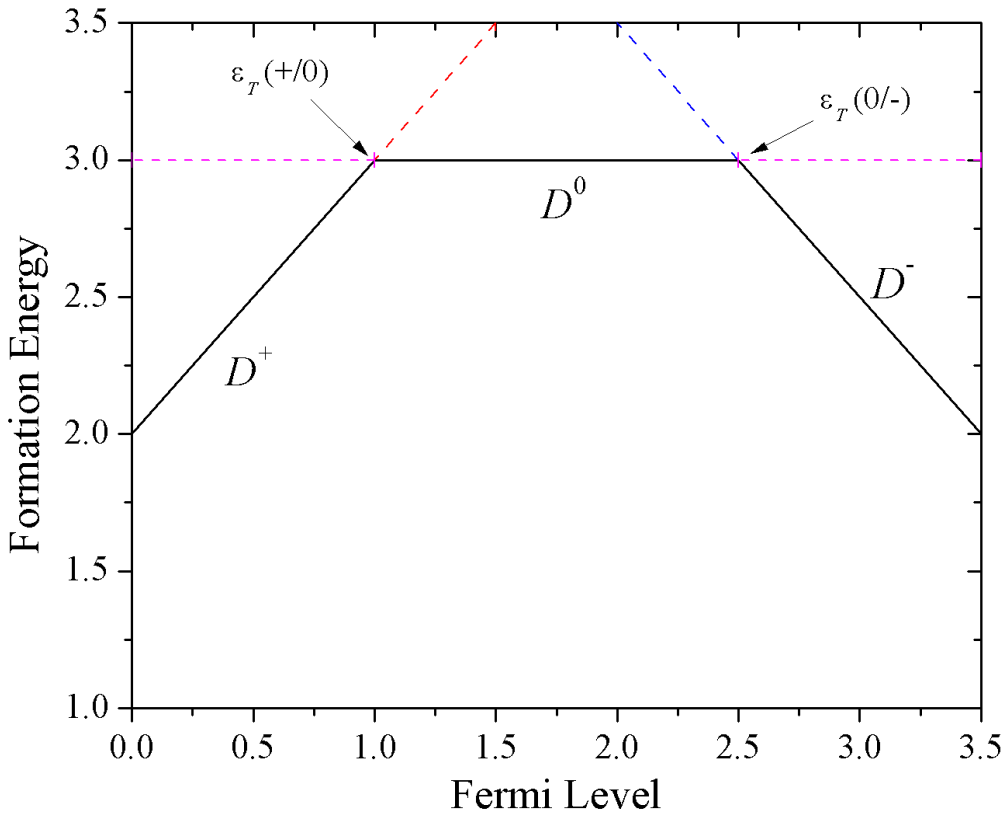

Figure 5: Schematic representation of the formation energy as a function of the Fermi level of a defect *D* according to Eq. 3.3.1. Here, the zero and maximum of the Fermi level axis correspond to the VBM and the CBM, respectively. The solid lines correspond to the formation energies for the most stable charge states of the defect *D*, while the dashed lines correspond to the higher energy charge states. The points where each line changes slope denotes the thermodynamic transition levels in the band gap. In *n*-type GaN (Fermi levels close to the CBM), the defect *D* behaves as a deep acceptor (negative charge state) with acceptor transition level at $\varepsilon_T(-/0)$. For the Fermi level closer to the VBM, the defect acts as a deep donor (positive charge state) defect with $\varepsilon_T(0/+)$.

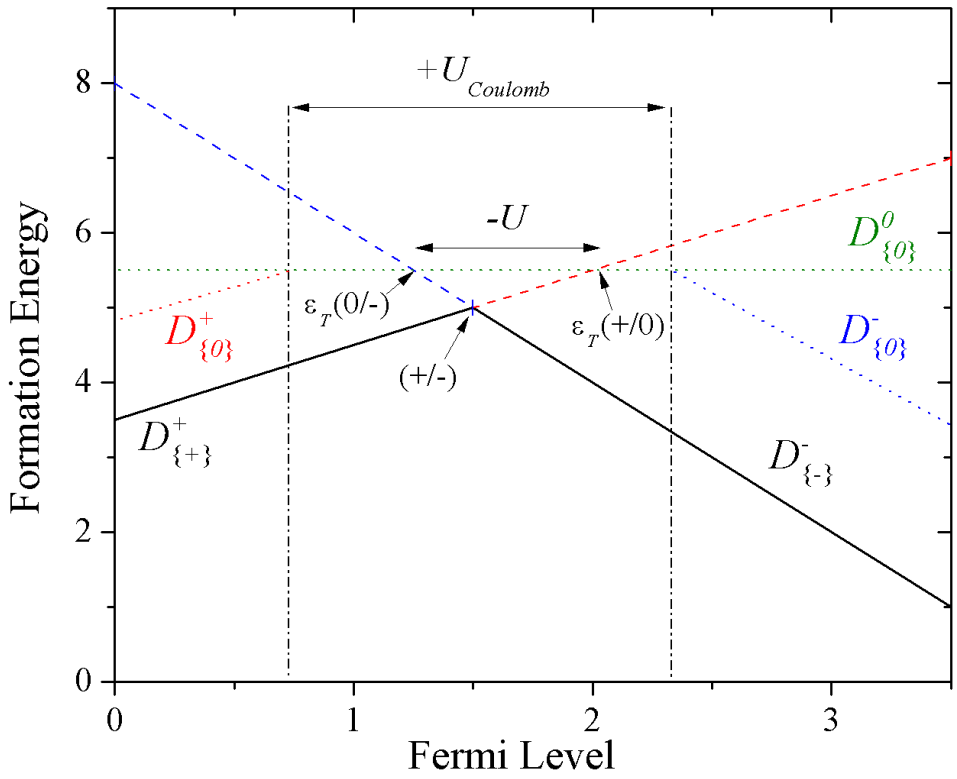

Figure 6: Schematic representation of the formation energy in function of the Fermi level for a negative-*U* behavior of a defect *D* based on Fig. 7 of Ref. [185]. $D^{q_i}_{\{q_j\}}$ denotes the defect *D* in its corresponding *q*-ith charge state while $\{q_j\}$ corresponds to the equilibrium atomic configuration of its *q*-jth charge state. The dotted lines correspond to the formation energy of the defect *D* in the frozen atomic configuration $\{0\}$ of the neutral charge state D^0. The dashed lines correspond to the higher formation energies of $D^+_{\{+\}}$ and $D^-_{\{-\}}$. Here $+U_{Coulomb}$ and $-U$ describe the positive electrostatic repulsion within the defect state and the negative *U* parameter, respectively. In the (-) charge state, the lattice relaxation is large enough that it overcomes the positive electrostatic Coulombic repulsion and makes the formation energy of the defect in the (-) charge state lower than in the neutral state (see Eq. 3.4.1.1). As a result, the in-between neutral state becomes unstable and the system transitions from the positive (+) charge state to the (-) charge state via the +/- crossover.

3.4.2. Optical Transition Levels and the Configuration Coordinate Diagram (CCD)

One of the most common approaches to describe optical transitions between different charge states of defects is the configuration coordinate diagram (CCD)[186,187]. The CCD used in this work is based on the harmonic approximation and describes the total energies of the system (*electronic energy* + *lattice energy*) containing defects as a function of a generalized configuration coordinate (schematically shown in Fig. 7). In such approximation, the potential curves are assumed to be simple parabolas and the configurational coordinate corresponds to a one dimensional mapping of the coordinates of a defect with its surrounding lattice in a three dimensional host lattice. The total energy *U* of the impurity in its ground (*g*) and excited (*e*) electronic states can be written as:

$$\begin{cases} U_g(r) = \frac{1}{2} k_g (r - r_A)^2 + U_g(r_A) & (3.4.2.1) \\ U_e(r) = \frac{1}{2} k_e (r - r_C)^2 + U_e(r_C) & (3.4.2.2) \end{cases}$$

Here, k_g and k_e are the generalized force constants (spring stiffness) of the ground and excited electronic states, respectively. The relaxed atomic configurations of the defect in its ground and excited electronic states are denoted by r_A and r_C.

In the ground state, the system is on the lower parabola $\left(U_g(r)\right)$ and oscillates around the minimum *A*. The analysis of possible optical transitions from the vibrational ground level of the lower parabola to the vibronic levels of the upper parabola is based on the Franck-Condon principle which is discussed in the following section.

3.4.2.1. Franck-Condon Approximation

In the Franck-Condon approach, the transitions between electronic states occur within a stationary nuclear framework. In other words, since the mass of electrons is much lighter than the nuclear masses, the nuclear configurations remain unchanged during absorption of a photon, and therefore, in the schematic diagram of Fig. 7 (or Fig. 8), it is represented by a vertical transition between points *A* and *B*. A more detailed justification of the vertical transition is based on the evaluation of the electric dipole moment operator between the ground vibrational state of the lower parabola and vibrational states of the upper parabola.

Let us assume that the overall ground electronic state $\left(\Psi_g^i\left(\vec{\tau},\vec{R}\right)\right)$ of the lower parabola corresponds to an electronic quantum number *g* and a vibrational quantum number *i* while the overall excited electronic state of the upper parabola $\left(\Psi_e^f\left(\vec{\tau},\vec{R}\right)\right)$ corresponds to an electronic quantum number *e* and a vibronic quantum number *f*. Here, $\vec{\tau}$ indicates both electronic spin and space coordinates while $\vec{R}$ denotes only nuclear coordinates. Within the Born-Oppenheimer approximation, the wave functions of these states are written as the product of the individual electronic and vibrational wave functions:

$$\begin{cases}\Psi_g^i\left(\vec{\tau},\vec{R}\right)=\psi_g\left(\vec{\tau},\vec{R}\right)\phi_{g,i}\left(\vec{R}\right)\\ \Psi_e^i\left(\vec{\tau},\vec{R}\right)=\psi_e\left(\vec{\tau},\vec{R}\right)\phi_{e,f}\left(\vec{R}\right)\end{cases}$$

The overall dipole moment operator $\left(\hat{D}\right)$ is characterized by the charges (+q) and nuclear positions (R_p) of the p-th nucleus, as well as the charges (-q) and the positions (r_m) of the m-th electron:

$$\hat{D} = q\sum_p R_p - q\sum_m r_m = \hat{D}_p + \hat{D}_m$$

where $\hat{D}_p$ and $\hat{D}_m$ are the corresponding nuclei and electronic dipole moment operators, respectively. The matrix elements of the overall dipole operator can therefore be written as:

$$\begin{aligned}
&\left\langle \Psi_e^f \middle| \hat{D} \middle| \Psi_g^i \right\rangle = \iint d\vec{\tau} d\vec{R} \left[\psi_e^* \left(\vec{\tau}, \vec{R}\right) \phi_{e,f}^* \left(\vec{R}\right) \left(\hat{D}_p + \hat{D}_m\right) \psi_g \left(\vec{\tau}, \vec{R}\right) \phi_{g,i} \left(\vec{R}\right) \right] \\
&\Leftrightarrow \left\langle \Psi_e^f \middle| \hat{D} \middle| \Psi_g^i \right\rangle = \iint d\vec{\tau} d\vec{R} \left[\psi_e^* \left(\vec{\tau}, \vec{R}\right) \phi_{e,f}^* \left(\vec{R}\right) \hat{D}_p \psi_g \left(\vec{\tau}, \vec{R}\right) \phi_{g,i} \left(\vec{R}\right) \right] \\
&\qquad + \iint d\vec{\tau} d\vec{R} \left[\psi_e^* \left(\vec{\tau}, \vec{R}\right) \phi_{e,f}^* \left(\vec{R}\right) \hat{D}_m \psi_g \left(\vec{\tau}, \vec{R}\right) \phi_{g,i} \left(\vec{R}\right) \right] \\
&\Leftrightarrow \left\langle \Psi_e^f \middle| \hat{D} \middle| \Psi_g^i \right\rangle = \int d\vec{R} \phi_{e,f}^* \left(\vec{R}\right) \hat{D}_p \phi_{g,i} \left(\vec{R}\right) \cdot \left[\cancelto{0}{\int d\vec{\tau} \psi_e^* \left(\vec{\tau}, \vec{R}\right) \psi_g \left(\vec{\tau}, \vec{R}\right)} \right] \\
&\qquad + \int d\vec{R} \phi_{g,i}^* \left(\vec{R}\right) \phi_{\mathrm{e,f}} \left(\vec{R}\right) \cdot \left[\int d\vec{\tau} \psi_e^* \left(\vec{\tau}, \vec{R}\right) \hat{D}_m \psi_g \left(\vec{\tau}, \vec{R}\right) \right]
\end{aligned} \tag{3.4.2.1}$$

Due to the orthonormality nature of the electronic orbitals in which:

$$\int d\vec{\tau} \psi_e^* \left(\vec{\tau}, \vec{R}\right) \psi_g \left(\vec{\tau}, \vec{R}\right) = \delta_{e,g},$$

the integral over the electronic coordinates in the first term of Eq. 3.4.2.1 is zero. Therefore the overall dipole operator can be simplified as:

$$\left\langle \Psi_e^f \middle| \hat{D} \middle| \Psi_g^i \right\rangle = \left[\int d\vec{R} \phi_{g,i}^* \left(\vec{R}\right) \phi_{e,f} \left(\vec{R}\right) \right] \left[\int d\vec{\tau} \psi_e^* \left(\vec{\tau}, \vec{R}\right) \hat{D}_m \psi_g \left(\vec{\tau}, \vec{R}\right) \right]$$

$$\Rightarrow \left\langle \Psi_e^f \middle| \hat{D} \middle| \Psi_g^i \right\rangle = D_{f,i} \left(\vec{R}\right) D_{e,g} \left(\vec{\tau}, \vec{R}\right), \tag{3.4.2.2}$$

$$\begin{aligned}
&\textit{where } D_{f,i} \left(\vec{R}\right) = \int d\vec{R} \phi_{g,i}^* \left(\vec{R}\right) \phi_{e,f} \left(\vec{R}\right), \\
&\textit{and } D_{e,g} \left(\vec{\tau}, \vec{R}\right) = \int d\vec{\tau} \psi_e^* \left(\vec{\tau}, \vec{R}\right) \hat{D}_m \psi_g \left(\vec{\tau}, \vec{R}\right)
\end{aligned}$$

The expression $D_{f,i}\left(\vec{R}\right)$ describes the overlap integral between the vibrational states in their respective electronic states *f* and *i*, while $D_{e,g}\left(\vec{\tau},\vec{R}\right)$ corresponds to the electric dipole moment of the electrons within the nuclear configuration $\vec{R}$.

Let us suppose that the electric dipole operator $D_{e,g}\left(\vec{\tau},\vec{R}\right)$ is weakly dependent on the nuclear coordinates, and subsequently perform a Taylor expansion of the matrix elements about the nuclear equilibrium position vector $\vec{R}$:[186]

$$D_{e,g}\left(\vec{\tau},\vec{R}\right)=D_{e,g}^{(0)}+D_{e,g}^{(1)}\cdot\vec{R}+\frac{1}{2!}D_{e,g}^{(2)}\cdot R^2+...$$

Within the Condon approximation, the matrix elements of $D_{e,g}\left(\vec{\tau},\vec{R}\right)$ are independent of the nuclear coordinates, given the atoms do not undergo large displacement from equilibrium. Therefore, the once complicated matrix $D_{e,g}\left(\vec{\tau},\vec{R}\right)$ is replaced by its zeroth order term (constant), yielding an approximated overall transition moment of:

$$\left\langle \Psi_e^f\left|\hat{D}\right|\Psi_g^i\right\rangle \approx D_{f,i}\left(\vec{R}\right)D_{e,g}^{(0)}=\left[\int d\vec{R}\phi_{e,f}^{*}\left(\vec{R}\right)\phi_{g,i}\left(\vec{R}\right)\right]D_{e,g}^{(0)}$$

According to the above equation, the optical transition moment is the largest between vibrational states that yield the highest overlap $D_{f,i}\left(\vec{R}\right)$. Therefore, the probability for optical transitions between different vibrational levels in their respective electronic states is proportional to the squares of the modulus of the vibrational wave functions of the initial and final states:

$$\left|\left\langle \Psi_e^f\left|\hat{D}\right|\Psi_g^i\right\rangle\right|^2=\left|\int d\vec{R}\phi_{e,f}^{*}\left(\vec{R}\right)\phi_{g,i}\left(\vec{R}\right)\right|^2\left|D_{e,g}^{(0)}\right|^2. \qquad (3.4.2.3)$$

More details regarding the Franck-Condon principle as a selection rule can be found in Ref. [186]. A schematic diagram shown in Fig. 7 shows a comparison of the transition probability

between vibronic states in different electronic states by arrow vectors $\overrightarrow{AB}$, $\overrightarrow{A'F}$ and $\overrightarrow{A''C}$. Here the $\overrightarrow{A'F}$ and $\overrightarrow{A''C}$ vertical transitions correspond to much lower values of the overlap integral between vibrational states and therefore yields lower transition probabilities. The vertical transition represented by $\overrightarrow{AB}$ corresponds to the maximum overlap integral and hence yields maximum transition probability between the vibrational state centered at *A* and the vibronic state peaking at the classical turning point *B*. Therefore, for a vertical transition $\overrightarrow{AB}$, a photon of energy equal to $E_{abs} = U_e(r_B) - U_g(r_A)$ is absorbed.

However, at the classical turning point *B*, the system is in non-equilibrium position, since it corresponds to one of the excited vibrational levels of the upper parabola. Subsequently, the system rearranges itself in order to minimize the interaction energy and gradually relaxes into the minimum *C* of the upper potential curve by emitting phonons:

$$E_e^{rel} = U_e(r_B) - U_e(r_C). \qquad (3.4.2.4)$$

Here E_e^{rel} corresponds to the relaxation energy of the excited electronic state. At the minimum of the upper potential curve (*C*), the system exists in its excited electronic state for a period determined by its lifetime. Eventually, the system transitions to the vibrational ground state of the lower potential curve (point *A*) radiatively or non-radiatively.

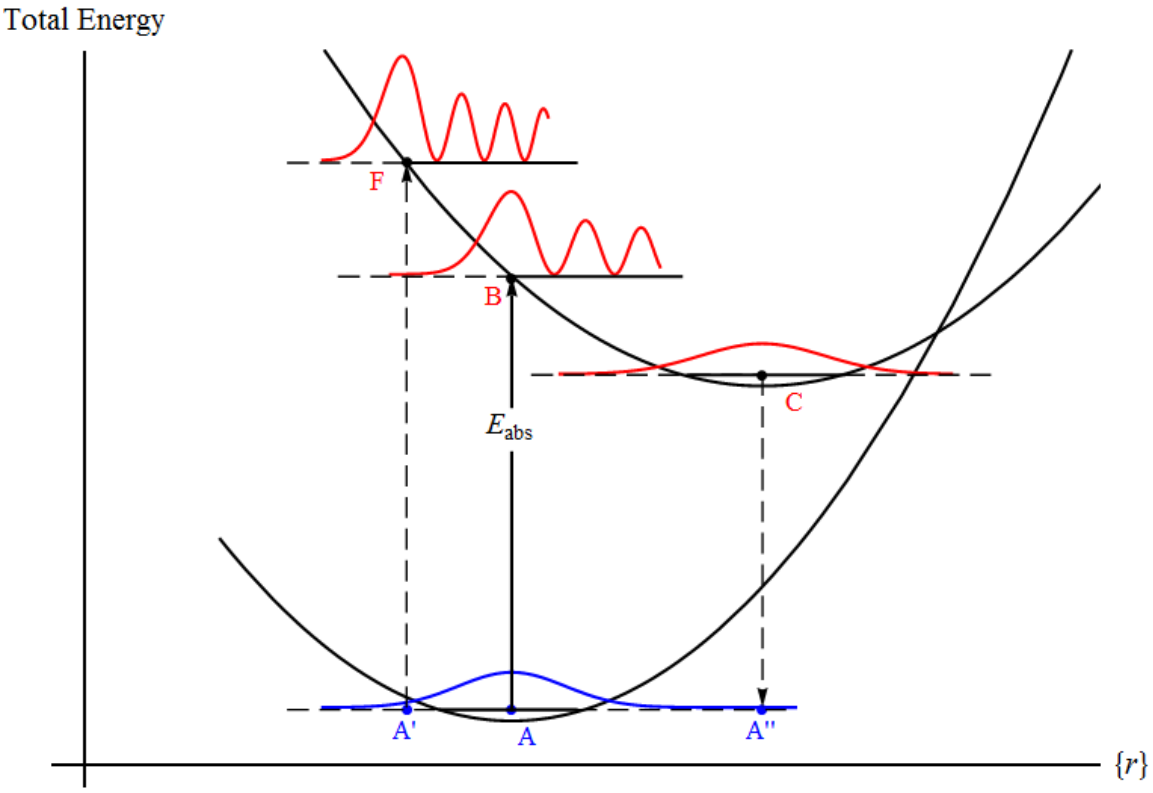

Figure 7: The quantum mechanical description of the FC approach. Only the lowest vibronic levels and corresponding wave functions are displayed. The vertical line shows the most probable optical transition (resonant excitation or E_{abs}) between the ground vibrational state of the ground electronic state and the excited vibrational state of the excited electronic state. There is a significantly less overlap between the vibrational wave functions of the ground electronic state and the vibrational state at points *F* and *C* which leads to lower probabilities for optical transition between the electronic states.

3.4.2.2. Non-radiative transitions using the CCD

The likelihood of a non-radiative transition can be estimated by finding the intersection (point *E*) of the two potential curves.[188,189] The energy of this intersection with respect to the vibrational ground state of the upper curve (excited state of the defect) represents the potential energy barrier given by:

$$E_b = U_e(r_E) - U_e(r_C), \tag{3.4.2.2.1}$$

for a non-radiative transition. Depending on the coherence lifetime and temperature of the sample for the resulting excited vibrational state (Fig. 8), the defect can either drop to point *D* of the lower parabola or transfers into the lower parabola at the crossover point *E*. In the former case, the recombination is radiative. When the barrier E_b is too high for thermal activation, or the lifetime of the excited state is short, (for example, due to high electron concentrations in *n*-type GaN) the system drops to point *D* of the lower parabola by emitting a photon of energy equal to $PL_{\max} = U_e(r_C) - U_g(r_D)$. Since point *D* corresponds to a non-equilibrium position of the ground electronic state, the system re-arranges its atomic configuration and relaxes to point *A* by emitting phonons with energies adding up to the FC shift:

$$FC = U_g(r_D) - U_g(r_A) = S_R \hbar \omega_g^i . \tag{3.4.2.2.2}$$

Here ω_g^i denotes the vibrational frequency of the ground electronic state (*g*) and the dimensionless term S_R is called the Huang-Rhys factor for radiative recombination. The parameter S_R corresponds to the mean number of phonons emitted from point *D* to *A*. In other words, S_R describes the strength of electron-phonon coupling between the defect and its host lattice. The stronger the electron-phonon coupling is, the larger the parameter S_R becomes and the greater is the shift between the equilibrium minima of the respective electronic

states $\Delta\vec{r}_{CA} = \vec{r}_C - \vec{r}_A$. Following the PL_{max}, the zero-phonon line (ZPL) which corresponds to direct optical transition between two vibrational ground states of the defect can be calculated as:

$$ZPL = PL_{\max} + FC \qquad (3.4.2.2.3)$$

Furthermore, the ZPL can also be expressed in terms of the previously discussed thermodynamic transition levels where:

$$ZPL = E_g - \varepsilon_T\left(q_1 / q_2\right) \qquad (3.4.2.2.3.a)$$

Regarding Eq. 3.4.2.2.1, if the barrier height E_b is not too high compared to the temperature of the sample, and the photoluminescence lifetime is long, the system can transition to the lower curve non-radiatively, by thermally jumping over the barrier and emitting dissipating phonons E_{NR}:[188]

$$E_{NR} = U_e\left(r_E\right) - U_e\left(r_A\right) = S_{NR}\hbar\omega_e^f\,, \qquad (3.4.2.2.4)$$

where ω_e^f describes the frequency of the vibrational ground state (*f=0*) of the excited electronic state (*e*) and S_{NR} is the Huang-Rhys factor for non-radiative recombination. Here, the dimensionless S_{NR} describes the mean number of phonons emitted via the non-radiative path (NR) depicted by a curved arrow in Fig. 8. Furthermore, one can estimate the frequency of attempts $\left(\omega_{NR}\right)$ to jump over a thermal potential barrier of energy E_b and transfer into the lower parabola with the Boltzmann probability equation[66]:

$$\omega_{NR} = \omega_e^f \exp\left(-E_b / k_B T\right), \qquad (3.4.2.2.5)$$

where $\omega_e^f \sim 10^{13} s^{-1}$ is the typical phonon frequency at which the vibrational ground state (*f =0*) of the excited state (*e*) of the defect oscillates. The frequency ω_e^f can be computed more accurately if one can determine experimentally or theoretically the phonon energy E_e^f of the vibrational ground state of the excited electronic state. In other words, from the expression of the

energy of the vibrational ground state of a simple quantum harmonic oscillator, one can relate the phonon energy E_e^f with the phonon frequency ω_e^f:

$$E_e^f = \frac{\hbar}{2}\omega_e^f .$$

Then, $\omega_e^f = \frac{2E_e^f}{\hbar}$ or $\omega_e^f = \frac{4\pi E_e^f}{h}$,

where *h* is the Planck's constant.

Now that we have obtained the frequency of attempts to jump over E_b, we can calculate the time it takes for non-radiative recombinations (t_{NR}) to occur as:

$$t_{NR} = \frac{2\pi}{\omega_{NR}}$$

If t_{NR} is much smaller than the photoluminescence lifetime (t_{PL}), we estimate that the defect is likely to be non-radiative. In this case, the ratio of t_{PL} over t_{NR} would correspond exactly to the amount of quenching that the observed PL band would undergo. In the other hand, if $t_{NR} \gg t_{PL}$, radiative recombination should prevail.

Accurate predictions of radiative versus non-radiative transition rates require calculations of electron-phonon interactions, and are currently a subject of intensive development (see for example recent Refs. [190,191]). The simple method presented here based on the intersection point between two potential curves for different charge states within the harmonic approximation, has been successfully used for predicting radiative versus non-radiative transitions in F-centers in the alkali halides.[188,192,193,194] Nonetheless, this method has not been thoroughly tested in GaN, which is why here we only draw preliminary conclusions about radiative or non-radiative nature of recombination via common native defects, based on the excitation and barrier energies obtained from CCD.

Now that we have shown that the CCD provides a way to estimate whether or not defects are radiative, we will show in the next section how to construct a CCD using the HSE electronic structure calculations.

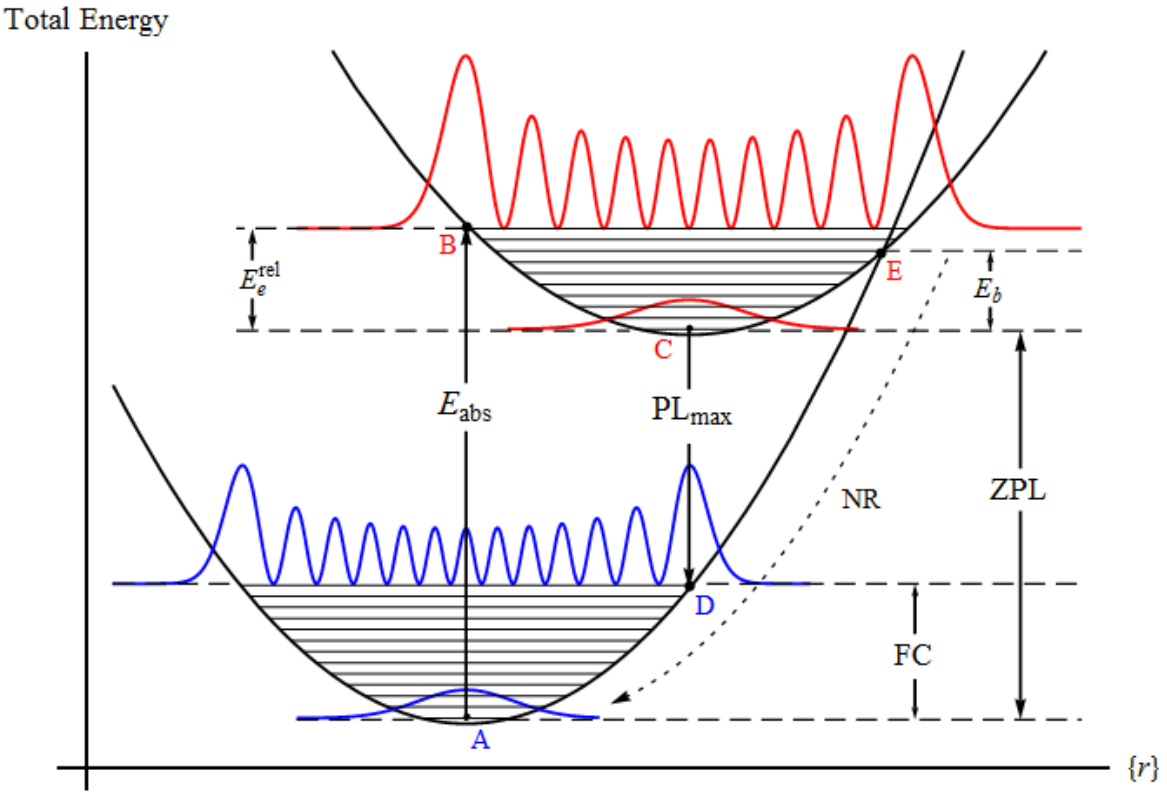

Figure 8: Schematic configuration coordinate diagram, displaying possible radiative and non-radiative transitions between the excited (*e*) and ground (*g*) electronic states of a defect. The potential curve of the excited state is vertically displaced from that of the ground state according to their formation energies and assuming the presence of an electron in the conduction band. The ZPL (zero-phonon line) describes the transition between the zero-point vibrational states in excited state and ground-state configurations. Here, E_{abs} denotes the resonant excitation energy (absorption energy) while PL_{max} corresponds to the peak of the PL band. E_e^{rel} $\Delta\epsilon_e$ and FC describe the relaxation energies of the excited state and the ground electronic states, respectively.

E_b denotes the energy barrier between the vibrational ground state of the upper curve and the crossover between the two curves. The dashed arrow labeled NR represents a non-radiative transition. The highest probability of occurrence for the vibrational ground state in the ground and excited electronic states occur at points *A* and *C*, respectively. Following the FC principle (see section 3.4.2.1), the vertical lines $\overrightarrow{AB}$ and $\overrightarrow{CD}$ correspond to the most probable optical transitions which respectively correspond to the absorption and photo-emission energies.

3.4.2.3. Construction of the CCD with the HSE06 formalism

Let us suppose we are investigating optical transitions of a deep defect (*D*) in GaN. By definition, the resonant excitation energy or absorption energy of a defect *D* is caused by the excitation of an electron $\left(+1e^{-}\right)$ from the defect level into the CBM. By using the formation energy formula (Eq. 3.3.1), the absorption energy (E_{abs}) is calculated as:

$$E_{abs} = \varepsilon_{Opt}\left[q_1 / \left(q_1 + 1e^{-}\right)\right] = E^f\left[D^{q_1+1}\right]_{\{q_1\}} - E^f\left[D^{q_1}\right]_{\{q_1\}} + E_g \qquad (3.4.2.3.1)$$

where $E^f\left[D^{q_1+1}\right]_{\{q_1\}}$ is the formation energy of the deep defect *D* in the q_1+1 state from the q_1 state atomic configuration and $E^f\left[D^{q_1}\right]_{\{q_1\}}$ is the formation energy of defect *D* in the q_1 charge state from the q_1 relaxed atomic structure. E_g is the amount of energy it costs to add one electron to the CBM during the transition which should correspond to the band gap of the system where the transition occurs. According to Eq. 3.4.2.2.3, the absorption energy expression can also be written in function of the previously discussed thermodynamic transition level as:

$$E_{abs} = ZPL + E_e^{rel} = E_g - \varepsilon_T\left[q_1 / \left(q_1 + 1e^{-}\right)\right] + E_e^{rel} \qquad (3.4.2.3.2)$$

As we have previously discussed, the atomic configuration of the newly obtained charged defect $\left[D^{q_1+1}\right]_{\{q_1\}}$ is not the most stable structure and loses the excess energy through phonons emission and therefore ends up in the relaxed structure of the q_1+1 state (point *C* in Fig. 8).

If radiative, the subsequent optical emission, which corresponds to the optical transition describing the recombination of the electron from the CBM (or shallow donor states in some cases) with the defect, is given by:

$$\mathrm{PL}_{\max} = \varepsilon_{Opt}\left[\left(q_1 + 1e^-\right)/q_1\right] = E^f\left[D^{q_1+1}\right]_{\{q_1+1\}} - E^f\left[D^{q_1}\right]_{\{q_1+1\}} + E_g \qquad (3.4.2.3.3)$$

where $E^f\left[D^{q_1}\right]_{\{q_1+1\}}$ is the formation energy of defect D in the q_1 state from the q_1+1 relaxed atomic structure and $E^f\left[D^{q_1+1}\right]_{\{q_1+1\}}$ is the formation energy of defect D in the q_1+1 charge state from the relaxed q_1+1 state atomic configuration. Maximum photoemission can also be calculated from the thermodynamic transition level as:

$$\mathrm{PL}_{\max} = ZPL + FC = E_g - \varepsilon_T\left[q_1/\left(q_1 + 1e^-\right)\right] + FC \qquad (3.4.2.3.4)$$

Now that we know how to calculate optical transition levels via the HSE method, let us further investigate possible computations of each component, i.e spring stiffness and equilibrium positions, of Eqs. 3.4.2.1 and 3.4.2.2 within the HSE06 method.

By substituting the expression of the potential energy of the ground electronic state (Eq. 3.4.2.1) into the equation defining the FC shift (Eq. 3.4.2.2.2), we obtain:

$$FC = U_g\left(r_D\right) - U_g\left(r_A\right)$$

$$\Leftrightarrow FC = \left[\frac{1}{2}k_g\underbrace{\left(r_D - r_A\right)^2}_{\Delta r_{DA}} + U_g\left(r_A\right)\right] - U_g\left(r_A\right)$$

$$\Leftrightarrow FC = \frac{1}{2}k_g\left(\Delta r_{DA}\right)^2$$

$$\Leftrightarrow \Delta r_{DA} = \left(\frac{2FC}{k_g}\right)^{\frac{1}{2}}, \qquad (3.4.2.3.3)$$

which is an equation of two unknown variables, Δr_{DA} and k_g. In order to calculate either unknown variables, we will insert the expression describing the potential energy of the excited electronic state (Eq. 3.4.2.2) into the equation describing the relaxation energy (Eq. 3.4.2.4):

$$E_e^{rel} = U_e\left(r_B\right) - U_e\left(r_C\right)$$

$$\Leftrightarrow E_e^{rel} = \left[\frac{1}{2}k_e\underbrace{\left(r_B - r_C\right)^2}_{\Delta r_{BC}} + U_e\left(r_C\right)\right] - U_e\left(r_C\right).$$

$$\Leftrightarrow E_e^{rel} = \frac{1}{2}k_e\left(\Delta r_{BC}\right)^2$$

$$\Leftrightarrow \Delta r_{BC} = \left(\frac{2E_e^{rel}}{k_e}\right)^{\frac{1}{2}} \qquad (3.4.2.3.4)$$

Since optical transitions are assumed to be vertical (FC principle), the displacement from the equilibrium position in both ground and excited electronic states are identical: $\Delta r_{BC} = \Delta r_{DA}$. As a result, the spring constants of both excited and ground electronic states are related by:

$$\left(\frac{2FC}{k_g}\right)^{\frac{1}{2}} = \left(\frac{2E_e^{rel}}{k_e}\right)^{\frac{1}{2}} \Rightarrow \frac{k_e}{k_g} = \frac{E_e^{rel}}{FC} \qquad (3.4.2.3.5)$$

The above equation describes the ratio between the stiffness of the excited and ground electronic states. Since neither spring constant's individual values can be obtained without the use of complex electron-phonon calculations, we can assign a random value to the spring constant of the ground (excited) electronic state and subsequently compute the value of the excited (ground) spring constant from the above equation. In other words, one can only obtain the ratio of the spring constants corresponding to the ground and excited electronic states. In fact, such inability to compute the individual values of the spring constant does not affect in any way the value of

the potential barrier (E_b) for NR transitions since at the intersection point (E) of the two potential curves:

$$U_e\left(r_E\right)=U_g\left(r_E\right) \Leftrightarrow \frac{1}{2}k_e\left(\Delta r_{EC}\right)^2+U_e\left(r_C\right)=\frac{1}{2}k_g\left(\Delta r_{EA}\right)^2+U_g\left(r_A\right)$$

$$\Leftrightarrow \frac{1}{2}k_e\left(\Delta r_{EC}\right)^2-\frac{1}{2}k_g\underbrace{\left(\Delta r_{EA}\right)^2}_{\Delta r_{BC}+\Delta r_{EC}}+\underbrace{U_e\left(r_C\right)-U_g\left(r_A\right)}_{ZPL}=0 \qquad (3.4.2.3.6)$$

However, since $E_b=U_e\left(r_E\right)-U_e\left(r_C\right)$

Then, $E_b=\left[\frac{1}{2}k_e\left(\Delta r_{EC}\right)^2+U_e\left(r_C\right)\right]-U_e\left(r_C\right)$

The above equation can be simplified as: $\Delta r_{EC}=\left(\frac{2E_b}{k_e}\right)^{\frac{1}{2}}$

By substituting the expression of Δr_{EC} into Eq. 3.4.2.3.6, we obtain:

$$\frac{1}{2}k_e\cdot\left(\frac{2E_b}{k_e}\right)-\frac{1}{2}k_g\left(\Delta r_{BC}+\Delta r_{EC}\right)^2+ZPL=0$$

$$\Leftrightarrow E_b-\frac{1}{2}k_g\left[\left(\Delta r_{BC}\right)^2+2\Delta r_{BC}\cdot\Delta r_{EC}+\left(\Delta r_{EC}\right)^2\right]+ZPL=0$$

$$\Leftrightarrow E_b-\frac{1}{2}k_g\left[\left(\frac{2E_e^{rel}}{k_e}\right)+2\left(\frac{2E_e^{rel}}{k_e}\right)^{\frac{1}{2}}\cdot\left(\frac{2E_b}{k_e}\right)^{\frac{1}{2}}+\frac{2E_b}{k_e}\right]+ZPL=0$$

$$\Leftrightarrow E_b-\frac{1}{2}\left[2\frac{k_g}{k_e}E_e^{rel}+4\frac{k_g}{k_e}\cdot\left(E_e^{rel}\cdot E_b\right)^{\frac{1}{2}}+2\frac{k_g}{k_e}E_b\right]+ZPL=0 \qquad (3.4.2.3.7)$$

From the above equation, we notice that the calculation of the potential barrier (E_b) only requires the value of the ratio of the spring constants and therefore is independent of the individual values of *k.* This consequently means that first-principles methods provide a valid way to construct the CCD accurately and hence can be used to estimate the probability of radiative versus non-radiative transitions of defects in GaN.

Now that we have provided a brief overview of the theoretical approach used for the analysis of defects in GaN, in the next section, we will describe the electronics and optical properties of native defects in GaN.

Section 4. Results

4.1. Theoretical investigation of Intrinsic Defects in GaN and their role in observed IR bands in electron-irradiated GaN samples

In section 1.4, we discussed the controversial issues regarding both the electronic properties of native defects in GaN and the microscopic sources of the observed IR bands in electron-irradiated GaN epilayers. In the following section, we shall perform a systematic study of the electronic and optical properties of common native defects, namely Ga vacancy (V_{Ga}), N vacancy (V_N), Ga-N divacancy ($V_{Ga}V_N$), interstitial Ga (Ga_i), Ga antisite (Ga_N), interstitial N (N_i), N antisite (N_{Ga}), the complex consisting of Ga interstitial and vacancy of Ga (Ga_iV_{Ga}), and the complex consisting of gallium antisite and vacancy of Ga (Ga_NV_{Ga}). We use the previously discussed exchange tuned HSE hybrid functional for the analysis of our native defects and compare our results to the most recent theoretical calculations and experimental observations.

4.1.1. Theoretical Methods

In order to study the electronic, structural and optical properties of native defects in GaN, we use the HSE06 method and the projector-augmented wave (PAW)[164] formalism as implemented in the VASP code.[195] Here, the Ga 3*d* valence electrons are not included in the PAW pseudopotentials. As typical for defect calculations in GaN, we adjusted the amount of exact exchange to 31% and the screening parameter is kept at typical 0.2 $Å^{-1}$. These parameters result in a band gap of 3.487 eV which is in a good agreement with the low temperature experimental value of 3.50 eV.[196] Calculated lattice parameters a = 3.210 Å, c = 5.198 Å and u = 0.377 Å for relaxed wurtzite GaN (see figure 9) are also in good agreement with experimental values.[197] Good convergence was achieved using the cutoff energy of 400 eV, the Γ-point only and hexagonal supercells containing 128 atoms. All structural relaxations were also performed within the same exchange tuned HSE to reduce forces to less than 0.05 eV/Å. Using larger supercells (up to 300 atoms) or denser **k**-point mesh (2×2×2), we estimate that above parameters produce errors of less than 0.05 eV in formation energies and transition levels.

In order to calculate the probability of defect formation in bulk GaN, we use the previously discussed concept of formation energy (cf. section 3.3) where:

$$E_f[D^q] = E_{tot}[D^q] - E_{tot}[bulk] + q\cdot(E_{VBM} + \Delta E_F) - \sum_p^P n_p\mu_p + \Delta E_{PA} + \Delta E_{LZ}$$

Although each the component of the above equation is discussed in details in section 3.3, the calculations of the elemental chemical potential of the *p*-th atom require careful attention. Here, the formation enthalpy $\Delta H(GaN)$ was calculated using total energies of a two-atom unit primitive GaN cell, orthorhombic metal Ga, and N_2 molecule, with volumes and atomic coordinates fully relaxed with HSE parameterization described above. The resulting

$\Delta H\left(GaN\right) = -1.249\ eV$ is in reasonably good agreement with previous theoretical calculations[36,72] and the experimental value of -1.34 eV reported recently.[198] In addition to the potential alignment correction $\left(\Delta E_{PA}\right)$ and Lany-Zunger corrections $\left(\Delta E_{LZ}\right)$[174], we are also using Madelung's corrections[175] for the case of neutral shallow defects. Details regarding each correction were given in section 3.3.2.1 and 3.3.2.2.

As previously discussed in section 3.3, formation energies cannot be expected to yield realistic defect concentrations due to the lack of entropic contributions and various other factors, which means that the values of defect formation energy should only be used as rough guidelines for defect formation. However, these complications do not affect the results of this work, since defect transition levels are calculated from formation energy differences.

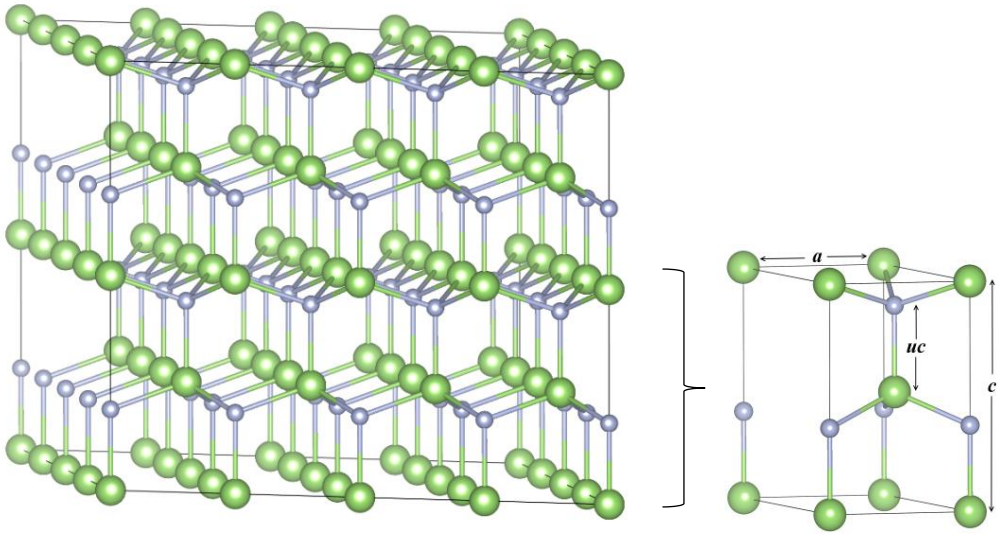

Figure 9: 128-atoms GaN supercell (left picture) with its corresponding wurtzite primitive cell (right picture) obtained with HSE lattice parameters of a = 3.210 Å, c = 5.198 Å and u = 0.377 Å. Small grey spheres represent N atoms and large green spheres represent Ga atoms.

4.1.2. Gallium vacancy (V_{Ga})

4.1.2.1. Formation Energy and Optical properties of V_{Ga}

Figure 10 (a, b) shows the formation energies of Ga vacancy in Ga-rich and N-rich growth conditions obtained from our HSE calculations. The slope of each line corresponds to a charge state of the defect, while the intersection points represent thermodynamic transition levels. We find that V_{Ga} exhibits both donor and acceptor properties, and four thermodynamic transition levels are predicted within the bandgap, namely the $\varepsilon_T(+/0) = 0.94$ eV, the $\varepsilon_T(0/-) = 1.73$ eV, the $\varepsilon_T(-/2-) = 1.87$ eV, and $\varepsilon_T(2-/3-) = 2.34$ eV, above the VBM. For Fermi levels close to the CBM (Fig. 10), Ga vacancies exhibit the lowest formation energy among all investigated native defects in all growing environments. This indicates that V_{Ga} can play a role as a compensating center in *n*-type GaN, which is consistent with previous experimental predictions.[43] In all other kinds of samples, i.e, high resistivity, *p*-type and compensated GaN, isolated vacancies of Ga exhibit fairly high formation energies and are unlikely to occur, unless created by electron irradiation.

While V_{Ga} has been widely discussed as a possible source of the yellow and green luminescence bands in GaN, the exact attribution of these PL bands is still not entirely settled. As we previously discussed in section 3.4.2, by fitting the CCD into calculated optical transitions and lattice relaxation energies, one can estimate whether the recombination via the defect is radiative or non-radiative. The CCD shown in Fig. 11 describes the optical transitions via the 2-/3- level of V_{Ga}. The lower potential curve (V_{Ga}^{3-}) is obtained by fitting a parabola into the two

calculated total energies of the V_{Ga}^{3-} in the relaxed defect lattices of the 3- and 2- charge states (this energy difference is the calculated Franck Condon shift of 0.54 eV). The upper potential curve is also obtained in similar manner, where a parabola is fitted into the calculated total energies of the V_{Ga}^{2-} in the relaxed defect lattices of the 3- and 2- charge states (the energy difference is the calculated relaxation energy of 0.46 eV). Note that in this CCD, all optical transitions are calculated directly from HSE, while the energy crossover for the non-radiative transition relies on the harmonic approximation (Eq. 3.4.2.2.1 and Eq. 3.4.2.3.7).

In order to demonstrate the validity of the harmonic approximation, we perform a linear interpolation between the two minima points in the 3- and 2- charge states. We subsequently compute HSE total energies of the intermediate geometries where the atomic configuration of each intermediate geometry is kept frozen (unrelaxed). We are not allowing atomic relaxation of the intermediate geometries because they would eventually relax into the ground state of V_{Ga}^{3-} (or V_{Ga}^{3-} in the 3- atomic configuration) making our computation quite trivial. As shown in Fig. 11, direct HSE calculations yields results that are very similar to the harmonic approximation, with average energy difference of 4 meV. In case of deep defect such as V_{Ga} where distortions tend to be large, the harmonic approximation still remains an accurate approach to describing the potential curves of the CCD.

Now that we have demonstrated that the CCD is a valid approach for the case of V_{Ga}, we calculated that the resonant excitation $V_{Ga}^{3-} \rightarrow V_{Ga}^{2-}$ is expected to have a maximum at 1.60 eV, which is 0.33 eV higher than the intersection of the two potential curves. Losing the excess energy through phonon emission, the system can either relax into the vibrational minimum of the V_{Ga}^{2-} state, or undergo a non-radiative transition to the ground state V_{Ga}^{3-}. In the former case, the

subsequent recombination of the hole localized on the vacancy and an electron from the CBM, returns the vacancy from V_{Ga}^{2-} to the V_{Ga}^{3-} state, with a PL maximum computed at 0.60 eV. This transition is then followed by lattice relaxation (FC shift) of 0.54 eV, resulting in a ZPL of 1.14 eV. However, based on the CCD (Fig. 11), the barrier for the non-radiative transition is 0.13 eV, suggesting that at room temperature the average time of the thermal jump over this barrier is several orders of magnitude shorter than a typical defect PL lifetime. Thus, the V_{Ga} is likely non-radiative.

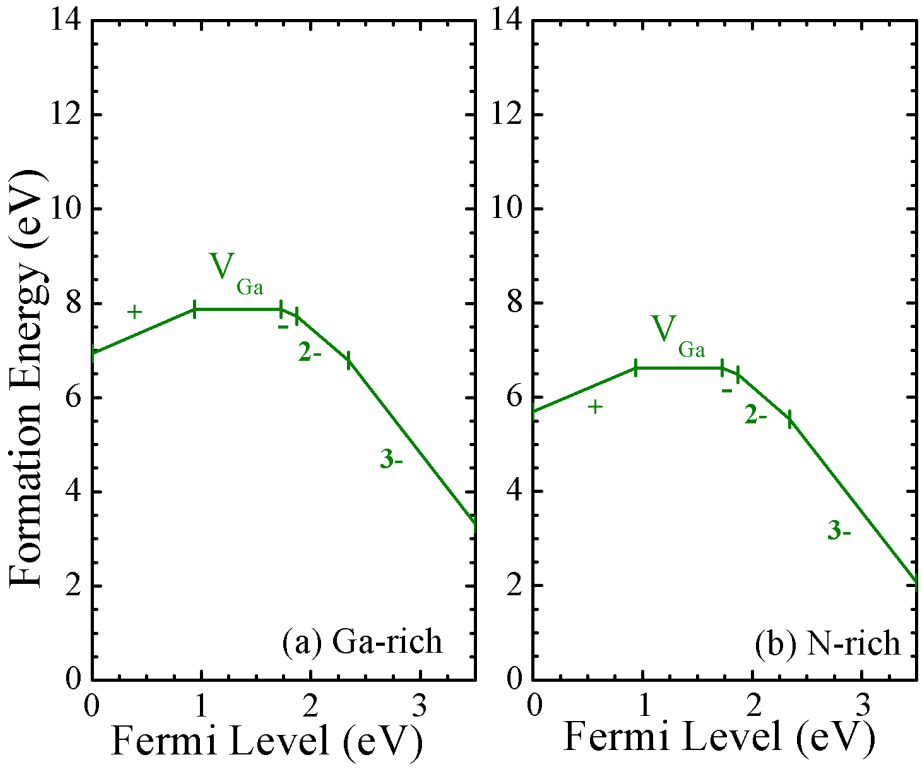

Figure 10: Formation energy of V_{Ga} as a function of the Fermi energy in (a) Ga-rich and (b) N-rich growth conditions. Ga vacancies display high formation energies in *p*-type and compensated GaN while it appears fairly energetically stable for Fermi level positions close to the CBM.

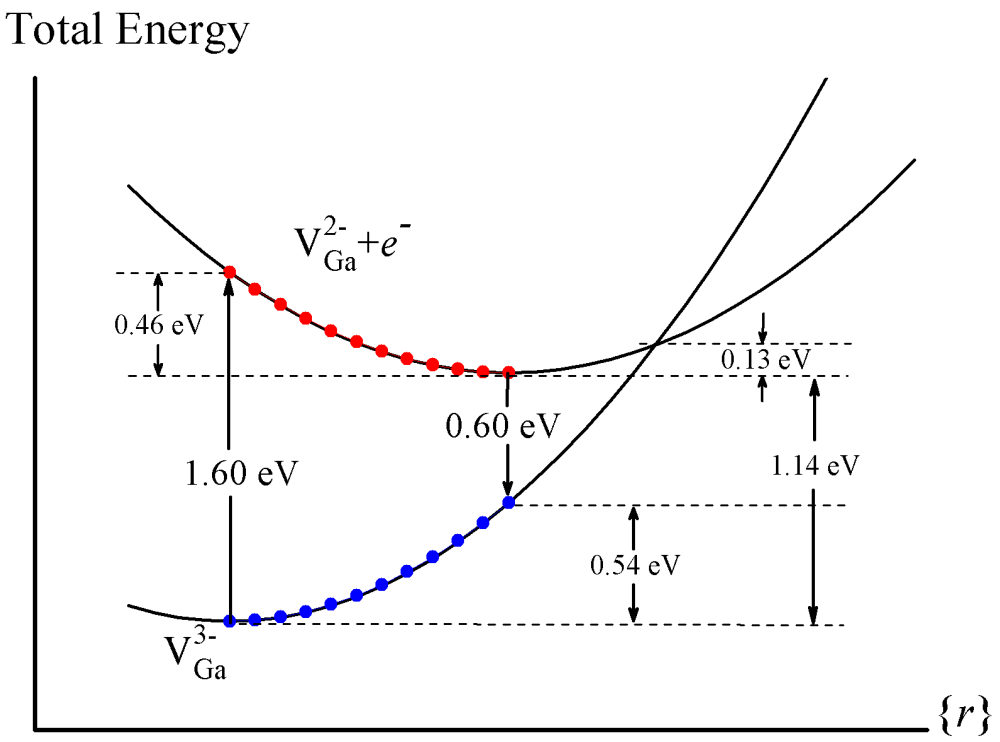

Figure 11: CCD of V_{Ga} obtained from the harmonic approximation fitting of total energies at relaxed defect lattices only (solid black lines), and direct HSE calculations (filled circles). The filled circles correspond to the total energies of ten intermediate defect geometries between two minima in the 3- and 2- charge states. An average difference in energy of 4 meV is found between the CCD based on the harmonic approximation and the CCD obtained from direct HSE calculations. A calculated emission of 0.60 eV, a FC shift of 0.54 eV and a ZPL of 1.14 eV are obtained. The energy barrier for a non-radiative transition is 0.13 eV, making the V_{Ga} likely non-radiative.

4.1.2.2. Atomic and Electronic Structure of V_{Ga}

HSE calculations show that in the singly positive charge state (+), the N atoms near the vacant Ga site, relax away from the vacancy by 12.3%, while with addition of electrons, the outward distortions decrease. For instance, in the 3- charge state (Fig. 12), the distances between neighboring N atoms and the vacant Ga site decrease by 9.38% (compared to the ideal bulk Ga-N bond length). Further calculations also show that removal of Ga atom in bulk GaN, i.e. breaking four bonds with nearest N atoms, introduces four defect levels within the bandgap.

The atomic structure and charge density of each of the four defect levels are displayed in Fig. 12. In this example, the V_{Ga} is in the 3- charge state, where all defect states are occupied by electrons. The Ga vacancies' defect states are linear combinations of nitrogen *p*-orbitals, which vary in the degree of localization. Each state displayed in Fig. 12 is degenerate with respect to spin. These defect states form four transition levels within the bandgap shown in Fig. 10.

The HSE calculated single electron energies and their changes with addition/removal of electrons to the Ga vacancy are shown in Fig. 13. For example, in the 3- charge state, the highest energy occupied defect states (spin-up and down) are located at 1.12 eV above the VBM. When V_{Ga} traps a hole, leading to 2- charge state of the defect, the highest spin-up state is shifted to 2.83 eV above the VBM. This, along with energy of accompanying atomic relaxations, results in the 2-/3- transition level occurring at 2.34 eV in Fig. 10. Another example is the spin-down defect state located at 1.29 eV above VBM in the 1- charge state (Fig. 11). Removing an electron from this defect state leads to the 0/- transition level at 1.68 eV in Fig 10.

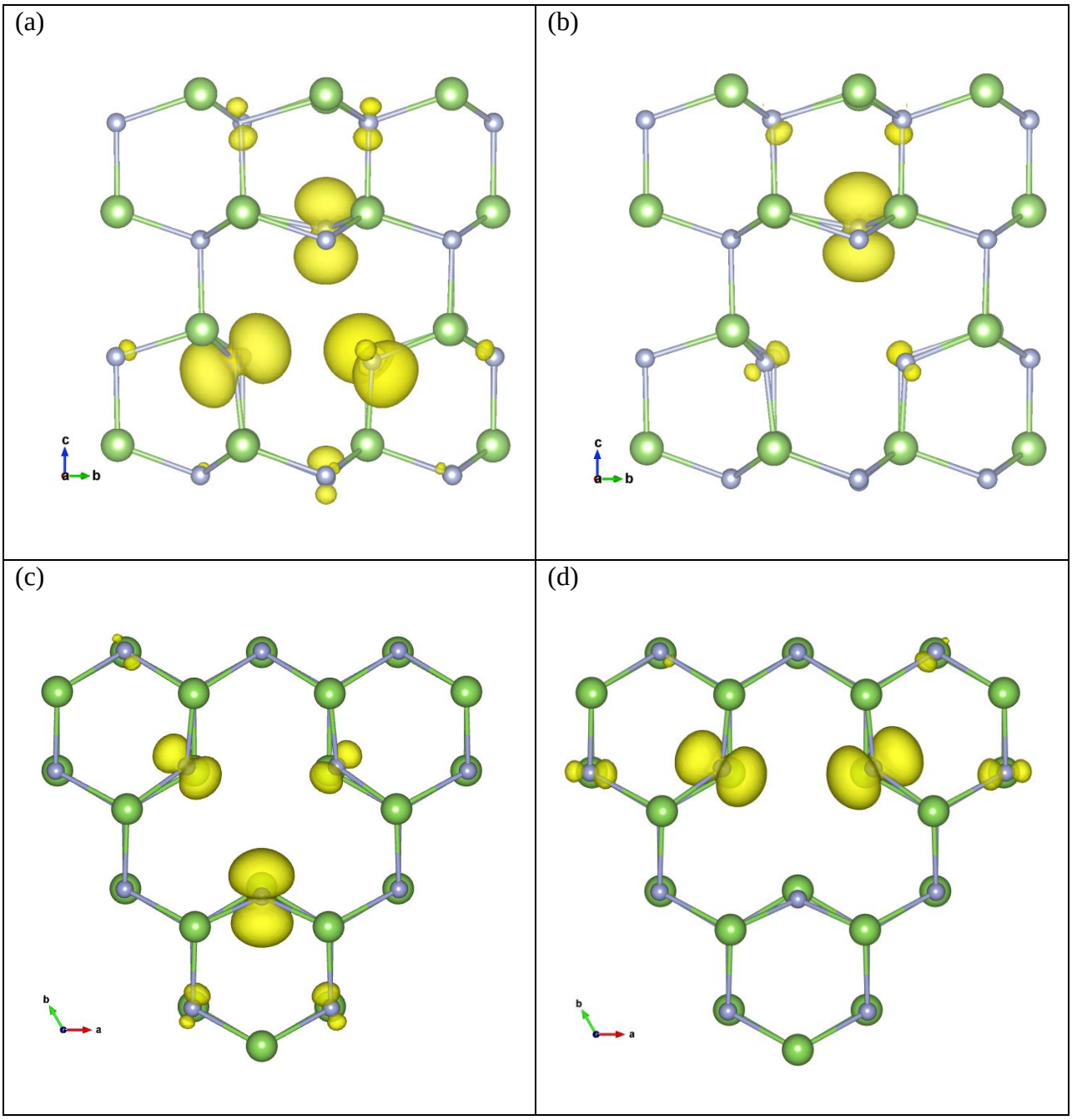

Figure 12: Charge density isosurfaces of the four defect states of V_{Ga}. The wave functions are calculated in the 3- charge state of the defect. For clarity, two different orientations are used for states (a, b) and (c, d), indicated by the lattice vectors. The (a)-(d) charge densities correspond to the eigenvalues shown in Fig. 13 (right panel, 3- charge state of V_{Ga}), from lowest (a) to highest

(d) energy. The small (grey) and large (green) spheres indicate the nitrogen and gallium atoms, respectively. The isosurface values are set at 5% of the maximum.

4.1.2.3. Magnetic properties of V_{Ga}

As shown in Fig. 13, unpaired spins of the electrons localized at the Ga vacancy lead to the local magnetic moment of V_{Ga}. When all defect states are occupied by electrons in 3- charge state, the charge density is equally distributed over four spin-up and four spin-down states, resulting in zero magnetic moment of the vacancy. Removing one electron, i.e. introducing one hole to the highest energy spin-up state leads to the magnetic moment of $1\,\mu_B$ for the 2- charge state of the defect. In this case, approximately 80% of magnetization is localized on one of the nearest nitrogen atoms. In the singly negative charge state, the magnetic moment is $2\,\mu_B$, with most of the magnetization density (about 90%) localized on the three neighboring nitrogen atoms, about $\sim 0.59\,\mu_B$ each. In this case, two of the four spin-down states are occupied by electrons (Fig. 13). Finally, in 0 and + charge states of V_{Ga}, the spin-down defect states have three and four localized holes, respectively, while in both cases all spin up states are occupied by electrons. This results in the vacancy magnetic moments of $3\,\mu_B$ and $4\,\mu_B$, in 0 and + charge states, respectively. In this case, calculations show that each of the four neighboring nitrogen atoms has local magnetic moments of $0.64\,\mu_B$ and $0.82\,\mu_B$ in 0 and + charge state respectively, and all local magnetic moments are ferromagnetically (FM) ordered. In both cases, the nearest neighbor nitrogen atoms provide around 85% of the total vacancy magnetic moment, while the remaining magnetization of the V_{Ga} comes from farther nitrogen atoms.

Large magnetic moments of V_{Ga} in + and 0 charge state raise questions about whether these ferromagnetic alignments of spins on neighboring N atoms are of the lowest energy. HSE calculations show that the antiferromagnetic (AFM) spin configuration on the four nearest N around the vacant Ga site in the + charge state is more energetically favorable than the FM

alignment by $\Delta E_f \approx 75$ meV. In AFM configuration, two of the spin-up and two of the spin-down defect states are occupied by electrons while the remaining two states of each spin have two localized holes, as shown in Fig. 13, leading to a net magnetic moment of $0\,\mu_B$. Comparison of the magnetization density of V_{Ga}^{+} in both AFM and FM spin configurations are displayed in Fig. 14. The defect states consisting of linear combination of *p*-orbitals localized at the four nearest N atoms create four aligned magnetic moments in the FM spin configuration (Fig. 14(a)). In the AFM configuration, each pair of the local magnetic moments is antiparallel (Fig. 14(b)). Similarly to FM case, local magnetic moments of each of the four neighboring N atoms are computed to be about $0.80\,\mu_B$, while the remaining magnetization mostly originates from next nearest N atoms. In contrast to the + charge state, the magnetic moment of $3\,\mu_B$ in the neutral charge state of V_{Ga} is more energetically favorable with respect to the possible AFM alignment by 47.5 meV. Thus, following Fig. 10, in *p*-type or high resistivity samples, V_{Ga} is predicted to have a magnetic moment of zero for Fermi energies up to 0.94 eV above the VBM, then a large $3\,\mu_B$ for Fermi levels between 0.94 eV and 1.73 eV, $1\,\mu_B$ for Fermi levels between 1.87 eV and 2.34 eV, and zero again for Fermi levels over 2.34 eV above the VBM.

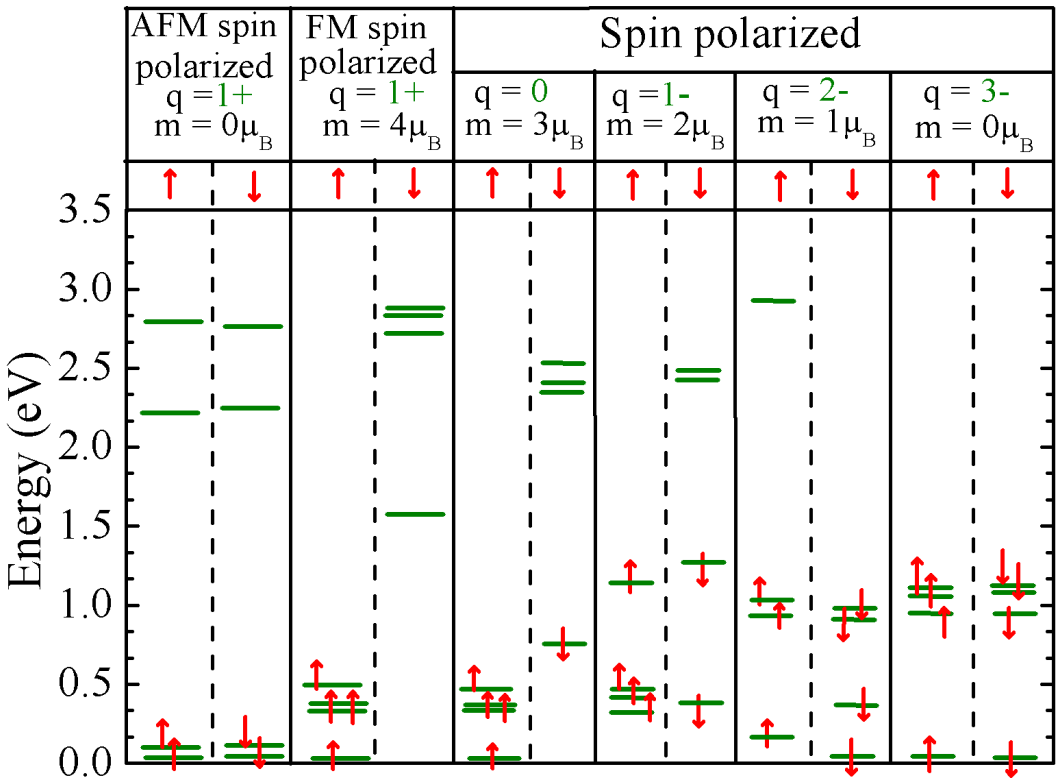

Figure 13: Single-electron energy levels of V_{Ga} for all the possible charge states q with their respective magnetic moments ***m***. Zero energy corresponds to the VBM.

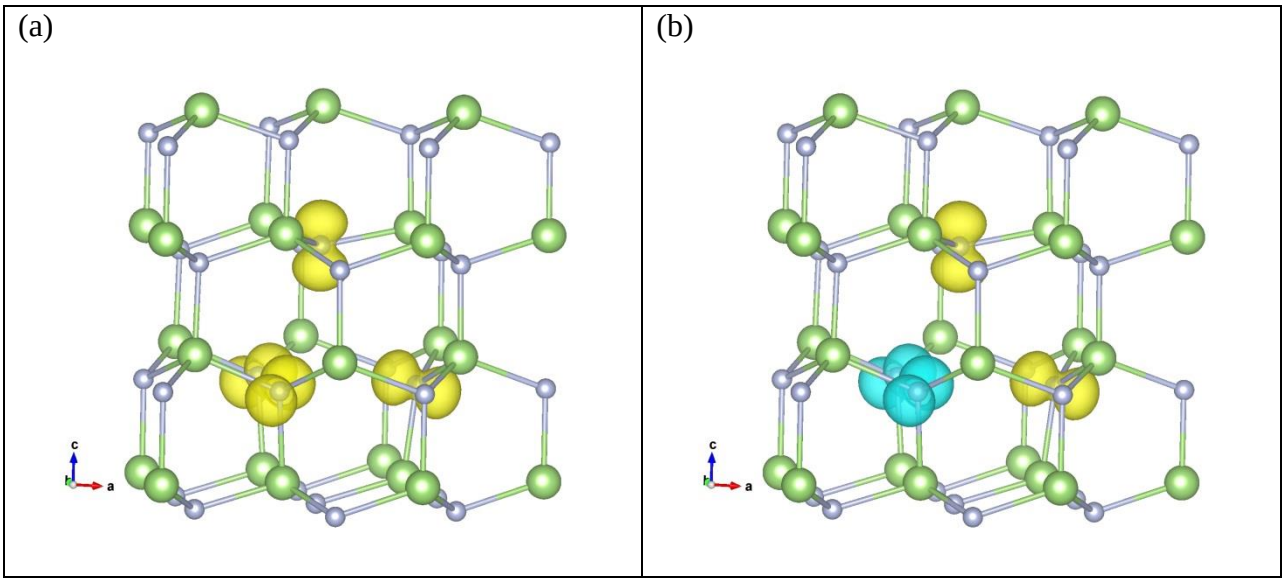

Figure 14: Magnetization density isosurfaces of V_{Ga} in the singly positive charge state in the (a) FM spin configuration and (b) AFM spin configuration. The positive (yellow) and negative (light blue) isosurface values are plotted at 10% of the maximum. AFM spin alignment has lower energy than FM alignment by 75 meV.

4.1.3. Nitrogen vacancy (V_N)

4.1.3.1. Formation energy and Electronic Structure of V_N

Figure 15 shows formation energies of the nitrogen vacancy (V_N) in the lowest energy charge states + and 3+. HSE calculated 2+/+ and 3+/2+ transition levels are located at 0.47 eV and 0.61 eV above the VBM, respectively. The formation energy of the 2+ charge state is always higher than the + and 3+ charge states, hence V_N exhibits properties of a negative-*U* center (with U = -0.13 eV, see section 3.4.1). The crossover 3+/+ occurs at 0.54 eV above the VBM, which is between the value of 0.47 eV reported in Refs. [72,76] and the value of 0.70 eV obtained in recent HSE calculations.[36] Overall, nitrogen vacancy is a donor defect with both deep and shallow levels. Relatively high formation energy of V_N for Fermi-level positions near the CBM indicates that nitrogen vacancies are unlikely to be an effective *n*-type source in GaN. However, in high resistivity and *p*-type samples, nitrogen vacancies can be a compensating defect.

A nitrogen vacancy in GaN introduces two nearly degenerate localized defect states within the electronic bandgap, and a weakly localized shallow donor state, which is too shallow to be accurately described in our supercell calculations. The charge density of one of the nearly degenerate localized defect states (computed in 3+ charge state of V_N) is displayed in figure 16. This defect state is a *spd*-hybridized orbital. The charge density strongly localized at the vacancy site consists of mostly *s*-character. The defect state also spreads to nearest Ga atoms (where it has 10% *s*-, 70% *p*-, and 20% *d*-character) and next nearest N atoms (80% *p*- and 20% *s*-character). The degeneracy of the two defect states along with large lattice relaxation causes the negative-*U* behavior of V_N. In the singly positive charge state, the Ga nearest neighbor along the *c*-axis relaxes away from the N vacant site by 2.15%, while the remaining three Ga neighbors

also undergo an outward relaxation of 1.60%, compared to ideal Ga-N bond length. In the +3 charge state (Fig. 16), the breathing relaxation is much larger, where the neighboring Ga atom moves away from the N vacant site by 22.1% (Ga atom along the *c*-axis) and 19.5% (atoms in Ga plane).

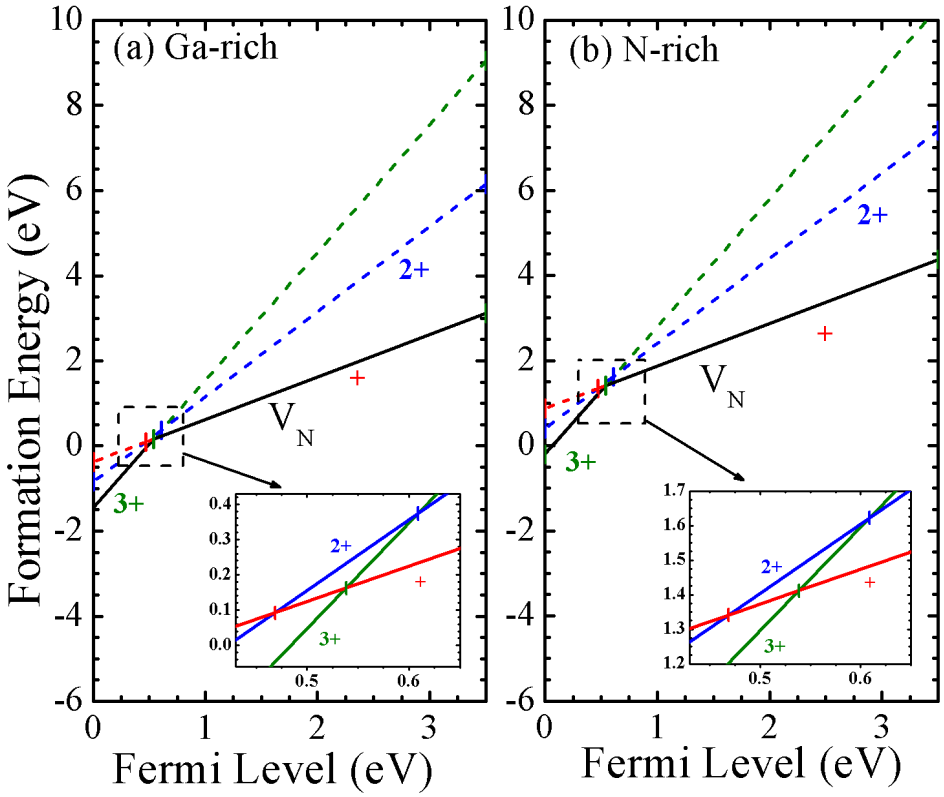

Figure 15: Formation energies of the V_N defect in GaN grown under (a) Ga-rich and (b) N-rich conditions. The dashed lines are used to emphasize the presence of negative-*U* behavior where $U = -0.13$ eV. The insets show the regions with the 2+/+ and 3+/2+ transition levels located at 0.47 eV and 0.61 eV above the VBM, respectively.

4.1.3.2. Optical Transitions Levels of V_N

A detailed experimental and theoretical analysis of the optical properties of V_N in bulk GaN has been previously published in Ref. [35]; here we briefly outline them to present a complete picture of native defects in GaN. The optical transition through the V_N 2+/+ level is internal, as previously suggested.[35] Initially, a positive V_N^+ ground state cannot efficiently trap a photo-generated hole, instead it traps an electron at the shallow +/0 level, making the defect overall neutral. It then traps a hole at a deep 2+/+ level, transferring the V_N into excited + charge state. The optical transition due to the recombination of the weakly localized electron at the shallow +/0 level and a hole localized at the 2+/+ level has a calculated energy of 2.24 eV. The FC shift of the + charge state following this transition is 0.78 eV, yielding a ZPL of 3.02 eV. The CCD (Fig. 17) fitted into the calculated optical transitions show that the resonant excitation energy is 0.77 eV below the energy of the crossover of the parabolas. The barrier for the thermal jump via this crossover from the vibrational ground state is 1.53 eV. This makes nitrogen vacancy most likely a radiative defect via the +/2+ transition level. The calculated optical transitions are in good agreement with the experimentally measured GL2,[36] i.e. PL maximum of 2.35 eV, and estimated ZPL in the range of 2.85-3.0 eV. The assumption that the GL2 band is caused by an internal transition between the excited and ground states of V_N^+, is supported by the experimentally observed exponential decay of the GL2 emission after pulsed excitation at low temperature, and invariance of the GL2 band's shape and position with changing excitation intensity.[35]

HSE calculations of optical transitions of V_N in *p*-type GaN, via the 3+/2+ level, can also be performed, where the emission line is computed to be 2.09 eV and the ZPL at 2.88 eV.[35] However, no experimental PL band for such transition has been observed.

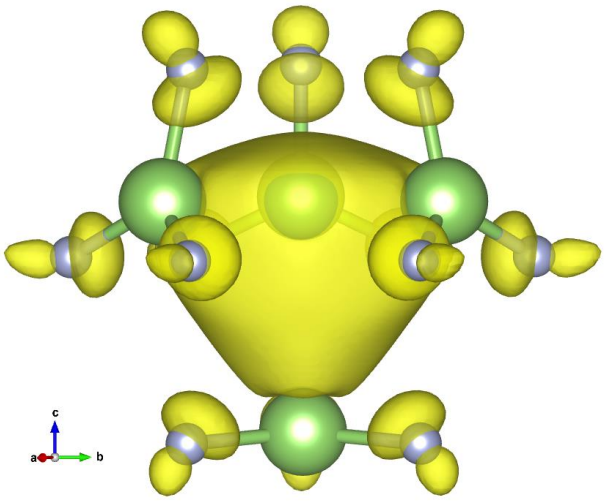

Figure 16: Charge density of the localized defect state of V_N calculated in the 3+ charge state. The isosurfaces with the value 6 % of the maximum are shown. There is a strongly localized charge density at the vacancy site, which is of *s*-character, while *s*- and *p*-hybridized parts of the defect state are formed at the neighboring N sites.

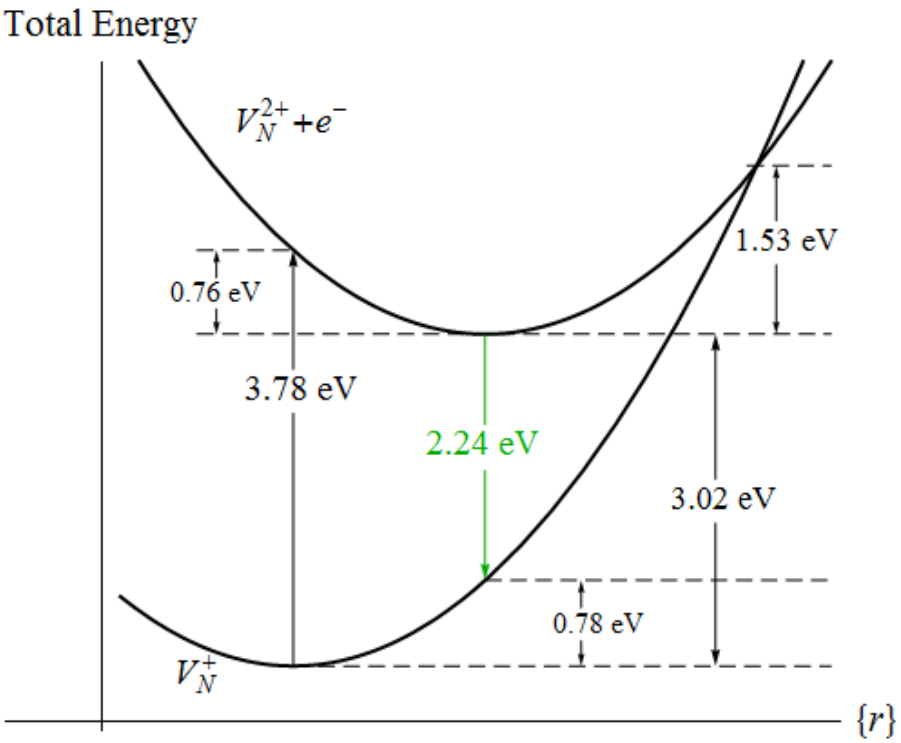

Figure 17: Configuration coordinate diagram of optical transitions for V_N. The PL maximum is 2.24 eV. Here the vibrational ground state of V_N^{2+} is 1.53 eV lower than the energy of the crossover of the potential curves, which makes transitions via the +/2+ level of V_N most likely radiative. The Franck–Condon shift of V_N^+ is computed to be 0.78 eV and the ZPL is 3.02 eV. These parameters are in close agreement with experimentally observed GL2 band.

4.1.4. Ga-N divacancy ($V_{Ga}V_N$)

4.1.4.1. Formation energy of $V_{Ga}V_N$

The relatively low formation energies of both V_{Ga} and V_N in *n*-type and *p*-type GaN, respectively, prompt the question of whether the isolated vacancies could bind into a stable complex ($V_{Ga}V_N$) divacancy. Figure 18 (a) displays the formation energies of the divacancy in its most stable charge states, i.e, 3+, +, 0, -, and 2-. The divacancy behaves as a deep donor as well as a deep acceptor, with calculated transition levels 3+/2+, 2+/+, +/0, 0/- and -/2- occurring at 0.81 eV, 0.68 eV, 0.98 eV, 1.48 eV and 1.95 eV above the VBM, respectively. The lack of stability of the 2+ charge state, being a characteristic of a negative-*U* defect, is a result of large charge-dependent atomic relaxations around the divacancy. In the 3+ charge state, the nearest N atoms around the vacant Ga site relax away by 11.7%, while the neighboring Ga atoms around the vacant N site move outwardly by 19.3%, compared to bulk Ga-N bonds. The 2- charge state is associated with smaller breathing relaxations around the empty Ga site (5.95 %) and vacant N site (5.33 %). According to Fig. 18 (a), divacancies ($V_{Ga}V_N$) display relatively low formation energies for Fermi levels close to the CBM. Note that the formation energy of the complex is comparable to that of V_{Ga} ($\Delta E_f \sim 0.08\,\text{eV}$) in *n*-type GaN, making it the second most probable intrinsic defect to occur in *n*-type GaN. In addition to the divacancy being fairly favorable in *n*-type GaN, it also displays a significantly large binding energy of 3.04 eV (Fig. 18 (b)) for Fermi levels above 2.34 eV from the VBM. In *p*-type GaN, the divacancy exhibits a binding energy of approximately 1 eV. Hence, divacancies are expected to be stable in bulk GaN if they are formed during growth or as a result of the electron irradiation. The stability of divacancies in *n*-

type GaN was also suggested in Refs. [65,70]; however no definite experimental identification of divacancies in *n*-type GaN has been demonstrated to date.

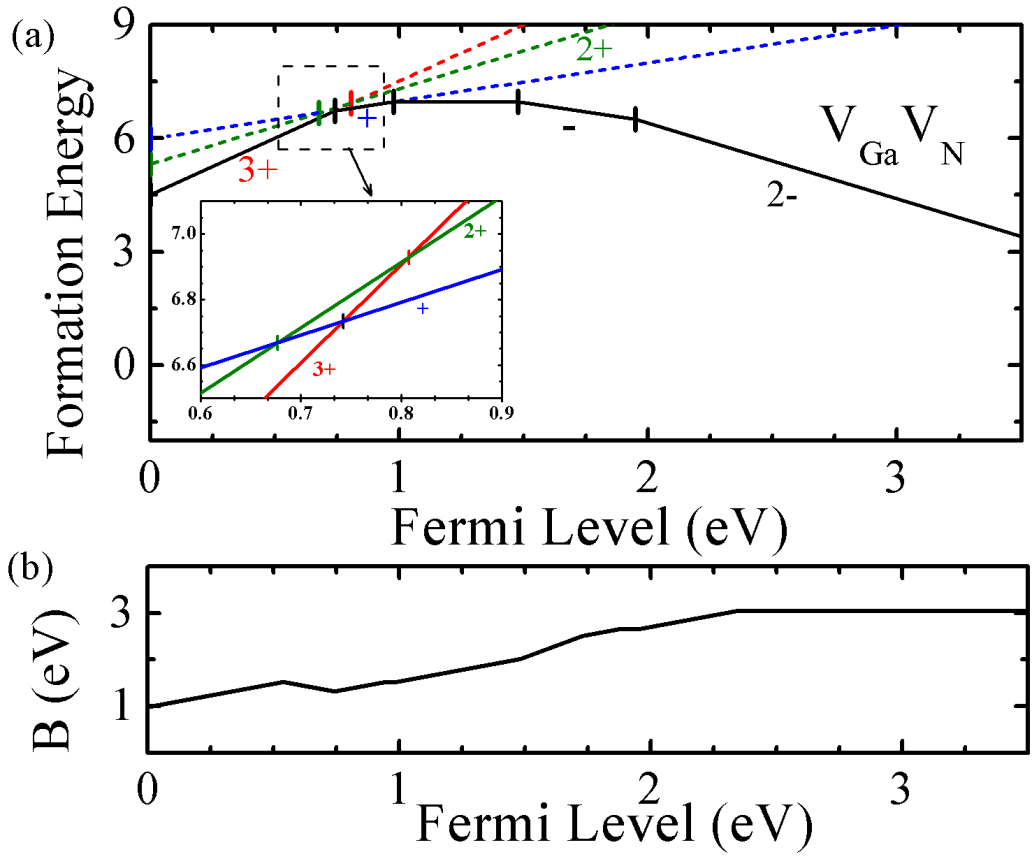

Figure 18: (a) Formation energy of the most stable charge states of the $V_{Ga}V_N$ defect (solid black line), in GaN grown under either Ga- or N-rich conditions. The dashed lines display the instability of the 2+ charge state or the negative-*U* behavior ($U = -0.13$ eV). The insets show the 2+/+ and 3+/2+ transition levels occurring at 0.68 eV and 0.81 eV, above the VBM, respectively. (b) Binding energy (B) of $V_{Ga}V_N$ as a function of the Fermi level across the band gap. In *n*-type GaN, the binding energy is calculated to be 3.04 eV.

4.1.4.2. Optics of $V_{Ga}V_N$

Figure 19 displays the CCD obtained by parabolic fits into the calculated optical transitions of divacancy via the 2-/- transition level. Starting from $V_{Ga}V_N^{2-}$ as the ground state in typical *n*-type GaN, the resonant excitation energy is predicted at 2.00 eV. A negatively charged $V_{Ga}V_N^{2-}$ captures a free hole at the 2-/- level. If radiative, the subsequent recombination of an electron in the conduction band (or bound to a shallow donor) and the hole localized at $V_{Ga}V_N^{-}$, has a calculated energy of 0.99 eV, with a Franck–Condon shift of 0.54 eV, and a ZPL of 1.53 eV. The calculated 0.99 eV emission peak is very close to the maximum of the experimentally observed broad structureless near-IR band (~0.95 eV) in electron irradiated GaN epilayers (Fig. 20),[51-56] thus making divacancy a possible candidate for this PL band. However, in the divacancy case, the calculated absorption energy is 0.12 eV larger than the energy of the crossover of the potential curves. The barrier for a non-radiative transition from the vibrational ground state of $V_{Ga}V_N^{-}$ via this crossover is 0.35 eV. This value of the energy barrier suggests significant temperature dependence of the non-radiative transition probability. Assuming a typical phonon frequency of 10^{13} s^{-1} and following Ref. [66], an estimation of the thermal jump probability suggests that at room temperature this defect is likely non-radiative. However, depending on the radiative lifetime of the PL, (the time the defect remains in the excited state), at a certain low temperature, the radiative recombination should prevail. Estimating this temperature requires the knowledge of the PL lifetime. This is in reasonable agreement with recent experimental studies of electron-irradiated GaN samples where appearance of the broad 0.95 eV PL band at low temperature (4.2 K) was observed and its near disappearance (~85%) after room temperature annealing (295 K) occurred.[56]

An alternative explanation of 0.95 eV near-IR band (shown in Fig. 20) has been discussed based on experiments (Refs. [54-56]). It was suggested that the migration of interstitial Ga (Ga_i defects will be analyzed in Section 4.5.2) near a vacant Ga site could play an important role in the broad 0.95 eV PL band. We performed HSE calculations of various configurations of defect complexes consisting of Ga_i with neighboring V_{Ga}. However, as a result of relaxation, interstitial Ga relocates into the vacancy making the complex $V_{Ga}Ga_i$ unstable. Therefore, if electron irradiation creates Ga_i in the immediate vicinity of V_{Ga}, the resulting complex quickly annihilates. In addition to $V_{Ga}Ga_i$, several atomic configurations of the complex Ga_NV_{Ga} were investigated. However, upon relaxation, Ga antisite also relocates into the Ga vacancy, yielding the isolated nitrogen vacancy.

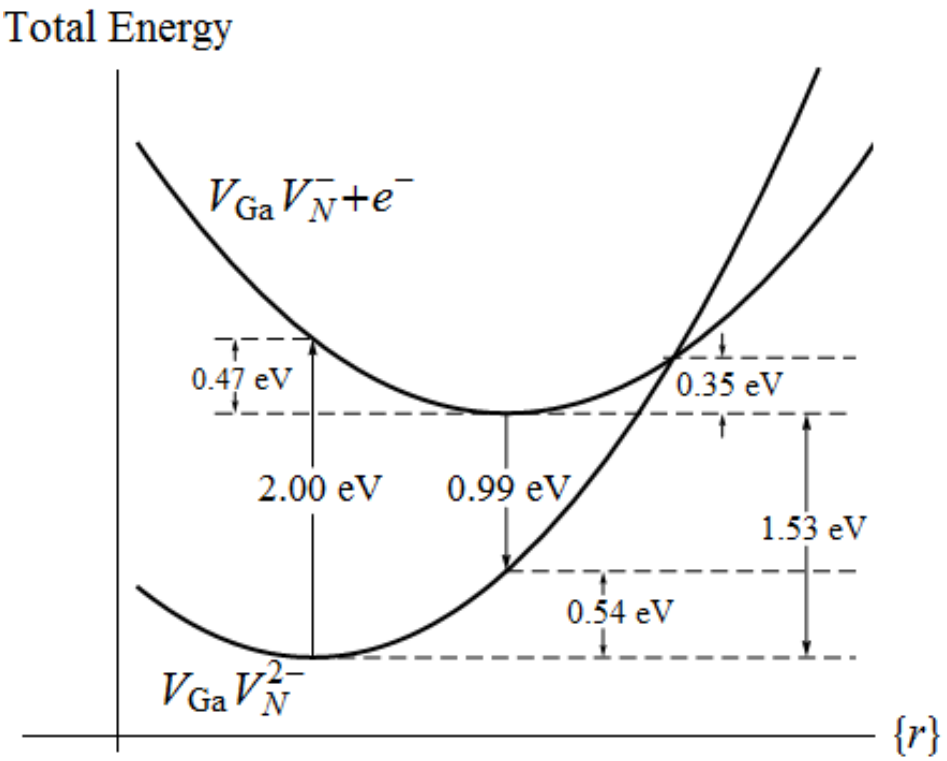

Figure 19: Configuration coordinate diagram for the optical transitions via the $V_{Ga}V_N$ defect in GaN. The emission is predicted to occur at 0.99 eV while the FC shift and ZPL are calculated to be 0.54 eV and 1.53 eV, respectively. Here the vibrational ground state of $V_{Ga}V_N^-$ is 0.35 eV below the energy of the crossover of the potential curves, suggesting that this defect is radiative only at low temperatures.

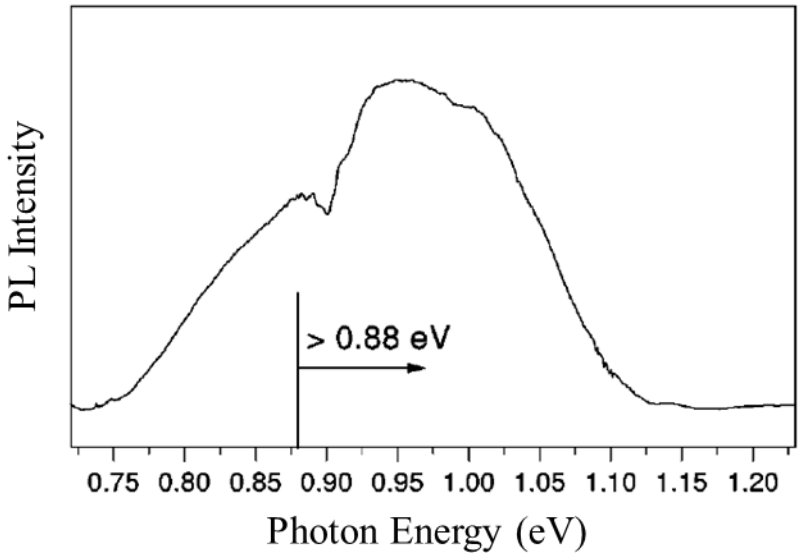

Figure 20: Broad PL band peaking at approximately 0.95 eV observed by Ref. [53] in electron-irradiated GaN samples. *The arrow indicates the range of filter used to distinguish the broad band from the PL-ODEPR spectra*.

4.1.5. Interstitial Ga (Ga_i)

4.1.5.1. Atomic Structure and Formation energy of Ga_i

HSE calculations show that interstitial Ga occurs in the +, 2+ and 3+ charge states and its most stable configuration is a near octahedral site, which is in agreement with recent HSE results.[75,76] Interstitial Ga at the tetrahedral site is found to be 93 meV higher in energy than a near octahedral Ga_i. Figure 21(a) displays the ideal atomic configurations of the tetrahedral and octahedral interstitial sites in the primitive wurtzite GaN unit cell. For comparison, the relaxed geometry of Ga_i in the + charge state is shown in Fig. 21(b). In a singly positive charge state (+), the distances from interstitial Ga to the two neighboring Ga planes are 2.52 Å and 2.39 Å, and the distances between Ga_i and N atoms in the two nearest N planes are 2.00 Å and 2.75 Å. In the 3+ charge state, when compared to the corresponding bonds of Ga_i^+, the Ga_i-Ga bonds exhibit slight outward breathing relaxation of 3.38% and 0.89%, respectively, while the Ga_i-N distances decrease by 1.76% and 12.2%. The attraction between interstitial Ga and nearest N atoms in the 3+ charge state obtained from HSE calculations also agrees well with previous DFT results.[66]

The formation energies and transition levels of interstitial Ga are shown in Fig. 22. This defect forms two deep donor transition levels 3+/2+ and 2+/+ at 2.33 eV and 2.64 eV, above the VBM, respectively. The +/0 transition level is found to be resonant with the CBM, suggesting the existence of a shallow donor state inaccessible in our supercell calculations. In both growing environments, for Fermi level values close to the VBM, Ga_i is the most energetically favorable defect among substitutional and interstitial native defects (i.e. excluding vacancies). Since Ga_i is a donor defect, it could therefore act as compensating center for *p*-type GaN. On the other hand,

interstitial Ga has high formation energy for Fermi levels near the CBM, suggesting that it is unlikely to form in *n*-type GaN.

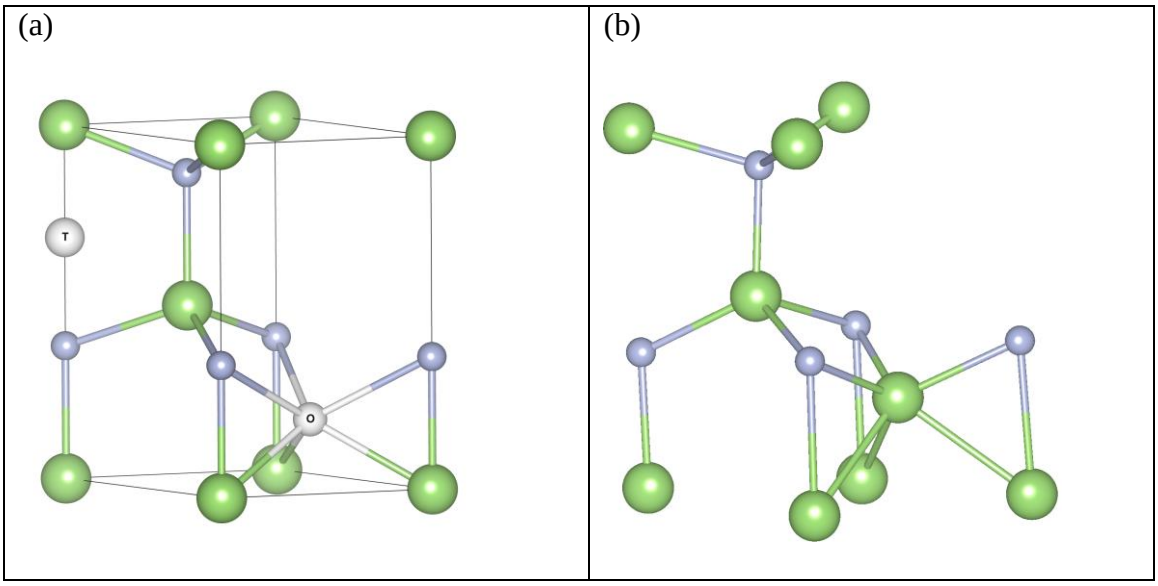

Figure 21: (a) Atomic configuration of the tetrahedral (labeled T) and octahedral (labeled O) interstitial sites in the wurtzite GaN. Large green spheres represent Ga atoms and small grey spheres represent N atoms. (b) Relaxed atomic structure of Ga_i in the + charge state. The Ga atom occupies a slightly distorted octahedral site, where the distances between Ga_i and nearest N atoms decrease by 4.28%, when compared to ideal octahedral configuration.

4.1.5.2. Optical Transitions of Ga_i

Our calculations of interstitial Ga can help explain the nature of the sharp near-IR band observed in 2.5 MeV electron-irradiated GaN samples. This PL band displayed in Fig. 24 peaks at 0.85 eV, and has a ZPL at 0.88 eV. Based on the slight shift of the peak of this PL band observed in different GaN epilayers, Buyanova et al.[52,199] argued that it could be caused by an internal transition between the excited state close to the CBM and the ground state of a deep isolated defect. Further experimental study on this PL band performed by Chen et al.[200] suggested that the sharp near IR PL peak is related to either isolated substitutional O_N impurity (which is unexpected since O_N is known to be a shallow donor) or a complex associated with O_N. Overall, to our knowledge, no conclusive experimental or theoretical explanation of the microscopic origin of the sharp 0.85 PL band in electron-irradiated GaN has been provided thus far. Our HSE calculations suggest that an internal transition between the excited and ground state of the + charge state of the interstitial Ga is responsible for this PL band.[51-56,199]

For Fermi levels close to the CBM, interstitial Ga has the lowest energy in the + charge state (Fig. 22). As mentioned above, in our calculations, the +/0 transition level is resonant with the conduction band, which implies a shallow donor level likely a few tens of meV below the CBM. At low temperatures and under the ultraviolet (UV) illumination, generating an electron-hole pair, the unoccupied shallow donor level of Ga_i^+ can capture the electron, transferring the defect into neutral charge state Ga_i^0. Since the captured electron is weakly localized on Ga_i, almost no relaxation of the lattice occurs. Subsequently, a free hole from the VBM is captured by the neutral defect at the 2+/+ transition level. Since this defect state is strongly localized, the hole capture leads to the lattice relaxation, which corresponds to the minimum energy lattice of Ga_i^{2+}.

Consequently, the defect is now converted into Ga_i^{2+} plus a weakly localized electron. Finally, an internal transition occurs, i.e. the recombination of the weakly localized electron from the shallow +/0 level and a hole strongly localized at the 2+/+ deep defect level. We calculate the energy of this transition to be 0.72 eV, which should cause a near IR PL band. The obtained relaxation energy (Franck-Condon shift) following this transition is 0.12 eV, yielding a ZPL of 0.84 eV, suggesting that this transition produces a sharp PL band.

The calculated CCD and the optical transitions via interstitial Ga are displayed in Fig. 23. In this case, the potential curves are found to never intersect; hence the path for non-radiative transitions described above in section 3.4.2.2 is unavailable. This defect is therefore likely to be radiative. Our calculated optical transitions (shown in Fig. 23) for interstitial Ga are in good agreement with the previously observed 0.85 eV PL band associated with a sharp ZPL of 0.88 eV in electron-irradiated GaN samples (Fig. 24).[51-56,199] Most recent HSE calculations of the migration mechanisms of interstitial Ga in the + charge state show a diffusion energy barrier of 1.6 eV.[76] In *n*-type GaN, Ga_i occurs in the + charge state (Fig. 22). Using the approximation for thermal jump rate (discussed in section 3.4.2.2), it can be estimated that Ga_i become mobile at temperatures around 620 K.

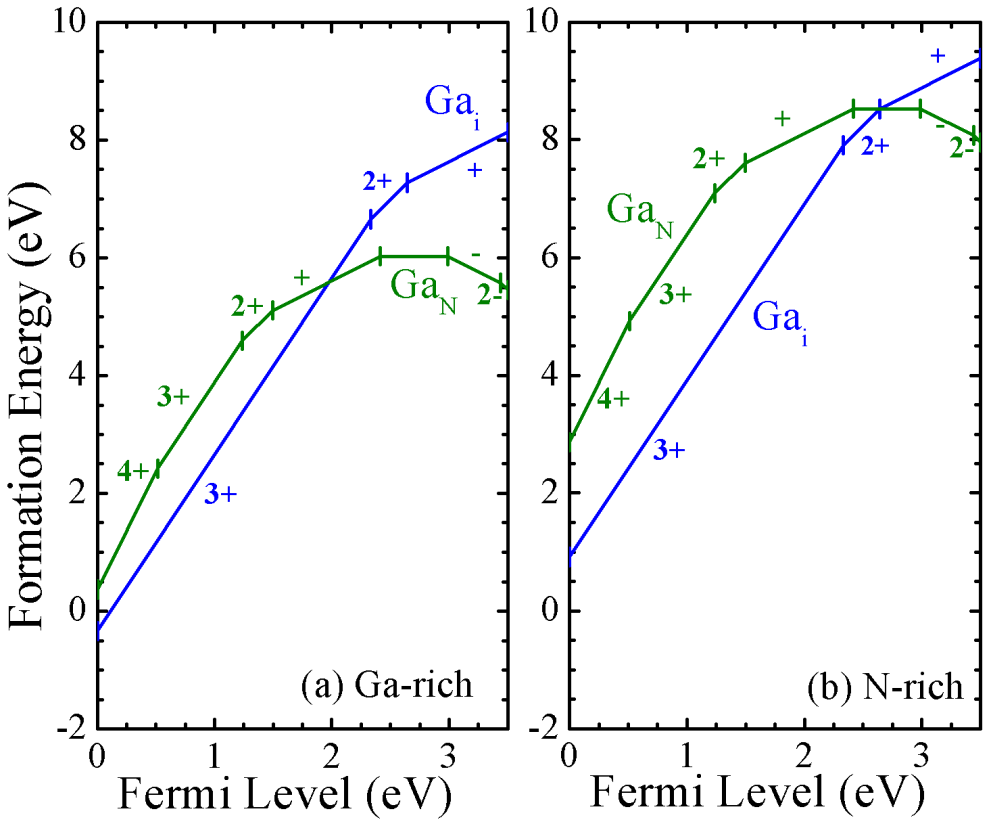

Figure 22: Formation energies of the Ga native defects as a function of the Fermi energy calculated for (a) Ga-rich and (b) N-rich growth conditions. In *p*-type GaN and Ga-rich environment, both interstitial and antisite Ga possess the lowest formation energies among investigated substitutional and interstitial native defects, while displaying very high formation energies in *n*-type GaN in both growth conditions.

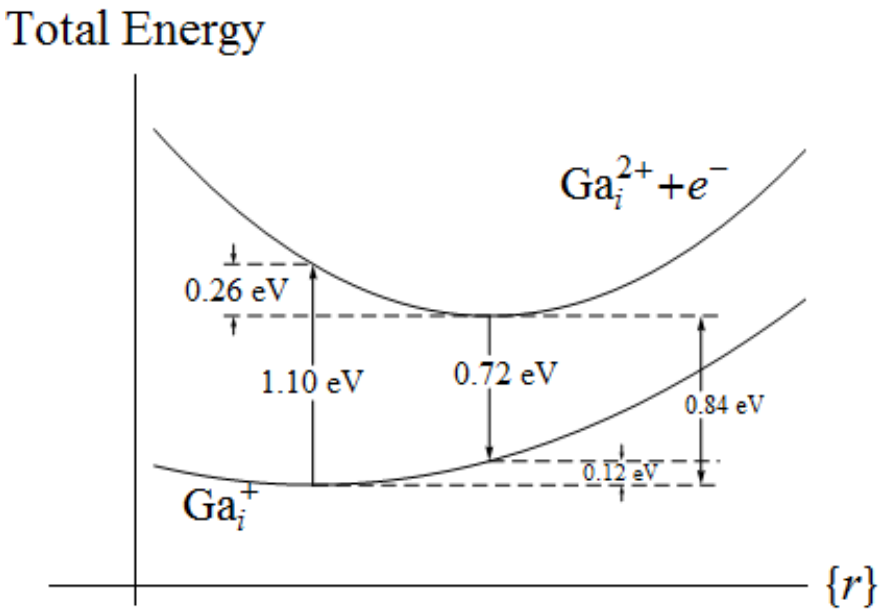

Figure 23: Configuration coordinate diagram for the isolated interstitial Ga, displaying calculated optical transitions via the +/2+ transition level. The peak of the PL band is at 0.72 eV. The two potential curves never intersect, making Ga_i likely a radiative defect. The ZPL is found to be 0.84 eV and the Franck-Condon shift (relaxation energy) is 0.12 eV.

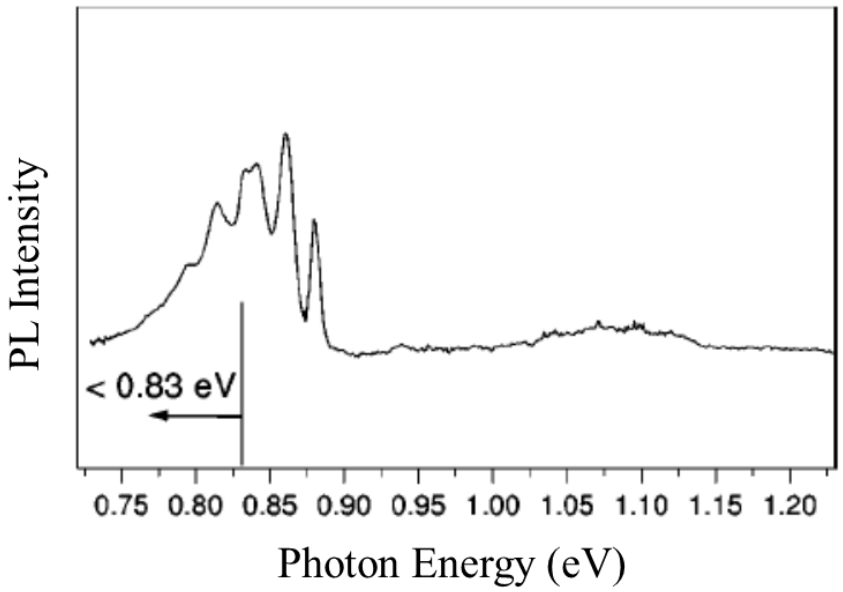

Figure 24: Sharp PL band with a ZPL at 0.88 eV observed by Ref. [53] in 2.5 MeV electron-irradiated samples. *The use of the arrow indicates the range of filter used in the experiment in order to separate the fine structure band from the PLODEPR band.*

4.1.6. Gallium antisite (Ga_N)

HSE calculations show that Ga antisite exhibits multiple stable charge states, depending on the Fermi level within the gap, i.e, 4+, 3+, 2+, +, 0, - and 2-. The 4+ charge state (shown in Fig. 25 (a)) is associated with considerable atomic distortions where the antisite defect relaxes along the *c*-axis which consequently pushes away its *c*-axis neighboring Ga atom by ~0.50 Å from its initial site. As a result, the Ga_N-Ga bond parallel to the *c*-axis is 36.8% longer than the ideal bulk Ga-N bond length. The three remaining Ga_N-Ga bonds are 24.3% longer. In contrast, in the 2- state (displayed in Fig. 25(b)), the Ga_N-Ga bonds undergo weaker outward relaxations of 11.1% and 10.3%, respectively.

Figure 26 displays the formation energy as a function of the Fermi level for Ga antisite in GaN grown under (a) Ga-rich and (b) N-rich conditions. Ga antisite behaves as a donor and an acceptor defect and exhibits six transition levels within the bandgap. The 4+/3+ occurs at 0.52 eV, 3+/2+ at 1.24 eV, 2+/+ at 1.50 eV, +/0 at 2.42 eV, 0/- at 2.99 eV and -/2- at 3.44 eV, above the VBM. Note that the -/2- transition level is very close to the CBM (~0.06 eV), making it difficult to confirm the stability of the 2- charge state. In Ga-rich conditions and for Fermi level close to the CBM, Ga antisites display formation energies lower to that of interstitial Ga ($\Delta E_f \sim 2.67\,\text{eV}$), whereas in *p*-type GaN, Ga_N is less energetically favorable than Ga_i by $\sim 0.7\,\text{eV}$.

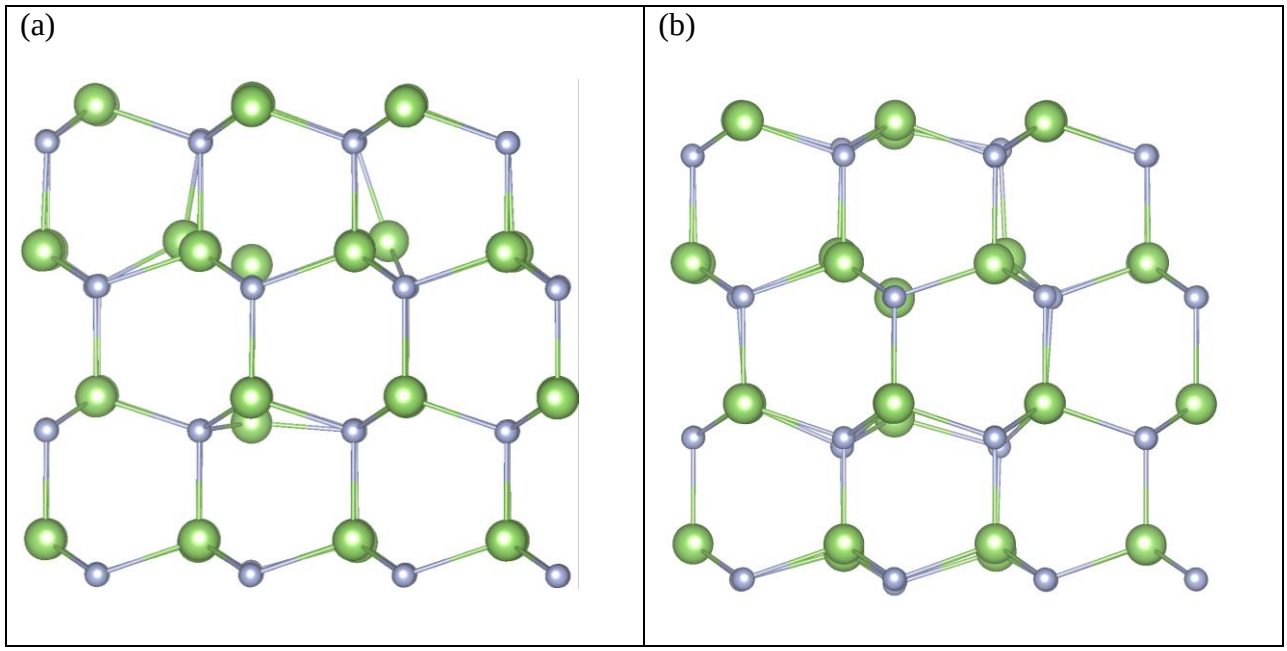

Figure 25: Relaxed atomic configurations of Ga_N in (a) the 4+ state and in the (b) 2- charge state. The 4+ charge state is accompanied with huge lattice distortions (~37%) in the neighborhood of the defect while the 2+ charge state induces smaller atomic relaxations (~11%).

4.1.7. Nitrogen antisite (Ga_N)

The formation energies of N antisite (N_{Ga}) are shown in Fig. 26 (a) and (b). N antisite is a deep donor, stable in the 2+ charge state for Fermi levels below ~2.40 eV from the VBM, and in the neutral state above that value. In the neutral charge state, the distance between N antisite and its three nearest N atoms decrease by 6.3 %, compared to bulk Ga-N bond length. The 2+ charge state of N_{Ga} is associated with substantial inward relaxation of 23.9% of N_{Ga}-N bonds, where the neighboring N atoms relax towards N antisite (Fig. 27). Note that the + charge state is higher than both 2+ and 0 charge states; this is a characteristic of a negative-*U* defect, which is associated with large lattice relaxations in 2+ charge state. Among all possible native defects investigated in this paper, N_{Ga} possesses the highest formation energy in both growing environments, and in both *n*-type and *p*-type GaN. Such high formation energy implies that N_{Ga} defects are unlikely to form.

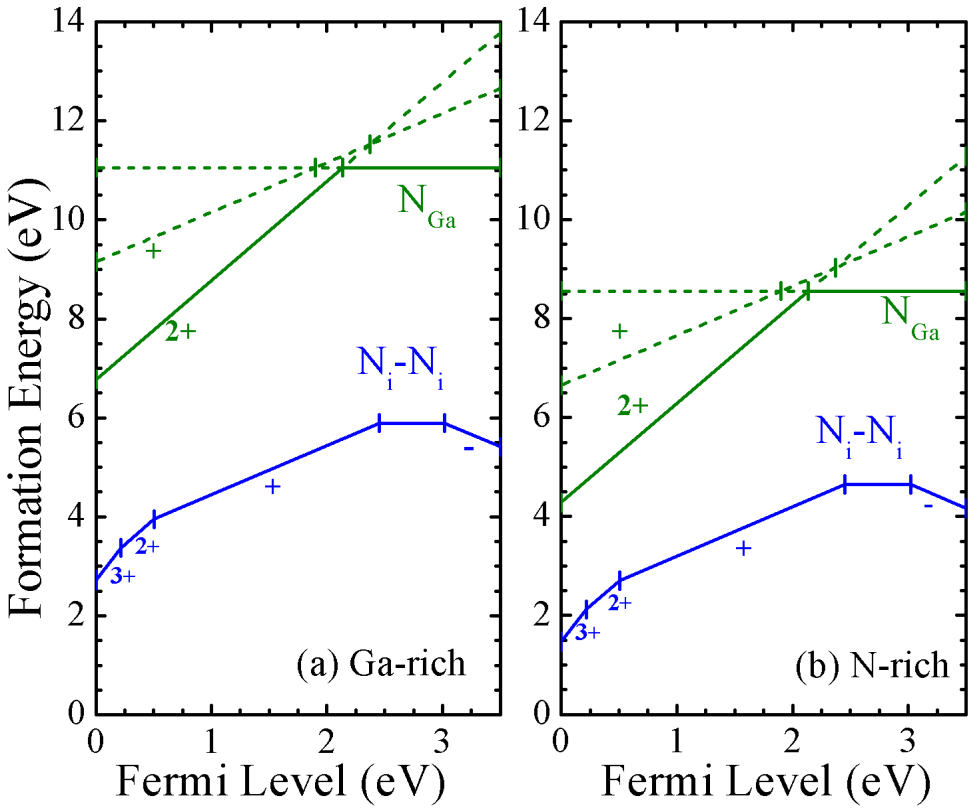

Figure 26: Formation energies of native nitrogen defects as a function of the Fermi energy for GaN grown in (a) Ga-rich and (b) N-rich environments. The dashed lines are used to show the instability of the + charge state for the N antisite, displaying a negative-*U* character ($U = -0.73$ eV).

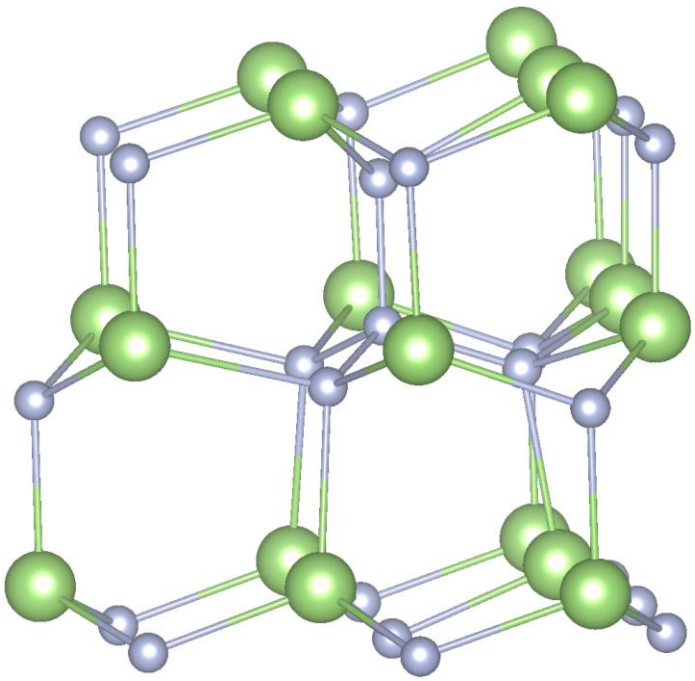

Figure 27: Relaxed atomic structure of N_{Ga} in the 2+ charge state. Here the next nearest N atom located along the *c*-axis and the neighboring N located in the basal plane relax inwardly towards the N antisite by approximately 24 %.

4.1.8. Interstitial Nitrogen

4.1.8.1. Formation Energy of Interstitial Nitrogen

Figure 26 (in the above section 4.1.7) shows calculated formation energies of the N split interstitial (N_i-N_i) as a function of the Fermi energy in Ga- and N-rich growth conditions. N split interstitial displays both acceptor 0/- transition level at 0.48 eV below the CBM, and several deep donor levels, namely, the +/0, 2+/+ and 3+/2+, respectively occurring at 2.45 eV, 0.51 eV and 0.22 eV above the VBM. Overall, nitrogen interstitial defect has relatively low formation energy, compared to other interstitial and antisite defects in GaN.

In addition to N split interstitial, the possible incorporation of the nitrogen molecule (N_2) into the GaN lattice was investigated. The N_2 molecule was placed at the center of the hexagonal cage in various directions relative to the *c*-axis and at the Ga-N bond center. HSE calculations of the various relaxed configurations of N_2 yield formation energies that are significantly higher than that obtained for the N split interstitial ($\Delta E_f \geq 4.81$ eV), hence making the interstitial N_2 molecules incorporation into bulk GaN unlikely.

4.1.8.2. Atomic Structure of Interstitial Nitrogen

As possible interstitial configurations of N in wurtzite GaN, we consider two distinct hexagonal sites ($N_{Hex\text{-}Ga}$ and $N_{Hex\text{-}N}$), the channel-centered site[201] (N_{CC}), the two octahedral sites (N_O), and the split interstitial geometry (N_i-N_i). The two interstitial hexagonal configurations, $N_{Hex\text{-}Ga}$ and $N_{Hex\text{-}N}$, are located at the centers of the Ga and N triangles in hexagonal planes, respectively. The channel-centered site (N_{CC}) is located at the center of the hexagonal channel between the two adjacent Ga and N planes. The two N_O sites are octahedrally coordinated by Ga

and N atoms respectively, and located symmetrically at ~0.32 Å away from the N_{CC} cite along the wurtzite *c*-axis. Finally, the split-interstitial site corresponds to the two neighboring N atoms sharing the same N site as shown in Fig. 28(b, c), compared to ideal GaN lattice (Fig. 28(a)).

Neutral interstitial N in the channel-centered site $\left(N_{CC}^{0}\right)$ is slightly distorted as a result of relaxation. Interstitial N atom moves along the wurtzite *c*-axis into a site located at about 2.10 Å and 2.22 Å from the Ga and N planes, respectively. Both neutral octahedral interstitial N sites are unstable, and the N atom relaxes into the above mentioned near channel-centered site. Similarly, nitrogen interstitial in the hexagonal site ($N_{Hex\text{-}Ga}$) is unstable, and upon relaxation, the N atom also relocates to the near channel-centered site (N_{CC}). The other hexagonal geometry $N_{Hex\text{-}N}$ in the neutral state is also distorted, i.e. the N atom is pushed away by one of three adjacent N atoms toward the other two by ~0.2 Å within the same nitrogen layer.

Our HSE calculations show that among all investigated interstitial N sites, the most stable configuration is the split-interstitial geometry (Fig. 28(b, c)). This defect was recently observed using high frequency EPR and electron nuclear double resonance (ENDOR) measurements in irradiated *n*-type GaN.[202] In the calculations, the initial N_i-N_i bond distance was chosen to be 1.13 Å, which is comparable to the bond distance of ideal N_2 molecule (1.1 Å).[66] N split interstitial defect exhibits multiple charge states, 3+, 2+, +, 0 and -. The relaxed atomic structures of the singly negative (-) and 3+ charge states of split interstitial N are displayed in Figs. 28(b) and 28(c). In the negative charge state, the N_i-N_i bond length is 1.41 Å while the distances from the N_i atoms to their nearest Ga atoms decrease by 7.48% and 6.82%, respectively, when compared to corresponding ideal Ga-N bonds. In the 3+ charge state, the N_i-N_i bond length is 1.11 Å and the N_i-Ga bonds undergo outward breathing relaxations of 29.7% and 17.4%, respectively. After relaxation, N split interstitial defect has formation energy that is 2.58 eV and

2.97 eV lower than those obtained in the $N_{Hex\text{-}N}$ and the near channel-centered sites, respectively. The obtained large differences in formation energies are also in good agreement with previous DFT calculations.[62,66]

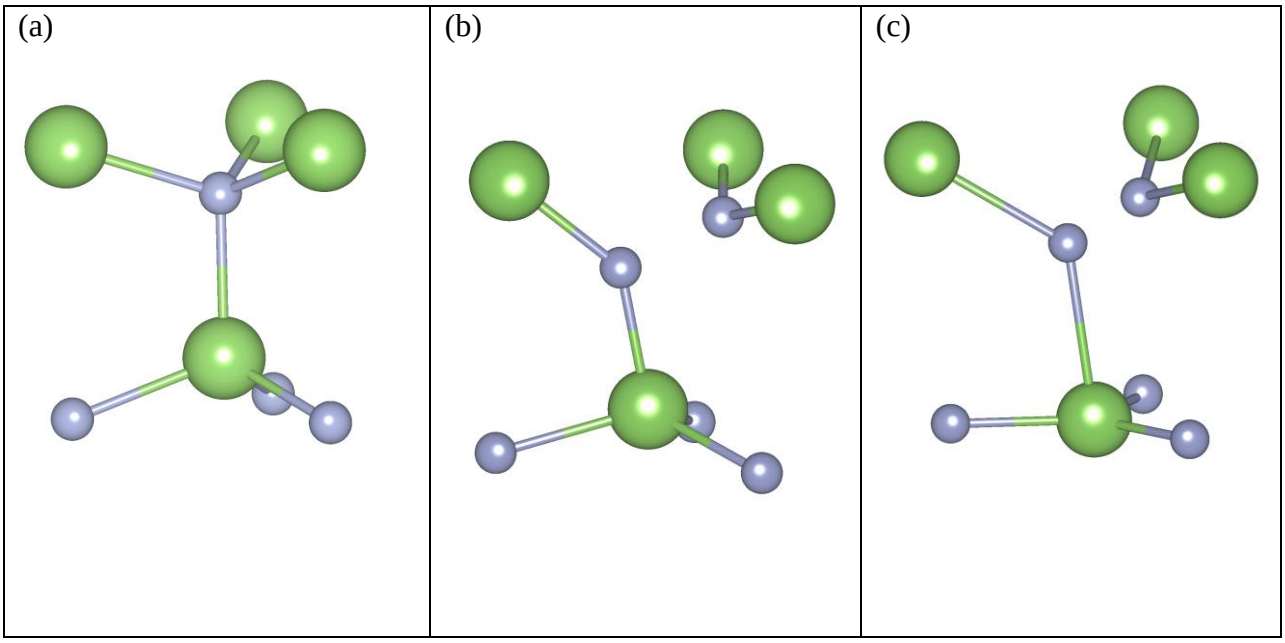

Figure 28: (a) Atomic structure of a section of ideal wurtzite GaN; (b) equivalent section of the relaxed N split interstitial (N_i-N_i) in the singly negative (-) charge state and (c) in the 3+ charge state. Large green spheres represent Ga atoms and small grey spheres represent N atoms. In the - charge state, the N_i-N_i bond is 1.41 Å; in the 3+ charge state, the N_i-N_i bond is reduced to 1.11 Å.

4.1.9. Summary of thermodynamic transition levels of native defects in GaN

Table 1: Thermodynamic transition levels $\varepsilon_T(q_1/q_2)$ of all investigated native defects in GaN, with the reference to the VBM, and their comparison with previous theoretical works.

Defects	q_1/q_2	$\varepsilon_T(q_1/q_2)$ in eV Present work	$\varepsilon_T(q_1/q_2)$ in eV HSE	$\varepsilon_T(q_1/q_2)$ in eV sX-LDA	$\varepsilon_T(q_1/q_2)$ in eV LDA and PBE-GGA
V_{Ga}	+/0	0.94	0.82[g]; 0.97[d]	…	…
	0/-	1.73	1.63[g]; 1.88[b]; 1.68[d]	1.37[e]	0.25[a]
	-/2-	1.87	2.09[g]; 2.10[b]; 2.33[d]	1.88[e]	0.64[a]
	3-/2-	2.34	2.3-2.4[g]; 3.13[b]; 2.80[d]	2.09[e]	1.10[a]
V_N	3+/2+	0.61	…	…	…
	3+/+**	0.54	0.50[g]; 0.68[b]; 0.47[f]; 0.70[i]	0.68[e]	1.18[a]
	2+/+	0.47	…	…	…
	+/0	CBM*	3.21[g]; 3.17[b]; 3.26[f]	…	…
	+/-**	…	…	3.32[e]	…
	0/-	…	3.4[g]		
$V_{Ga}V_N$	3+/2+	0.81	…	…	…
	3+/+**	0.74	…	…	…
	2+/+	0.68	…	…	0.5[c]
	+/0	0.98	…	…	0.65[c]
	0/-	1.48	…	…	…

	0/2-[**]	…	…	…	~0.66-0.7[c]
	-/2-	1.95	…	…	…
Ga_i	3+/2+	2.33	2.18[g]; 2.43[b]	…	2.55[a]
	2+/+	2.64	2.42[g]; 2.83[b]	…	2.39[a]
	+/0	CBM*	…	…	…
Ga_N	4+/3+	0.52	1.56[b]	…	0.93[h]
	3+/2+	1.24	1.60[b]	…	1.86[h]
	2+/+	1.50	2.13[b]	…	2.08[h]
	+/0	2.42	2.50[b]	…	2.27[h]
	0/-	2.99	3.23[b]	…	2.70[h]
	-/2-	3.44	…	…	…
N_i	3+/2+	0.22	VBM[ǂ(g)]; 0.77[b]	…	0.74[a]
	2+/+	0.51	0.50[g]; 1.80[b]	…	0.90[a]
	+/0	2.45	2.16[g]; 2.20[b]	…	1.48[a]
	0/-	3.02	2.82[g]; 3.28[b]	…	2.00[a]
N_{Ga}	2+/+	2.76	1.70[b]	…	0.88[h]
	+/0	2.03	2.67[b]	…	1.68[h]
	2+/0[**]	2.40	…	…	…
	0/-	…	…	…	2.70[h]

[*]: Shallow transition level resonant with the CBM.

[**]: Crossover due to *–U* behavior.

ǂ: Transition level resonant with the VBM.

[a] Reference 66.

[b] Reference 75.

[c] Reference 70.

[d] Reference 74.

[e] Reference 73 with Freysoldt corrections.[203,204]

[f] Reference 72.

[g] Reference 76.

[h] Reference 64.

[i] Reference 36.

Transition levels of all investigated intrinsic defects in this paper are shown in Fig. 29. Also, a comparison between defect transition levels obtained here and in previous theoretical works is displayed in Table 1.

In the case of V_{Ga}, the absence of the +/0 transition level in Ref. [66] (using LDA) and Ref. [73] (using sX-LDA), and the differences in other transition energies could be attributed to the use of different exchange-correlation functionals in these works. The application of different functionals can also lead to different atomic relaxations of the nearest N atoms around the vacant Ga site. For example, in the 3- charge state, our HSE calculations yield an outward breathing relaxation of ~9-10%, which is approximately twice of that previously obtained with LDA (4%).[66] The differences in lattice relaxations of a defect also contribute to the discrepancies in calculated transition levels. However, our calculated relaxations of nearest N atoms around V_{Ga}^{3-} are similar to the 11-12% relaxation values obtained by sX-LDA in Ref. [73], while the resulting

2-/3- transition levels differ by 0.25 eV. Therefore the difference in defect relaxations in different methods is not the only contributing factor in the transition level discrepancies.

The transition levels for vacancy of Ga described here mostly agree with most recent HSE calculations (see Table 1). However, Ga vacancy transition levels 2-/3- and -/2- obtained in Refs. [74,75], were found to occur deeper with respect to VBM in the band gap. Although the Ga 3*d* electrons were used in those calculations, our tests show that the inclusion of Ga 3*d* electrons results in negligible differences for calculated transition levels. Other possible sources of the discrepancies are the **k**-point sampling methods and supercell sizes. For example, in Ref. [38], using a 2×2×2 **k**-point mesh and 96-atom supercells, a formation energy of 4.42 eV (calculated at CBM) was obtained for V_{Ga}^{3-} which is 1.10 eV higher than our obtained values of 3.32 eV (calculated at Γ-point in 128-atom hexagonal supercell). In Ref. [75], HSE calculation with an off-center single **k**-point and 108 atom-supercells, formation energy of 3.2 eV (at CBM) was obtained for V_{Ga}^{3-} which only differs from our result by 0.1 eV. Our tests also show that using the above **k**-point sampling methods produces formation energy differences of ~0.05 eV.

Finally, the difference in the results could be due to the use of different electrostatic correction schemes. Two common approaches are the Freysoldt corrections[203,204], used in Refs. [74-76], and the Lany and Zunger corrections[174] used in this dissertation. As could be seen from Table 1, our results tend to be similar to HSE results of Refs. [74,75] for transition levels between the low charge states, and significantly different for high charge state transition levels. However, the discrepancies between our results and the results from Ref. [76] tend to be small even for high charge state transition levels (in the latter work, a slightly smaller amount of exact exchange of 29% was used).

For the nitrogen vacancy, the crossover 3+/+ (negative-*U* center) is also similar in most recent HSE calculations.[72,76] Nevertheless, in contrast to this work, the shallow +/0 transition level was not obtained, rather a +/0 level was predicted to be deep, occurring at 0.15-0.2 eV below the CBM.[72,75,76]

Although the negative-*U* character of divacancies was also described in Ref. [70] using GGA, authors obtained a 0/2- crossover at ~0.66-0.7 eV above the VBM, while we obtain a 3+/+ crossover at 0.74 eV. The formation energy calculated in Ref. [70] for the divacancy in the 2- charge state is ~1.2 eV lower than in this work.

Previous LDA predictions of negative-*U* center of interstitial Ga are not reproduced in this paper.[64,66] The atomic relaxations obtained here are very similar to previous LDA results for interstitial Ga in the 2+ charge state. The use of the LDA functional in Refs. [64, 66] is the source of the transition levels discrepancies. Recent HSE calculations[75] of Ga_i show a difference in transition levels of the 3+/2+ and 2+/+ of 0.1 eV and 0.2 eV, in the + and 3+ charge states, respectively, when compared to our work. Here, we also notice that the differences are not following the charge states of the defects. Comparison of transition levels of antisite defects (Ga_N and N_{Ga}) and N split interstitial with previous HSE results show similar trends as previously discussed.

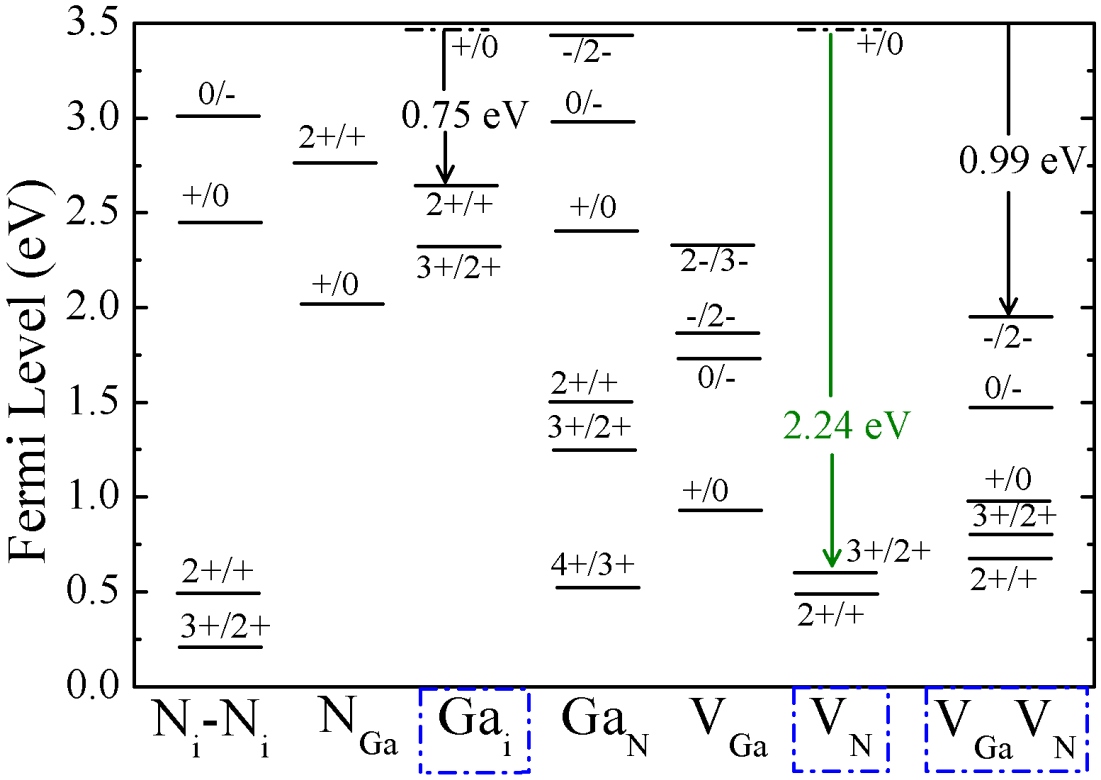

Figure 29: Thermodynamic transition levels $\varepsilon_T(q_1/q_2)$ of all investigated native defects in GaN, with the reference to the VBM. The solid lines denote the positions of the deep defect transition levels. The +/0 transition levels of Ga_i and V_N are calculated to be resonant with the conduction band, which suggests that experimentally, shallow donor levels (black dashed lines) of these defects should be observed. The straight arrows display HSE calculated optical transitions (emission lines) of Ga_i, V_N and $V_{Ga}V_N$ defects.

4.1.10. Concluding remarks regarding the analysis of native defects in GaN

In this section, we have performed theoretical investigation of the electronic and optical properties of most common native defects, i.e, V_{Ga}, V_N, $V_{Ga}V_{N,}$ Ga_i, Ga_N, N_i-N_i, N_{Ga}, Ga_iV_{Ga} and Ga_NV_{Ga}. As predicted, the use of exchange tuned HSE leads to significant changes in the predictions of thermodynamic transition levels and optical transitions for several intrinsic defects in GaN compared to local approximations to the DFT.

Analysis of the configuration coordinate diagram constructed from the computed HSE transition energies suggests that Ga vacancy is likely a non-radiative defect. Our calculations of Ga vacancy show that V_{Ga} in the neutral state exhibits a large magnetic moment of 3 μ_B, while in the + charge state, the spins on the four neighboring nitrogen atoms around the vacant Ga site favor an AFM order over a FM spin configuration with energy difference of 75 meV.

Nitrogen vacancy is found to be the most energetically stable native defect in *p*-type GaN and for Fermi levels up to ~3 eV above the VBM. Calculated transitions via the 2+/+ level of V_N show an emission occurring at 2.24 eV and a zero-phonon line of 3.02 eV, which is in good agreement with recent experimental data on observed GL2 band.[35] The PL band of V_N is predicted to originate from internal transitions between a shallow and a deep level of the same defect. The HSE fitted configuration coordinate diagram suggests that V_N is a radiative defect.

In *n*-type GaN, we show divacancies to be fairly energetically stable with a binding energy of 3.04 eV. We also calculate an optical transition (emission) at 0.99 eV via the 2-/- level of the divacancy, with ZPL of 1.53 eV. This transition could be related to the near IR broad 0.95 PL band, frequently observed in 2.5 MeV electron irradiated GaN samples.[51-56] The calculations

also suggest that the transitions via this defect level are radiative at low temperatures and become non-radiative at room temperature.

Among investigated interstitial and antisite defects, both Ga_i and Ga_N were found to be energetically favorable in *p*-type GaN while split interstitial N_i-N_i exhibits the lowest formation energy in *n*-type GaN. Since calculations predict both a shallow donor level and deep levels of Ga_i, the optical transition via this defect is also suggested to be internal, with an electron weakly localized on the shallow level and a hole localized on a deep level of same defect prior to the recombination. Configuration coordinate diagram fitted into the HSE computed optical transitions and lattice relaxations, suggests that Ga interstitial is a radiative defect. Furthermore, interstitial Ga was found to be a good candidate for a defect responsible for the sharp near-IR 0.85 eV PL band associated with a ZPL at 0.88 eV observed in electron-irradiated GaN epilayers.[51-56,199] Calculations also suggest that N antisites are unlikely to occur in bulk GaN.

In addition to intrinsic defects, external defects also play a major role in the electrical and optical properties of bulk GaN. In the following section, we experimentally analyze the BL2 band in semi-insulating GaN grown by hydride vapor phase epitaxy (HVPE) and present theoretical results which allow the identification of the defect responsible for BL2. We attribute the BL2 band to the C_NO_N-H complex, or possibly to the C_N-H complex.

4.2. Experimental and theoretical analysis of hydrogen-carbon complexes and the blue luminescence band in GaN

4.2.1. Experimental Method

Here, we observed the BL2 band in several high-resistivity GaN samples grown by MOCVD and HVPE techniques. In all cases, the intensity of the BL2 band gradually decreased under continuous UV exposure, and simultaneously with this bleaching the YL band intensity increased. For detailed study in this work, we selected a freestanding GaN sample grown by HVPE at Kyma Technologies. The 200-μm-thick sample was doped with iron to make it semi-insulating. The presented PL spectra are obtained from the Ga face, which was chemically-mechanically polished.

Steady-state PL was excited with an unfocused He-Cd laser (30 mW, 325 nm), dispersed by a 1200 rules/mm grating in a 0.3 m monochromator and detected by a cooled photomultiplier tube. Calibrated neutral-density filters were used to attenuate the excitation power density (P_{exc}) over the range of 10^{-5} - 0.2 W/cm^2. The absolute internal quantum efficiency of PL, η, is defined as $\eta = I^{PL} / G$, where I^{PL} is the integrated PL intensity from a particular PL band and G is the concentration of electron-hole pairs created by the laser per second in the same volume. To find η for a particular PL band, we compared its integrated intensity with the PL intensity obtained from a calibrated GaN sample.[205,206]

4.2.2. Experimental Results and Discussion

Figure 30 shows a PL spectrum at photon energies above 2.6 eV at 18 K and with relatively high excitation intensity (P_{exc} = 200 mW/cm^2). The BL2 band has a maximum at about 3.0 eV and a characteristic fine structure at its high-energy side. In the excitonic region, the strongest line at 3.479 eV with the full width at half maximum (FWHM) of about 5 meV is presumably due to annihilation of an exciton bound to a neutral shallow donor. The lines at 3.446 eV and 3.356 eV labeled R4 and R5, respectively, are most probably the resonant Raman lines because their separations from the HeCd laser line (at 3.814 eV) are multiples of the LO phonon energy in GaN (about 91-92 meV). The line at 3.326 eV is identified as the ZPL of the BL2 band. The fine structure at the high-energy side of the BL2 band is identical to that observed in other samples, such as undoped or C-doped GaN grown by MOCVD.[79,80]

The shape of the BL2 band is asymmetric, corresponding to the case of a defect with a moderately strong electron-phonon coupling. The shape can be modeled with the following formula derived from a one-dimensional configuration coordinate model:[35]

$$I^{PL}(\hbar\omega) = I^{PL}_{\max} \exp\left[-2S_e\left(\sqrt{\frac{E_0 + 0.5\hbar\Omega - \hbar\omega}{E_0 + 0.5\hbar\Omega - \hbar\omega_{\max}}} - 1\right)^2\right], \qquad (4.2.2.1)$$

where $I^{PL}_{\max}$ is the PL intensity in the maximum of the broad band, S_e and $\hbar\Omega$ are the Huang-Rhys factor (previously discussed in section 3.4.2.2) and the dominant phonon energy for the excited state, respectively, $\hbar\omega$ is the photon energy, and E_0 is the ZPL energy. The value of $\hbar\Omega$ for the BL2 band is unknown and will be assumed to be 0.04 eV for definiteness. Then, the value of E_0 in this fit (3.33 eV) practically coincides with the experimentally observed ZPL. The value of S_e (4.3) is smaller than that for the YL band in GaN (7.4);[93] i.e., the electron-phonon coupling for the BL2 defect is weaker.

The low-temperature PL spectra at low excitation intensity (P_{exc} = 1 mW/cm^2), shown in Fig. 31, were measured before and after prolonged exposure with a HeCd laser at P_{exc} = 200 mW/cm^2. The scans with P_{exc} = 1 mW/cm^2 were taken with a relatively large step in wavelength (2 nm) and wide slits of a monochromator. These conditions were chosen to keep a high signal-to-noise ratio and to avoid any distortion of the PL spectra during the scan due to changes of the PL intensities caused by UV exposure. Because of this, the fine structure at photon energies above 3 eV cannot be observed in Fig. 31.

Before the UV exposure with P_{exc} = 200 mW/cm^2, the PL spectrum at photon energies below 2.6 eV contains a green band with a maximum at 2.36 eV, identified as the GL2 band caused by the nitrogen vacancy (V_N).[35] The shape of the GL2 band can be fitted well with Eq. (4.2.2.1), and in the fit shown in Fig. 31, we used previously reported parameters.[35] The YL band with a maximum at ~2.2 eV is not observed before the UV exposure, but it may contribute as a low-temperature shoulder to the GL2 band. The inclusion of the YL band with relatively low peak intensity and other parameters found in Ref. [93] greatly improves the fit of the overall PL spectrum (Fig. 31).

During the UV exposure with P_{exc} = 200 mW/cm^2, the BL2 band intensity gradually decreased, very similarly to our previous observations of the BL2 band in MOCVD-grown high-resistivity GaN samples.[78,79,80] Simultaneously with the bleaching of the BL2 band, the YL band intensity increased. After 290 min of the high-intensity UV exposure, the BL2 band intensity significantly decreased, whereas the YL band intensity greatly increased and became higher than that of the GL2 band.

In the fit shown with open circles in Fig. 31, all the parameters for the YL and GL2 bands remain unchanged except for the peak intensities. In the energy range above 2.6 eV, the best fit

was obtained when $\hbar\omega_{max}$ for the BL2 band was changed from 2.98 eV (before the UV exposure) to 2.96 eV (after 290 min of UV exposure). This indicates that under illumination, the source of the BL2 band undergoes some small changes, which could be possibly a rearrangement of H atoms around the defect, as will be discussed below. Other parameters describing the shape and width of the BL2 band remained unchanged.

By performing deconvolution of the PL spectrum while changing only the peak intensities of the YL, GL2 and BL2 bands and keeping intact the parameters describing the shapes and positions of these PL bands, we obtained the dependences of the quantum efficiencies for the major PL bands as a function of the UV exposure time (Fig. 32). The intensity of the GL2 band is practically independent of the UV exposure time, indicating a stable defect. The BL2 band intensity gradually decreased, while that of the YL band increased. The absolute reduction of the BL2 quantum efficiency (about 1.8×10^{-4}) is just slightly larger than the absolute increase of the YL quantum efficiency (about 1.0×10^{-4}). The difference can be explained by gradually increasing the non-radiative recombination efficiency (the total PL intensity decreased by 28% and the exciton emission intensity decreased by 57% during 5 hours of UV exposure). This suggests that the source of BL2 band converts into the source of YL band, as the sum of the two remains nearly constant. However, the reverse process is also possible, since the intensities of the BL2 and YL bands are restored to their original values after storing the sample at room temperature. The restoration times vary in different samples, from several hours for some samples to several days for other samples.

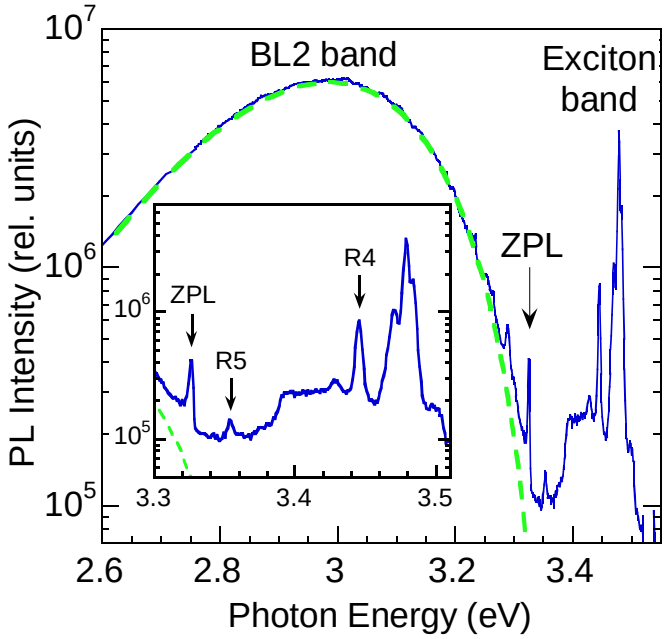

Figure 30: Low-temperature (T = 18 K) PL spectrum at P_{exc} = 200 mW/cm^2. The ZPL of the BL2 band at 3.326 eV is indicated with an arrow. The dashed line is a fit using Eq. (4.2.2.1) with the following parameters: $I_0^{PL} = 6\times10^6$, S_e = 4.3, $E_0 + 0.5\hbar\Omega = 3.35$eV, $\hbar\omega_{max} = 2.985$ eV. The inset shows the high resolution of the ZPL of the BL2 band and the higher energy lines.

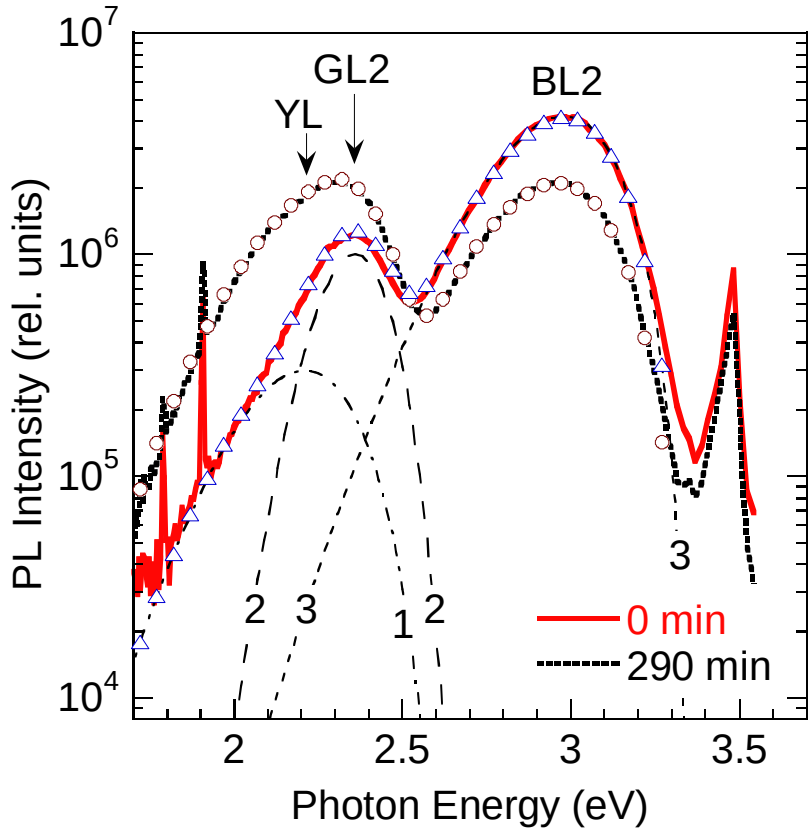

Figure 31: Low-temperature (T = 18 K) PL spectra measured at P_{exc} = 1 mW/cm^2 before (thick solid line) and after (thick dotted line) UV illumination with P_{exc} = 200 mW/cm^2 for 290 min. The PL intensity is divided by the excitation intensity. The contributions of three PL bands to the PL spectrum before illumination are shown with dashed and dash-dotted lines. The dash-dotted thin line 1 represents the shape of the YL band and is calculated using Eq. 4.2.2.1 with the following parameters: $I_0^{PL}(\text{YL}) = 3.0\times10^5$, S_e = 7.4, $E_0 + 0.5\hbar\Omega = 2.67$ eV, $\hbar\omega_{max} = 2.21$ eV. The long-dashed thin line 2 represents the shape of the GL2 band and is calculated using Eq. 4.2.2.1 with the following parameters: $I_0^{PL}(\text{GL2}) = 1.0\times10^6$, S_e = 26.5, $E_0 + 0.5\hbar\Omega = 2.87$ eV, $\hbar\omega_{max} = 2.36$ eV. The short-dashed thin line 3 represents the shape of the BL2 band and is calculated using Eq. 4.2.2.1 with the following parameters: $I_0^{PL}(\text{BL2}) = 4.2\times10^6$, S_e = 4.5,

$E_0 + 0.5\hbar\Omega = 3.35\text{eV}$, $\hbar\omega_{\max} = 2.98$ eV. The sum of the three band shapes is shown with empty triangles. The contributions of the individual PL bands to the PL spectrum after illumination are not shown for clarity, but their sum is shown with open circles. The individual band shapes after illuminations were calculated using Eq. (1) with the same parameters as before illumination, except for the following parameters: $I_0^{PL}(\text{YL}) = 1.47\times10^6$, $I_0^{PL}(\text{GL2}) = 1.04\times10^6$, $I_0^{PL}(\text{BL2}) = 2.1\times10^6$, and $\hbar\omega_{\max}(\text{BL2}) = 2.96$ eV. The small shift in the PL band maximum is needed to obtain a good fit.

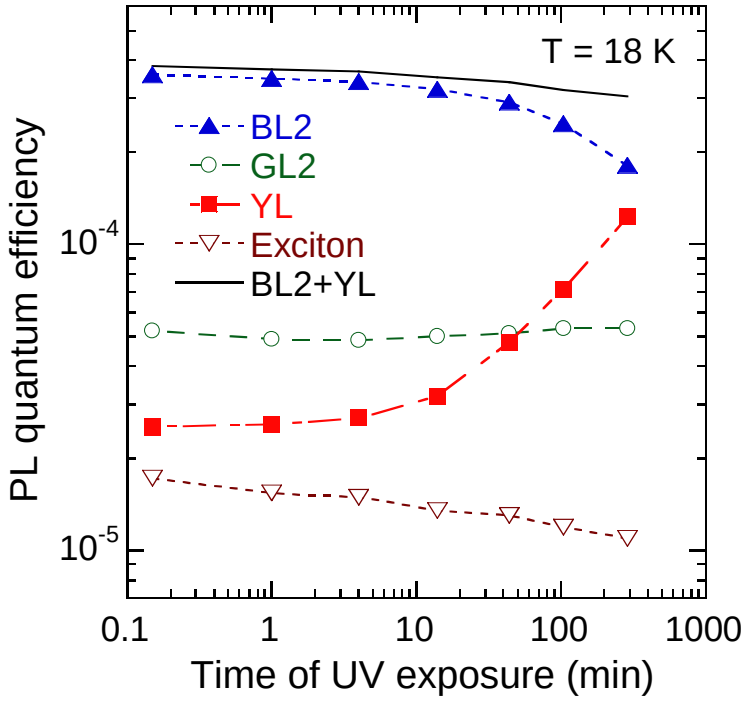

Figure 32: Evolution of the PL quantum efficiency for the main PL bands at T = 18 K and P_{exc} = 1 mW/cm^2 with time of UV exposure with P_{exc} = 200 mW/cm^2.

4.2.3. Theoretical Approach

The theoretical method used to investigate the behavior of the observed BL2 band is identical to the one used in the analysis of intrinsic defects in GaN (cf. section 4.1.1). However, a brief comment regarding the chemical potential of oxygen needs to be given. Contrary to our earlier work[92], in which the chemical potential of oxygen was obtained from the O_2 molecule, here we are deriving the chemical potential of O_2 using the HSE computed formation enthalpy of Ga_2O_3 (see section 3.3.1). The change in method in the calculation of the chemical potential of oxygen is based on a description given in Ref. [93].

4.2.4. Theoretical Results and Discussion

Since the PL experiments suggest that the defects responsible for the BL2 band under the UV illumination are being converted into the defects responsible for YL band, it is logical to assume that both BL2 and YL sources are carbon related. YL has been proposed to be originating from either the C_N acceptor or the C_NO_N donor-acceptor pair complex, depending on the amount of oxygen in the samples.[93] The two sources of YL in different samples can be distinguished by the emergence of the secondary green PL band (labeled GL) with increased excitation intensity, which is possible only in samples where the isolated C_N acceptor is generating YL. This is due to the C_N defect creating two transition levels in the bandgap, 0/+ and –/0. The recombination via –/0 levels creates YL, while GL emerges with increased excitation intensity when the 0/+ transition levels are activated upon saturation of the –/0 levels with holes. On the other hand, the C_NO_N complex creates a single 0/+ optically active transition level in the bandgap;[207] therefore, in samples where C_NO_N is the primary source of YL, the secondary GL is not observed (detailed analysis of carbon related YL and GL can be found in Ref. 93). The

samples studied in this work do not exhibit GL for any excitation intensity, suggesting that YL in this case is generated by C_NO_N complexes. Taking this as a starting point, and taking into account that the HVPE growth method leaves significant amounts of hydrogen in the studied samples, in the next section we analyze the possibility of hydrogen being bound to carbon-related defects, and show that this complex, either C_NO_N-H_i or C_N-H_i can explain the experimentally observed BL2 band.

4.2.4.1. Properties of Isolated Hydrogen

The technological importance of hydrogen in the fabrication of *p*-type GaN has been understood, since the passivation of Mg acceptors by hydrogen has been demonstrated.[208] Because of this, interstitial hydrogen has been extensively investigated by the first principles methods in the past two decades. DFT studies revealed complex behavior of hydrogen in GaN,[209,211,212] suggesting that it is a negative-*U* center with a very large *U* value of -2.4 eV. Here, we calculate the properties of interstitial hydrogen in GaN using the HSE06 hybrid functional. Figure 33 shows the formation energies of interstitial hydrogen H_i, for several high symmetry sites in the GaN wurtzite lattice.

Our results are qualitatively similar to previous DFT findings, with some notable quantitative differences. Hydrogen has multiple metastable interstitial sites, with relatively small energy differences (~0.2 eV), separated by barriers varying from 0.2 to ~1 eV in height. Here, we show the results for only a few of the lowest energy sites. We find the positive charge state of interstitial hydrogen (H_i^+) to be the most stable for Fermi levels below 3 eV in the bandgap. The most stable site for the + charge state (as well as the neutral state) is the bond center site along the wurtzite *c*-axis (labeled as $BC_{\parallel}$ in Fig. 33). Recent hybrid functional calculations showed H_i in the $BC_{\perp}$ site (one of the other three Ga-N bonds) to be 0.2 eV higher in energy.[213] The anti-

bonding nitrogen (AB(N)) site only slightly higher in energy (~0.1 eV) than the bond-center site, indicating roughly similar stability of the two for all values of E_F. For the Fermi energies above 3 eV, H_i exhibits acceptor type properties, with the anti-bonding gallium (AB(Ga) in Fig. 33) site being the most stable in singly negative charge state. The neutral charge state has higher formation energy for any position of the Fermi level. This negative-*U* behavior, however, is not as strong as was previously found in local approximations to the DFT. We estimate the value of *U* to be -0.77 eV. We also find the formation of the H_2 molecule to be unfavorable for any Fermi energy. As shown in Fig. 33 (dotted line), the H_2 molecule has a formation energy that is at least 1.5 eV higher than that of interstitial hydrogen for E_F ~3 eV and this difference is larger elsewhere.

Previous studies using GGA found that the potential barrier for H^+ to jump from the $BC_{\|}$ site to the AB(N) site is 0.22 eV, and a barrier between two adjacent AB(N) sites is 0.85 eV.[213] Assuming the typical phonon frequency of 10^{13} s^{-1} (section 3.4.2.2) it can be estimated that these migration barriers indicate that H^+ is mobile at room temperature. This means that interstitial hydrogen atoms, which are expected to be present in large concentrations in HVPE grown samples, will migrate and could form defect complexes due to an attractive interaction with other defects.

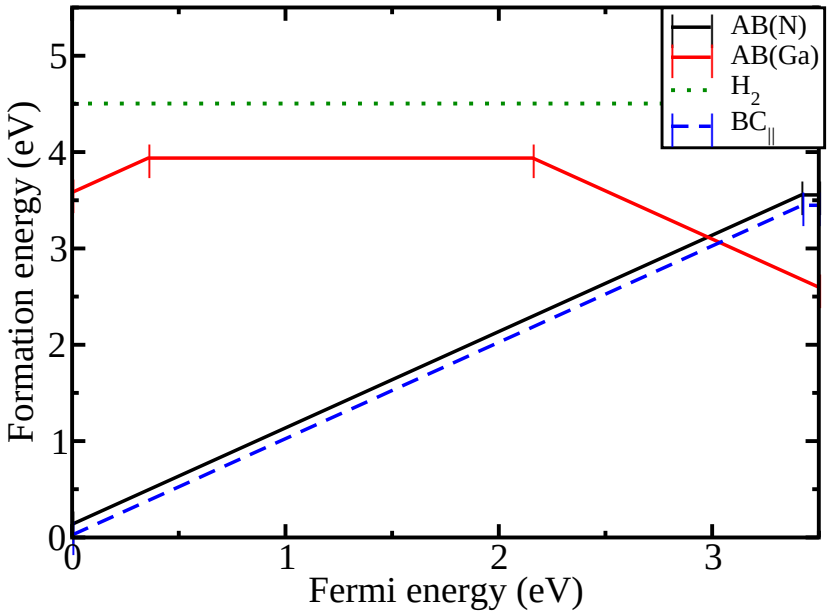

Figure 33: Formation energies of several configurations of interstitial hydrogen H_i as a function of the Fermi level. Labels correspond to AB(N) – hydrogen in anti-bonding nitrogen site, AB(Ga) – anti-bonding gallium site, hydrogen molecule H_2 (dotted line), and $BC_{\|}$ - bond-center site along the wurtzite *c*-axis.

4.2.4.2. General Properties of Carbon in GaN

Carbon can form several defects in GaN, the most stable of which are the C_N acceptor and the C_NO_N deep donor complex. Our calculations of electronic and optical properties of the C_N acceptor and the C_NO_N complex, and their comparison to experimental measurements have been published elsewhere.[92,93] Here, we only briefly summarize them in order to discuss the properties related to their possible interaction with hydrogen. The C_N defect creates two transition levels in the bandgap: the 0/+ transition level at 0.48 eV above the VBM, and the –/0 transition level at 1.04 above the VBM. This suggests that for Fermi level above ~1 eV in the bandgap, a negatively charged C_N defect can attract a positively charged hydrogen interstitial H_i, forming the C_N-H_i complex. Another carbon defect is the C_NO_N complex, which is a deep donor, with the 0/+ transition level calculated at 0.75 eV above the VBM (as well as the +/2+ level at ~0.14 eV above the VBM). Optical transitions via the 0/+ transition level of C_NO_N complex and –/0 transition level of C_N defect were suggested to lead to two slightly different YL bands.[93] As we show below, the interstitial hydrogen can form a weakly stable complex with C_NO_N and C_N, which also explains the experimentally observed photo-bleaching of the BL2 band.

4.2.4.3. Properties of the C_NO_N-H_i complex

Here, we consider the possibility of hydrogen creating a stable complex with one of the sources of the YL band, the C_NO_N complex. The C_NO_N-H_i complex in turn could be responsible for the BL2 band. As was mentioned above, the BL2 band is observed in high-resistivity GaN, including undoped, Fe-doped, and C-doped samples.[79] According to early PL excitation studies, Fe_{Ga} acceptors in GaN have the -/0 transition level at ~1.0 eV below the CBM.[214] More recent measurements on high-quality GaN:Fe samples suggested the value of 0.68±0.06 eV below the CBM for this transition level.[215] The Fe_{Ga} acceptors compensate the shallow donors (such as oxygen), making the samples semi-insulating.[216] In this case, the Fermi level in GaN:Fe should be located in the vicinity of this transition level. In carbon-doped GaN samples, the Fermi level is expected to be located near the C_N or C_NO_N transition levels; i.e., 0.9-1.0 eV above the VBM. However, the BL2 band is not observed in some carbon-doped GaN samples,[93,217] and is observed in all iron-doped samples studied in this work. This indicates that the BL2 band is sensitive to the position of the Fermi energy (similarly to the case of the nitrogen vacancy[35]). In other words, the BL2 related defect is not formed when the Fermi level is pinned by the carbon-related transition levels (~0.9-1.0 above the VBM), and it is formed when the Fermi level is pinned by the Fe_{Ga} acceptor levels (~0.6 below the CBM). As we show in Fig. 35, within certain range of Fermi energies interstitial hydrogen can bind to the C_NO_N complex (or C_N) as mobile H_i diffuses throughout the sample.

4.2.4.3.a. Atomic Configuration of the C_NO_N-H_i complex

Our calculations show that hydrogen has numerous metastable sites around the carbon atom in the C_NO_N complex and around isolated C_N defect. These complexes with hydrogen occupying different sites show similar electronic and optical properties. Figure 34 shows the three lowest-energy basic structures of the C_NO_N-H_i complex, with three locations of hydrogen. The lowest energy configuration is formed when H is attached to the C_NO_N complex on the carbon side at the carbon anti-bonding site, closest to oxygen (lower right H in Fig. 34). Other C anti-bonding sites, where the C-H bond is pointing away from the oxygen (lower left H in Fig. 34), are only 0.1 eV higher in energy. There are other lower symmetry sites, for example an off C anti-bonding site, that is ~0.2 eV higher in energy (higher H in Fig. 34), which has a number of equivalent sites around the C atom. The lowest energy bond-center site is the Ga-N bond center site, where H is in the bond center between the Ga atom (common to C and O) and N along the wurtzite *c*-axis (not shown here). However, this bond center C_NO_N-H_i complex has an energy of 0.78 eV higher than the lowest energy anti-bonding geometry, indicating that bond-center C_NO_N-H_i complexes are energetically unfavorable.

We have also performed calculations for hydrogen occupying all possible bond-center, as well as O and Ga anti-bonding sites, which are nearest neighbors of the C_NO_N complex; however, these hydrogen geometries lead to energies higher by 1 eV or more. Thus, interstitial hydrogen is most likely attached to carbon, with numerous carbon anti-bonding H_i geometries of the C_NO_N-H_i complex of similar energies that could be realized in experiment. Hydrogen exhibits similar behavior around the isolated carbon acceptor C_N and can form stable C_N-H_i complexes. However, the presence of oxygen modifies the electronic states, leading to somewhat different optical properties for the C_NO_N-H_i and C_N-H_i defects, as discussed below.

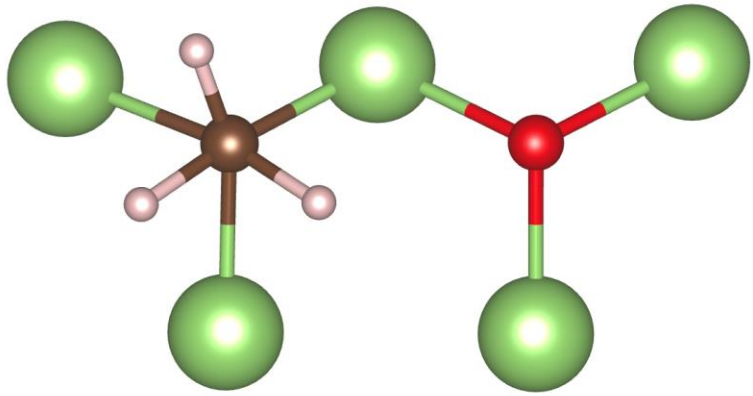

Figure 34: Three low energy structures of the C_NO_N-H_i complex. Large green atoms are Ga while the medium sized atom on the left (in brown color) is C and the atom in the right (in red color) is O. Small atoms are H, occupying C anti-bonding sites (lower H atoms), and off anti-bonding site (higher H atom) which has a number of equivalent positions.

4.2.4.3.b. Formation energy of the C_NO_N-H_i complex

Figure 35 shows the formation energies of the C_NO_N-H_i and C_N-H_i complexes in their lowest energy configurations, compared to the lowest energy configurations of interstitial hydrogen, C_N acceptor, and the C_NO_N complex. The C_NO_N-H_i and C_NO_N complexes have similar formation energies, for Fermi levels below ~0.8 eV in the gap, where both complexes are positively charged in the lowest energy charge state. For higher Fermi levels, the C_NO_N complex is more energetically favorable, since it becomes neutral, while the C_NO_N-H_i complex behaves as a shallow donor with a computed donor level at ~0.1 eV below the conduction band. The C_N-H_i complex is more favorable for Fermi levels above 2 eV in the bandgap, where the formation energy of the shallow donor C_NO_N-H_i complex increases considerably due to its + charge state. The C_N-H_i complex creates the 0/+ deep donor transition level at 0.3 eV above the VBM. This transition level is in the energy range that makes it a possible candidate for the BL2 band. The C_NO_N-H_i complex creates the +/2+ transition level due to a weakly localized defect state close to the valence band in the + charge state, and a localized hole in the 2+ state, leading to large relaxation energies between the two states. The weakly localized nature of the electron in the defect state makes it problematic to calculate this transition level accurately (as with all shallow states). HSE calculations with supercells containing 300 atoms suggest this +/2+ transition level is at roughly 0.06-0.1 eV above the VBM. Subsequent calculations show that optical transitions through this +/2+ level of the C_NO_N-H_i complex also can explain the observed properties of the BL2 band.

The C_NO_N complex is neutral for Fermi levels above 0.8 eV, therefore there is no long-range electrostatic attraction between this complex and interstitial hydrogen. Nevertheless, the calculated binding energy of the C_NO_N-H_i complex (lower panel in Fig. 35) is ~0.9 eV for most

Fermi levels in the bandgap, indicating a relatively stable complex. The isolated C_N is negatively charged for Fermi levels above 1 eV, and can attract positive hydrogen for Fermi levels below 3 eV, above which hydrogen becomes negative. The C_N-H_i complex is slightly more stable for a similar range of Fermi energies; i.e., between 1 and 3 eV in the bandgap, and is ~1.2 eV. Since the binding energy provides information about the energy differences between the equilibrium lowest energy configurations of the complex and its constituents, the diffusion barriers for hydrogen dissociation are also needed to estimate the stability of the complex (discussed below).

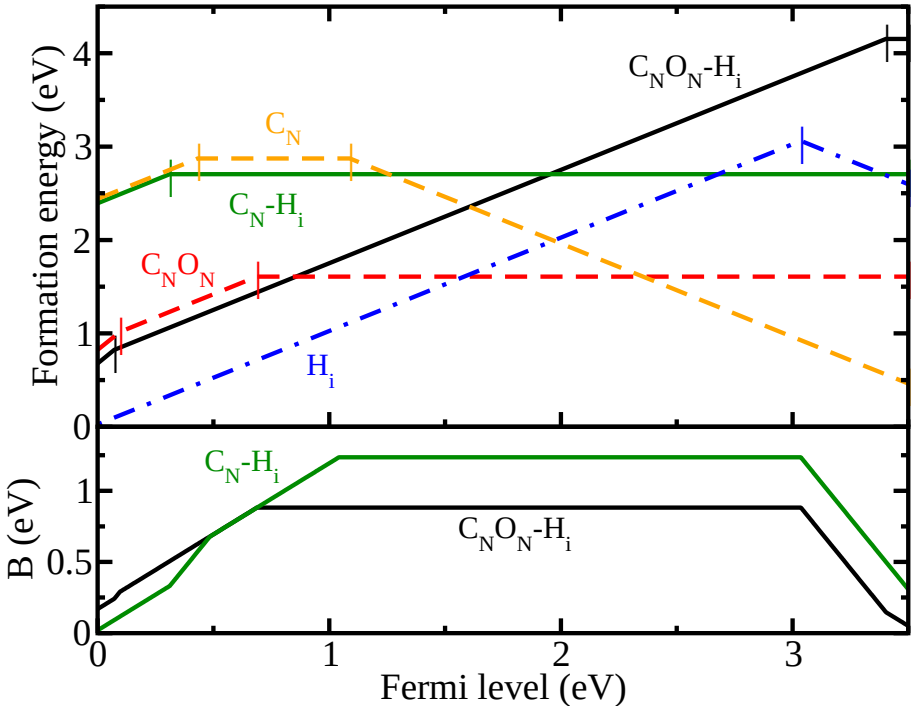

Figure 35: Formation energies of C_NO_N-H_i, C_N-H_i, C_NO_N, C_N, and hydrogen interstitial H_i as a function of the Fermi level in the GaN band gap (upper panel). The Fermi energies where lines change slope correspond to the thermodynamic transition levels. Binding energy (B) of C_NO_N-H_i and C_N-H_i complexes as a function of the Fermi level in the gap (lower panel).

4.2.4.3.c. Optics of the C_NO_N-H_i complex

Calculated results for optical transitions via the C_NO_N-H_i complex are shown in Figure 36. After optical band-to-band excitation (upward vertical arrow), a free hole is unlikely to be captured by this defect because it is positively charged in its ground state, and the efficiency of such a recombination channel would be very low. However, since the C_NO_N-H_i complex also has a shallow donor state, a free electron can be captured first by the $(C_NO_N\text{-}H_i)^+$ defect at its 0/+ shallow donor level. This would transfer the C_NO_N-H_i complex into a neutral excited state; i.e., $(C_NO_N\text{-}H_i + e^*)^0$, where e^* represents a weakly bound electron. Subsequently, a hole can be captured by the neutral complex to the +/2+ defect state, lowering the system's energy by roughly 0.1 eV. The electron-hole recombination will then occur as an internal transition, where a weakly localized electron recombines with the hole localized on the same C_NO_N-H_i defect. Preliminary time-resolved PL experiments indicate that at least in some high-resistivity GaN samples, the decay of the BL2 band after a pulse excitation is nearly exponential, with a characteristic lifetime of about 0.4 μs at temperatures between 15 and 100 K,[80] which agrees with the above-described internal transition mechanism. Detailed studies of the time-resolved PL are underway.

The optical transition calculations performed for the lowest energy structure of this complex yield a PL maximum at 3.03 eV, which is very close to the experimentally observed BL2 band maximum (at 2.98-3.03 eV in different samples). The calculated excitation energy of 3.69 eV is larger than the bandgap, and therefore cannot be observed experimentally. The thermodynamic transition level for the +/2+ transition level responsible for this optical line is very close to the valence band, leading to a significant delocalization error. Examining the electronic structure of the defect shows one distinct defect state roughly 0.1 eV above the

valence band, which in the + charge state has a delocalized (or weakly localized) wave function. As a consequence, it can only be estimated that this electronic state is within 0.1 eV from the VBM. This leads to the estimated value of the ZPL of about 3.4 eV. The Franck-Condon shift is calculated to be 0.37 eV. These values are in good agreement with the experimentally observed values of 3.33-3.34 eV for the ZPL of the BL2 band, suggesting that the C_NO_N-H_i complex is the defect responsible for it.

The existence of other possible carbon anti-bonding sites for hydrogen in the complex leads to varying emission energies within ~0.15 eV (the bond-center H_i geometry yields significantly lower PL maximum at 2.75 eV), which may lead to additional broadening of the PL band due to statistical averaging of different positions of hydrogen in the sample. Moreover, defects with different positions of hydrogen would have different barriers for the complex dissociation. Thus, we may expect a slight shift of the BL2 band with prolonged UV exposure. To verify the above predictions, we have carried out an experiment with a focused laser beam (with a diameter of the spot of about 0.2 mm) to cause greater bleaching of the BL2 band. The results are shown in Fig. 37. After 2.5 hours of this UV exposure, the intensity of the BL2 band decreased by a factor of four, and the intensity of the YL band increased by the same factor. Interestingly, the maximum of the BL2 band shifted to lower photon energies by about 50 meV and the full width at half maximum of the BL2 band increased from 386 to 460 meV after the prolonged UV exposure (Fig. 37). This indicates that various similar configurations of BL2 defect could exist, as discussed above. Some of these are being destroyed under UV illumination, while others appear to be more stable which leads to the broadening and shift of the BL2.

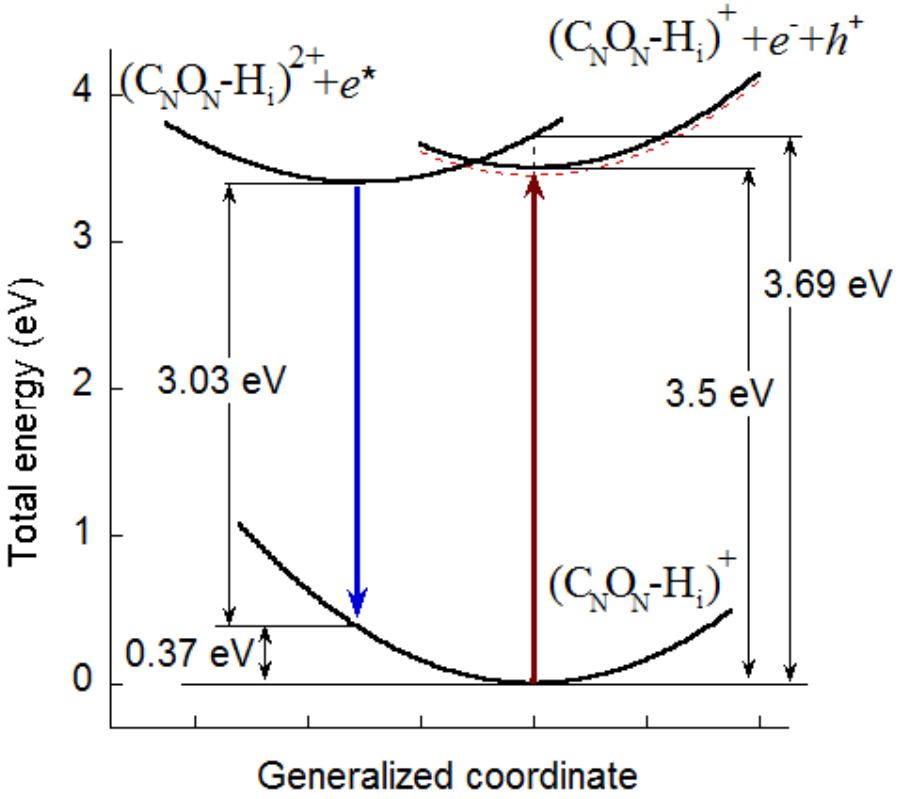

Figure 36: Configuration coordinate diagram and calculated optical transitions for the C_NO_N-H_i complex. The upward vertical arrow represents the band-to-band excitation, with the generation of an electron-hole pair. The following transition of the system from the solid parabola to the dashed parabola corresponds to the capture of a free electron at the 0/+ shallow level of the C_NO_N-H_i defect. The transition from the upper-right parabola (solid one if the electron is free or dashed one if the electron is captured at the 0/+ level) to the upper-left parabola corresponds to the nonradiative capture of a free hole at the +/2+ transition level of the C_NO_N-H_i complex. The thermodynamic +/2+ transition level is at ~0.1 eV above the VBM, and the Franck-Condon shift is 0.37 eV. The downward arrow represents the optical recombination producing a PL band with a maximum at 3.03 eV and ZPL at 3.4 eV. Resonant excitation of the C_NO_N-H_i complex is expected to produce a PL excitation band with a maximum at 3.69 eV, which cannot be observed experimentally since the energy is higher than the GaN bandgap.

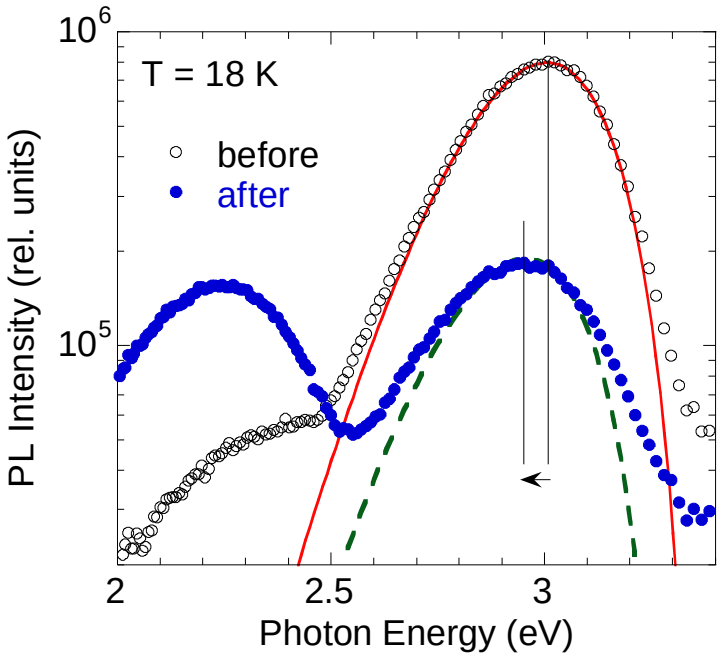

Figure 37: The YL and BL2 bands before and after 2.5 hours of UV exposure with a focused HeCd laser at $P_{exc} \approx 100$ W/cm^2 and $T = 18$ K. The measurements are done at $P_{exc} \approx 0.4$ W/cm^2. The solid line is calculated using Eq. (1) with the following parameters: $I_0^{PL} = 8\times10^5$, $S_e = 4.5$, $E_0 + 0.5\hbar\Omega = 3.35$eV, $\hbar\omega_{\max} = 3.005$ eV. The dashed line is identical to the solid line but shifted vertically (by a factor of 4.5) and horizontally (by 50 meV). The vertical lines show positions of the BL2 band maximum, and the arrow indicates the shift of the BL2 band maximum by about 50 meV.

4.2.4.4. Properties of the C_N-H_i complex

According to the above-discussed model in section 4.2.4.3.a, hydrogen binds to the carbon side of the C_NO_N complex, whereas the complexes with H_i near oxygen have substantially higher energy. A question arises concerning what role the oxygen plays in the electronic and optical properties of the defect responsible for the BL2 band. Below, we examine the C_N-H_i complex as another possible candidate for the BL2 band.

As in the previous case, hydrogen has numerous metastable locations around C_N, which appears to be common for interstitial hydrogen. The lowest energy C_N- H_i complex is formed for H_i located ~1.1 Å away from carbon, at one of the three equivalent anti-bonding sites. H_i occupying fourth anti-bonding site along the wurtzite *c*-axis has an energy that is 0.27 eV higher. The C-Ga bond center site has 0.6 eV higher energy, with the C-Ga bond stretched from ~2 Å to 2.8Å. The situation is similar to the case of the C_NO_N-H_i complex, creating a number of complex geometries with slightly different transition levels, all of which could be realized according to their formation energies.

4.2.4.4.a. Optical Transitions of the C_N-H_i complex

Figure 38 shows optical transitions via the C_N-H_i complex. After band-to-band excitation (the upward arrow), the neutral C_N-H_i defect captures a free hole (transition from the upper right parabola to the upper left parabola) due to the existence of the 0/+ level at 0.3 eV above the VBM. The defect becomes positively charged and the free electron recombines with the hole localized on the defect, producing a PL band. The lowest formation energy C_N-H_i defect would produce the PL band with a maximum at 2.72 eV, which is ~0.3 eV lower than the observed BL2 band maximum in experiment. However, some slightly higher energy complex geometries do show optical transitions closer to the experimental results for the BL2 band. For example, C_N-H_i with hydrogen at the anti-bonding site along the wurtzite *c*-axis has a thermodynamic transition level at 0.24 eV above the VBM, and it would produce a PL band with a maximum at 2.95 eV and a ZPL at 3.26 eV. This site is only 0.23 eV higher in energy than the lowest energy anti-bonding site described above, and therefore it is possible that it contributes to the BL2 band. In contrast to the C_NO_N-H_i case, the excitation energy for this defect is below the band gap, and has a value of 3.38 eV.

In all geometries, the C_NO_N-H_i defect creates two transition levels in the bandgap, leading to the predicted internal optical transition with electron-hole pair localized on the same defect. At the same time, the C_N-H_i defect creates only one deep donor level, suggesting an external transition, with conduction band electron (or a shallow donor-bound electron) recombining with the hole localized on the defect. It should be noted that in its excited state, positively charged C_N-H_i defect can also create hydrogenic levels below the CBM, since it can weakly bind a conduction band electron. In this case an internal transition is also possible. Early experiments on time-resolved PL revealed that the decay of the BL2 intensity after pulse excitation is nearly

exponential, even at 15 K. The characteristic PL lifetime for the BL2 band in undoped GaN has been determined to be about 400 ns.[79] This is in contrast to typical transitions from a shallow donor to a deep acceptor for a majority of defect-related PL bands in GaN.[4] These donor-acceptor pair (DAP)-type transitions produce non-exponential and a very slow decay of PL intensity at low temperature due to random separations of bound electrons and holes in DAPs. Transitions from the conduction band to the deep defect levels can be ignored at low temperatures for both *n*-type conductive GaN samples and the high-resistivity samples, because non-radiative capture of photo-generated electrons by shallow and deep donors is much faster than the radiative recombination. Our preliminary experimental results on time-resolved PL indicate an internal transition, because the decay of the BL2 intensity in time-resolved PL measurements is nearly exponential at low temperatures.

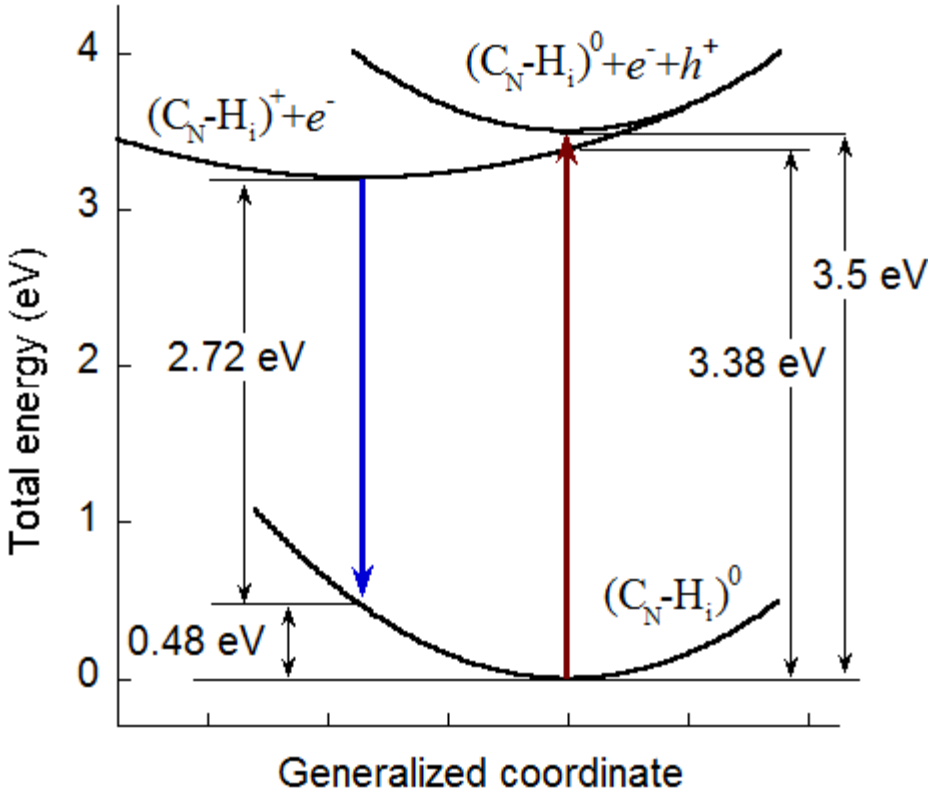

Figure 38: Configuration coordinate diagram and calculated optical transitions for the C_N-H_i complex. The upward vertical arrow represents the band-to-band excitation, with the generation of an electron-hole pair. The transition from the upper-right parabola to the upper-left parabola corresponds to the nonradiative capture of a free hole at the 0/+ transition level of the C_N-H_i complex. The thermodynamic 0/+ transition level is at 0.3 eV above the VBM, and the Franck-Condon shift is 0.48 eV. The downward arrow represents the optical recombination producing a PL band with a maximum at 2.72 eV and ZPL at 3.2 eV. Resonant excitation of the C_N-H_i complex would produce a PL excitation band with a maximum at 3.38 eV.

4.2.4.5. Summary of thermodynamic and optical transitions of various configurations of the carbon-hydrogen related complexes

Figure 39 shows the schematic band diagram with the thermodynamic transition levels and optical transitions via the C_NO_N-H_i and C_N-H_i complexes. As mentioned above, hydrogen has numerous locations around carbon, with similar formation energies, leading to a number of possible transition levels close to each other that can contribute to the PL, since all of these possible geometries can be realized in experiment. This would lead to a broadening of the PL bandwidth, in addition to the common electron-phonon coupling. In both cases, a set of the most typical hydrogen sites produces transition levels within ~0.2 eV from each other, as shown with the shaded areas in Fig. 39. Hydrogen sites with formation energies larger than 1 eV can have different electronic structure, and are not shown here. For both defects, the lowest energy defects create transition levels that are the lowest in the gap (~0.1 and 0.24 eV for the C_NO_N-H_i and C_N-H_i complexes, respectively). The lowest energy structures for both defects are anti-bonding sites: in the C_N-H_i complex, interstitial H is located in one of the three equivalent anti-bonding Ga sites and in the C_NO_N-H_i complex the hydrogen-carbon bond is pointing toward the oxygen. Next in formation energy, there are a few alternative anti-bonding sites, that are about 0.1 eV higher for each complex, also producing transition levels ~0.1 higher. Finally, the highest formation energy defects states with transition levels of 0.35 and 0.44 eV shown in Fig. 39, correspond to the Ga-N bond center site for the C_NO_N-H_i, complex and the C-Ga bond-center site in the C_N-H_i complex. These configurations are 0.78 eV and 0.6 eV higher in formation energy than the lowest energy geometry, respectively. The presence of oxygen lowers all transition levels by ~0.1-0.15 eV, bringing the PL maximum created by the C_NO_N-H_i complex closer to experiment.

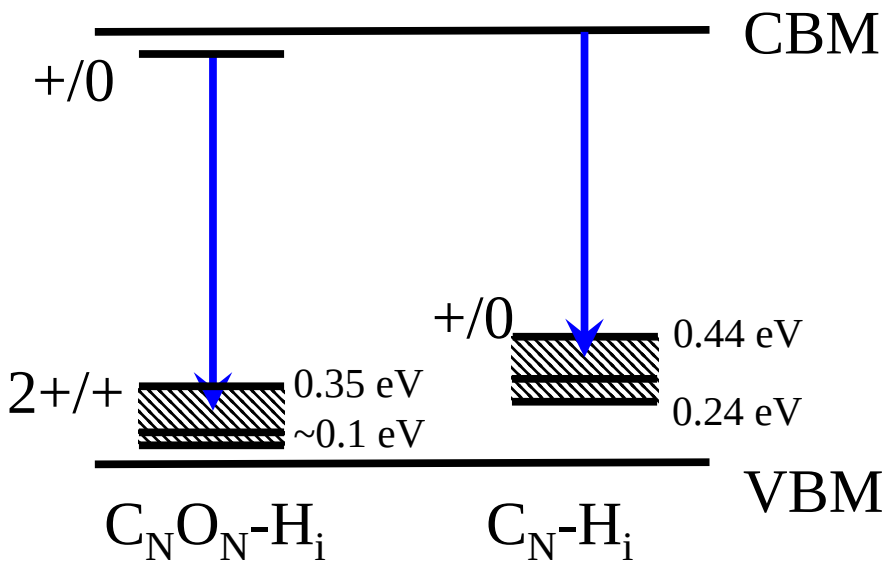

Figure 39: Schematic band diagram illustrating the optical transitions via C_NO_N-H_i and C_N-H_i complexes. A variety of possible positions of interstitial hydrogen leads to a variety of similar transition levels, which would be distributed in the sample according to their formation energies. Shaded areas represent the energy range within which most low energy defect configurations vary.

4.2.4.6. Stability of the C_NO_N-H_i complex and PL photo-bleaching

The experimentally observed BL2 band exhibits photo-bleaching under laser exposure, with the PL intensity dropping by a factor of four after 290 min of UV illumination, while the YL band intensity increases by approximately the same amount (Fig. 37). This can be explained with the assumption that the BL2 source is unstable and is being destroyed, either by the UV laser or subsequent non-radiative recombination, leaving the source of YL as a byproduct of this decomposition. Since our HSE calculations indicate that the C_NO_N-H_i complex is a possible source of the BL2 band, we have performed diffusion barriers calculations, for the dissociation of hydrogen from the complex. The diffusion barriers were calculated using the nudged elastic band method[218] within the GGA approximation to the DFT. The initial most stable geometry for the C_NO_N-H_i complex is taken from above HSE calculations, with hydrogen occupying the anti-bonding C site. The isolated hydrogen in the + state has similar formation energies in both the bond-center and nitrogen anti-bonding sites; therefore, these two final geometries are also calculated. All calculations were performed in the + charge state of the complex.

Figure 40 shows the diffusion barriers, as well as the diffusion path of the H atom in GaN containing the C_NO_N-H_i complex. When being dissociated from the C_NO_N-H_i complex, hydrogen has to overcome the potential barrier of 1.1 eV. It should be noted that these GGA results imply slightly lower binding energy of the hydrogen atom to C_NO_N within GGA approximation, about 0.75 eV (roughly the energy difference between points AB(N) and AB(C) points in Fig. 40). More accurate HSE calculations (when possible to perform) usually yield larger diffusion barriers for hydrogen in GaN by ~0.1-0.2 eV,[213] suggesting that actual dissociation barrier could be slightly larger. An additional jump from the nitrogen anti-bonding site into the bond-center site requires an additional 0.25 eV. As shown in Fig. 40, the bond-center site is unstable in GGA

(there is no energy minimum); however, HSE calculations show that it has an energy that is actually a little lower than that of the nitrogen anti-bonding site. Therefore a more accurate HSE calculation is expected to predict the 0.25 eV barrier between these two sites of similar energy.

The potential barrier of ~1.1 eV for detachment of hydrogen from the C_NO_N-H_i complex suggests that it is weakly stable at room temperature. For optical transitions corresponding to the PL maximum, the thermal energy released following the radiative recombination via the C_NO_N-H_i complex (equal to Franck-Condon shift of 0.37 eV) is not sufficient for the dissociation. However, the significant PL bandwidth allows for the following explanation of the hydrogen detachment from the complex via the recombination enhanced defect reaction mechanism,[219,187] as shown schematically in Fig. 41. Upon photon absorption the defect is transferred into the 2+ charge state. In this 2+ charge state there is a certain spread of the defect vibrational wave function, which defines the PL bandwidth. Therefore, radiative recombination events occur (albeit with lower probabilities) at both sides of the upper parabola. Thus, the vibrational lattice energy (ZPL minus the photon energy), larger than 1.1 eV dissociation barrier, can be released following the radiative recombination. This corresponds to the radiative transitions (left downward arrow in Fig. 41) with photon energies lower than 2.3 eV, leaving enough lattice relaxation energy for complex dissociation. Thus, complex dissociation via relaxation following the radiative transition, turns the C_NO_N-H_i or C_N-H_i (the BL2 band source) into the C_NO_N or C_N (the YL band source). At room temperature, mobile hydrogen will be moving throughout the sample until it is attached again to a more stable site, such as the C_NO_N or C_N defect. Thus, photo-bleaching under illumination should be followed by a slow recovery of the BL2 band in the dark at room temperature. In experiment, the BL2 band is bleached in a few hundred minutes at low temperature, followed by the BL2 recovery, which takes from several hours to several

days at room temperature (the restoration time was sample-dependent) Thus, optical transitions and energetics of the C_NO_N-H_i complex (or the C_N-H_i complex) explain all the experimentally observed aspects of BL2, both qualitatively and quantitatively.

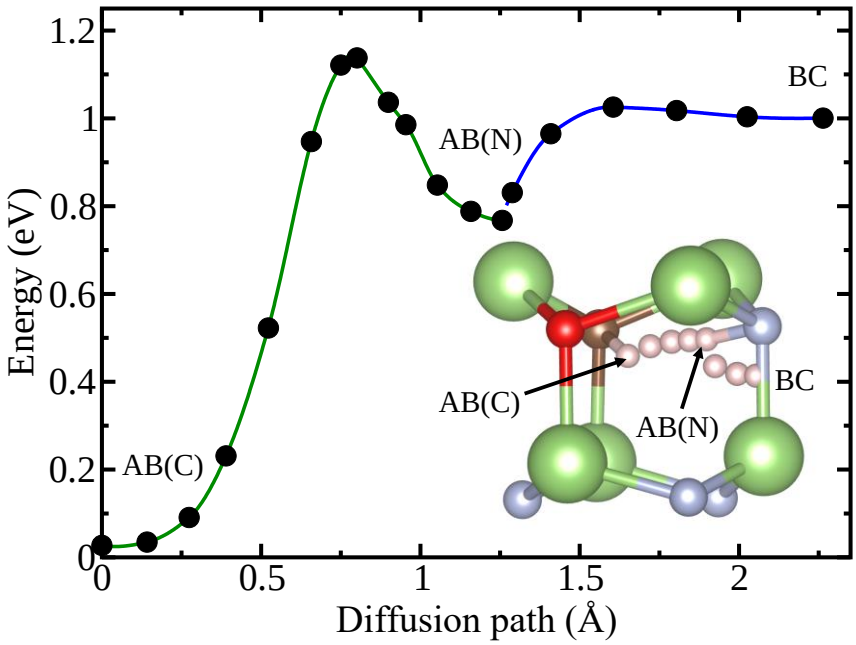

Figure 40: GGA Calculated hydrogen diffusion barriers determining the dissociation energies of the C_NO_N-H_i complex, and the corresponding diffusion path. Initially bound to the complex at the anti-bonding carbon site, labeled AB(C), the hydrogen atom can jump into the neighboring anti-bonding nitrogen site AB(N). Subsequently, the hydrogen can jump into the Ga-N bond-center site, labeled BC.

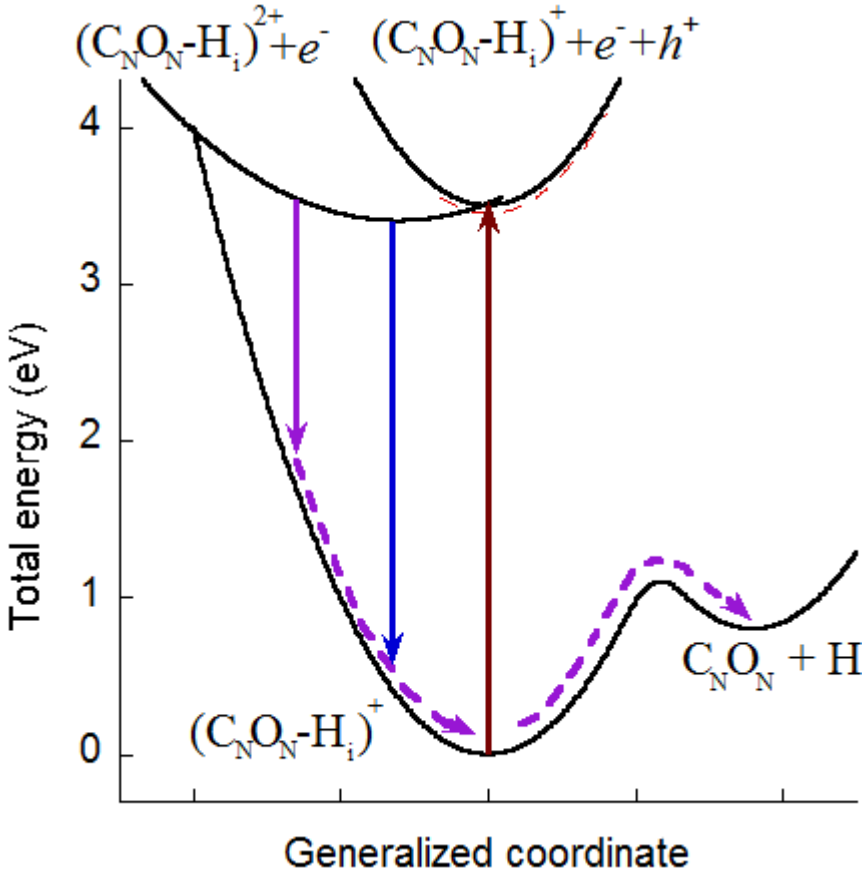

Figure 41: The configuration coordinate diagram schematically explaining the bleaching of the BL2 band. The radiative transition producing the BL2 band is shown with right downward arrow. However, a smaller fraction of recombinations occurs with lower photon energies, shown with left downward arrow, which can cause the dissociation of the complex (the processes shown with dashed arrows).

4.2.5. Concluding remarks regarding the BL2 band in GaN

In the last section of this dissertation, we performed first-principles calculations to explain the properties of the BL2 band that peaks at 3.0 eV and is only observed in high-resistivity GaN samples grown by MOCVD or HVPE techniques. The characteristic feature of the BL2 band is its significant bleaching under UV exposure, with the concurrent emergence of the YL band around 2.2 eV. Since the YL band is attributed to either the C_N defect or the C_NO_N complex, and both the MOCVD and HVPE growths may produce the abundance of hydrogen in GaN samples, we have explored the possibilities that the C_N-H_i and C_NO_N-H_i complexes dissociate under UV exposure by releasing a hydrogen atom. The calculated and experimental results agree very well, especially for the C_NO_N-H_i complex as the source of the BL2 band. In particular, the calculated PL band maximum and the position of the zero-phonon line for the C_NO_N-H_i complex (3.03 eV and 3.4 eV, respectively) are almost identical to the values observed experimentally (3.0 eV and 3.33 eV, respectively), while the predicted values for the C_N-H_i complex (2.73 eV and 3.1 eV, respectively) are off by 0.2-0.3 eV. Moreover, we predict that an electron is first captured by a shallow donor level of the C_NO_N-H_i complex, located at about 0.1 eV below the conduction band, and it recombines with a hole captured at the level close to the valence band. Such internal transition would produce PL decaying exponentially in time after a laser pulse. Preliminary time-resolved PL experiments indicated such behavior. Furthermore, in Fe-doped GaN, where the BL2 band is usually observed, the Fermi level is expected to be located at about 0.6 eV below the conduction band. In these conditions, both C_N-H_i and C_NO_N-H_i complexes are stable with binding energies of 1.2 and 0.9 eV, respectively. The barrier for detachment of hydrogen in C_NO_N-H_i complex is about 1.1 eV. We suggest that the bleaching of the BL2 band occurs via the photo-induced defect reaction, whereby a fraction of electron-hole

recombinations at the C_N-H_i or C_NO_N-H_i complex, occurring at the lower energy tail of PL spectrum, leave enough lattice relaxation energy to detach the hydrogen atom.

Future work

Throughout this dissertation, we have seen that theoretical predictions of thermodynamic transitions $\varepsilon_T\left(q_1/q_2\right)$ or optical transition of a defect between two different charge states q_1 and q_2 can either be obtained from (a) the single particle eigenvalue based on Janak's theorem,[220,221,222] or (b) the formation energy differences of the defect in charge states q_1 and q_2.

The latter method (method (b)) requires total energy calculations in SCs which involves spurious interactions between the compensating background (jellium) and the periodically repeated images of the defect in non-neutral charged systems[173]. Although using the formation energy method reproduces the experimental bandgap of 3.5 eV for GaN, it yields incorrect ionization potential (*IP*) since neither the CBM nor the VBM values computed within the KS method are physically meaningful. Furthermore, it is also found that the position of calculated transition levels of deep defects such as V_{Ga} (Table 1) or carbon related impurities,[91,92,93] is dependent on the method used for fictitious interactions corrections. An identical trend is also observed for defects close to band edges, i.e V_N, in which transitions levels are underestimated by approximately 0.2-0.3 eV across the bandgap, when compared to experiment.[73,75] Therefore, accurate computations of thermodynamic or optical transition levels of deep and shallow defects via the formation energy method might not be the most reliable of approaches.

The former method (b) involves Janak's theorem[134] which states that the *k*-th Kohn-Sham eigenvalue $\left(\varepsilon_k\right)$ is equal to the change in the KS total energy $\left(E_J\left[n_J\left(\vec{r}\right)\right]\right)$ with respect to the electron occupancy (q_k) of the *k*-th level,

$$\frac{\partial E_J\left[n_J\left(\vec{r}\right)\right]}{\partial q_k}=\varepsilon_k$$

By implementing Janak's theorem[134] in the SC formalism subjected to the "simplest" correction, i.e Madelung corrections, we obtain:

$$\frac{\partial E_J\left[n_J\left(\vec{r},q_k\right)\right]}{\partial q_k}=\underbrace{\int\left[\frac{-\nabla^2}{2}+u_{eff}^J\left(\vec{r}\right)\right]\frac{\partial n_J\left(\vec{r}\right)}{\partial q_k}d\vec{r}}_{\varepsilon_k}+\frac{\partial}{\partial q_k}\left[\frac{q^2\alpha}{2\varepsilon\left(V_{SC}\right)^{1/3}}\right]$$

$$\Leftrightarrow\frac{\partial E_J\left[n_J\left(\vec{r},q_k\right)\right]}{\partial q_k}=\varepsilon_k+2q\frac{\alpha}{2\varepsilon\left(V_{SC}\right)^{1/3}}$$

According to the above equation, by dropping or adding an electron into a SC subjected to Madelung corrections, its corresponding *k*-th KS eigenvalue will be shifted by a value of $\pm\,2\frac{\alpha}{2\varepsilon\left(V_{SC}\right)^{1/3}}$. In case of defects fairly close to band edges, this might consequently make the eigenvalue of the defect state resonant with the CBM/VBM, which would hence yield inaccurate thermodynamic and optical transition levels.

In order to circumvent the difficulties encountered using either previously discussed methods [(a) or (b)], future work regarding the restoration of the experimental ionization potential of GaN by tuning the screening parameter *w* from the HSE functional while using the standard value of the amount of exact exchange a_1 = 0.25 can be performed. Although in such method, the position of the CBM is incorrect, by reestablishing the correct *IP*, we are defining the correct value of the VBM with respect to vacuum level and hence possibly setting up a benchmark for which better comparison between experiment and theory could be done for future investigation of thermodynamic and optical transition levels of defects in semiconductors.

References

[1] S. Nakamura, T. Mukai, and M. Senoh, Appl. Phys. Lett. **64**, 1687 (1994).
[2] S. Nakamura, and G. Fasol, *The Blue Laser-Diode—GaN Based Light Emitters and Lasers* (Springer-Verlag, Berlin, 1997).
[3] R. Dahal, J. Li, K. Aryal, J. Y. Lin, and H. X. Jiang, Appl. Phys. Lett. **97**, 073115 (2010).
[4] M. A. Reshchikov and H. Morkoc, J. Appl. Phys. **97**, 061301 (2005).
[5] W. Kohn and L. J. Sham, Phys. Rev. Lett. **140**, 1133 (1965).
[6] R. O. Jones, *Introduction to Density Functional Theory and Exchange-Correlation Energy Functionals* Computational Nanoscience: Do It Yourself! (Jülich, NIC Series Vol. **31**, 2006).
[7] R. M. Martin, *Electronic Structure: Basic Theory and Practical Methods*, (Cambridge University Press, New York, 2004).
[8] J. Paier, M. Marsman, K. Hummer, G. Kresse, I. C. Gerber , and J. G. Ángyán, J. Chem. Phys. **124**, 154709 (2006).
[9] J. P. Perdew, Phys. Rev. Let. **55**, 1665 (1985).
[10] F. Herman, J. P. Dyke, and I. P. Ortenburger, Phys, Rev. Lett. **22**, 807 (1969).
[11] P. S. Svendsen and U. von Barth, Phys. Rev. B **54**, 17402 (1996).
[12] J. Tao, J.P. Perdew, V.N. Staroverov and G.E. Scuseria, Phys. Rev. Lett. **91**, 14640 (2003).
[13] J. P. Perdew, A. Ruzsinszky, G. I. Csonka, O. A. Vydrov, G. E. Scuseria, L. A. Constantin, X. Zhou, and K. Burke, Phys. Rev. Lett. **100**, 136406 (2008).
[14] J. P. Perdew, K. Burke, M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).
[15] K. Hummer, J. Harl, and G. Kresse, Phys. Rev. B **80**, 115205 (2009).
[16] W. Kohn, Rev. Mod. Phys. **71**, 1253 (1999).
[17] R. G. Parr and W. Yang, Phys. Chem. **46**, 701(1995).
[18] P. G.-Giorgi, M. Seidl, Phys. Chem. Chem. Phys. **12**, 14405 (2010).
[19] R. Armiento and A. E. Mattsson, Phys. Rev. B **72**, 085108 (2005).
[20] A. E. Mattsson, R. Armiento, and T. R. Mattsson, Phys. Rev. Lett. **101**, 239701 (2008).
[21] F. Aryasetiawan and O. Gunnarsson, Rep. Prog. Phys. **61**, 237 (1998).
[22] A. D. Becke, J. Chern. Phys. **96**, 2155 (1992); A. D. Becke, J. Chem. Phys. **98**, 1372 (1993); A. D. Becke, J. Chem. Phys. **107**, 8554 (1997).
[23] J. P. Perdew, M. Ernzerhof, and K. Burke, J. Chem. Phys. **105**, 9982 (1996).
[24] M. Ernzerhof, J. P. Perdew, and K. Burke, Int. J. Quantum Chem. **64**, 285 (1997).
[25] J. Heyd, Thesis for PHD, Rice University, (2004).
[26] J. Heyd, G. E. Scuseria, and M. Ernzerhof, J. Chem. Phys. **118**, 8207 (2003); J. Heyd, G. E. Scuseria, and M. Ernzerhof, J. Chem. Phys. **124**, 219906 (2006).
[27] J. E. Peralta, J. Heyd, G. E. Scuseria, and R. L. Martin, Phys. Rev. B **74**, 073101 (2006).
[28] J. Heyd and G. E. Scuseria, J. Chem. Phys. **121**, 1187 (2004).
[29] K. Hummer, A. Grüneis, and G. Kresse, Phys. Rev. B **75**, 195211 (2007).
[30] P. Deák, B. Aradi, T. Frauenheim, E. Janzén, and A. Gali, Phys. Rev. B **81**, 153203 (2010).
[31] A. V. Krukau, O. A. Vydrov, A. F. Izmaylov, and G. E. Scuseria, J. Chem Phys. **125**, 224106 (2006).
[32] A. F. Izmaylov and G. E. Scuseria, J. Chem. Phys. **127**, 144106 (2007).
[33] J. L. Lyons, A. Janotti, and C. G. Van de Walle, Appl. Phys. Lett. **97**, 152108 (2010).
[34] J. L. Lyons, A. Janotti, and C. G. Van de Walle, Phys. Rev. Lett. **108**, 156403 (2012).

[35] M. A. Reshchikov, D. O. Demchenko, J. D. McNamara, S. Fernández-Garrido, and R. Calarco, Phys. Rev. B **90**, 035207 (2014).
[36] G. Miceli and A. Pasquarello, Phys. Rev. B **93**, 165207 (2016).
[37] Y. Y. Sun, T. A. Abtew, P. Zhang, and S. B. Zhang, Phys. Rev. B **90**, 165301 (2014).
[38] S. T. Pandelides, Rev. Mod. Phys. **50**, 797 (1979).
[39] A. J. R. deKock, Philips Res. Rep. Suppl. **1**, 1 (1973).
[40] P. M. Petroff and B. L. Hartmann, Appl. Phys. Lett. **23**, 469 (1973).
[41] J. Lento, J-L. Mozos, and R. Nieminen, J. Phys.: Condens. Matter **14**, 2637 (2002).
[42] C. Persson, Y. –J. Zhao, S. Lany, and A. Zunger, Phys. Rev. B **72**, 035211, (2005).
[43] K. Saarinen, T. Laine, S. Kuisma, J. Nissilã, P. Hautojãrvi, L.Dobrzynski, J. M. Baranowski, K. Pakula, R. Stepniewski, M.Wojdak, A. Wysmolek, T. Suski, M. Leszczynski, I. Grzegory, and S. Porowski, Phys. Rev. Lett. **79**, 3030 (1997).
[44] K. Saarinen, T. Suski, I. Grzegory, and D. C. Look, Phys. Rev. B. **64**, 233201, (2001).
[45] J. Oila, V. Ranki, K. Saarinen, P. Hautojãrvi, J. Likonen, J. M. Baranowski, K. Pakula, M. Leszczynski, and I. Grzegory, Phys. Rev. B **63**, 045205 (2001).
[46] J. Oila, J. Kivioja, V. Ranki, K. Saarinen, D. C. Look, R. J. Molnar, S. S. Park, S. K. Lee, and J. Y. Han, Appl. Phys. Lett. **82**, 3433 (2003).
[47] A. Sedhain, J. Li, J. Y. Lin, and H. X. Jiang, Appl. Phys. Lett. **96**, 151902 (2010).
[48] J. Neugebauer and C. G. Van de Walle, Appl. Phys. Lett. **69**, 503 (1996).
[49] R. Armitage, W. Hong, Y. Qing, H. Feick, J. Gebauer, E. R. Weber, S. Hautakangas, and K. Saarinen, Appl. Phys. Lett. **82**, 3457 (2003).
[50] F. Reurings and F. Tuomisto, Proc. SPIE **6473**, 64730M (2007).
[51] M. Linde, S. J. Uftring, and G. D. Watkins, Phys. Rev. B **55**, R10177 (1997).
[52] I. A. Buyanova, Mt. Wagner, W. M. Chen, J. L. Lindström, B. Monemar, H. Amano, and I. Akasaki, MRS Internet J. Nitride Semicond. Res. **3**, 18 (1998).
[53] C. Bozdog, H. Przybylinska, G. D. Watkins, V. Härle, F. Scholz, M. Mayer, M. Kamp, R. J. Molnar, A. E. Wickenden, D. D. Koleske, and R. L. Henry, Phys. Rev. B **59**, 12479 (1999).
[54] G. D. Watkins, K. H. Chow, P. Johannesena, L. S. Vlasenkoa, C. Bozdoga, A. J. Zakrzewskia, M. Mizutab, H. Sunakawab, N. Kurodab, A. Usuib, Physica B **340-342**, 25-31 (2003).
[55] K. H. Chow, G. D. Watkins, A. Usui, and M. Mizuta, Phys. Rev. Lett. **85**, 2761 (2000).
[56] K. H. Chow, L. S. Vlasenko, P. Johannesen, C. Bozdog, G. D. Watkins, A. Usui, H. Sunakawa, C. Sasaoka, and M. Mizuta, Phys. Rev. B **69**, 045207 (2004).
[57] S. Hautakangas, J. Oila, M. Alatalo, K. Saarinen, L. Liszkay, D. Seghier, and H. P. Gislason, Phys. Rev. Lett. **90**, 137402 (2003).
[58] S. Zeng, G. N. Aliev, D. Wolverson, J. J. Davies, S. J. Bingham, D. A. Abdulmalik, P. G. Coleman, T. Wang, and P. J. Parbrook, Phys. Stat. Sol. **3**, No. 6, 1919–1922 (2006).
[59] D. W. Jenkins and J. D. Dow, Phys. Rev. B **39**, 3317 (1989).
[60] F. Gao, E. J. Bylaska, and W. J. Weber, Phys. Rev. B **70**, 245208 (2004).
[61] P. Perlin, T. Suski, H. Teisseyre, M. Leszczynski, I. Grzegory, J. Jun, S. Porowski, P. Boguskawski, J. Bernholc, J.C. Chervin, A. Polian, and T. D. Moustakas, Phys. Rev. B **75**, 296 (1995).
[62] J. Neugebauer and C. G. Van de Walle, Phys. Rev. B **50**, 8067 (1994).
[63] T. Matilla and R. M. Nieminen, Phys. Rev. B **54**, 16676, (1996).
[64] C. G. Van de Walle and J. Neugebauer, J. Appl. Phys. **95**, 3851 (2004).
[65] M. G. Ganchenkova and R. M. Nieminen, Phys. Rev. Lett. **96**, 196402 (2006).

[66] S. Limpijumnong and C. G. Van de Walle, Phys. Rev. B **69**, 035207 (2004).
[67] K Laaksonen, M. G. Ganchenkova and R. M. Nieminen, J. Phys.: Condens. Matter **21**, 015803 (2009).
[68] I. Gorczyca, A. Svane, and N. E. Christensen, Phys. Rev. B **60**, 8147 (1999).
[69] O. Gunnarsson, O. Jepsen, and O. K. Andersen, Phys. Rev. B **27**, 7144 (1983).
[70] Y. Gohda and A. Oshiyama, J. Phys. Soc. Jpn. **79**, 083705-3 (2010).
[71] Y. S. Puzyrev, T. Roy, M. Beck, B. R. Tuttle, R. D. Schrimpf, D. M. Fleetwood, and S. T. Pantelides, J. Appl. Phys. **109**, 034501 (2011).
[72] Q. Yan, A. Janotti, M. Scheffler, and C. G. Van de Walle, Appl. Phys. Lett. **100**, 142110 (2012).
[73] R. Gillen and J. Robertson, J. Phys.: Condens. Matter **25**, 405501 (2013).
[74] J. L. Lyons, A. Alkauskas, A. Janotti, and C.G. Van de Walle, Phys. Stat. Sol. **252**, 900 (2015).
[75] G. Miceli and A. Pasquerello, Microelectron. Eng. **147**, 51-54 (2015).
[76] A. Kyrtsos, M. Matsubara, and E. Bellotti, Phys. Rev. B. **93**, 245201 (2016).
[77] D. O. Demchenko and M. A. Reshchikov, Phys. Rev. B **88**, 115204 (2013).
[78] M. A. Reshchikov, Y. T. Moon, and H. Morkoç, Phys. Stat. Sol. (c) **7**, 2716 (2005).
[79] M. A. Reshchikov, Y. T. Moon, X. Gu, B. Nemeth, J. Nause, and H. Morkoç, Physica B **376-377**, 715 (2006).
[80] M. A. Reshchikov and H. Morkoç, Physica B. **376-377**, 428 (2006).
[81] C. H. Seager, A. F. Wright, J. Yu, and W. Götz, J. Appl. Phys. **92**, 6553 (2002).
[82] S. Dhar and S. Ghosh, Appl. Phys. Lett. **80**, 4519 (2002).
[83] S. A. Brown, R. J. Reeves, C. S. Haase, R. Cheung, C. Kirchner, and M. Kamp, Appl. Phys. Lett. **75**, 3285 (1999).
[84] S. Xu. G. Li, S. J. Chua, X. C. Wang, and W. Wang, Appl. Phys. Lett. **72**, 2451 (1998).
[85] R. Armitage, Q. Yang, and E. R. Weber, J. Appl. Phys. **97**, 073524 (2005).
[86] A. Y. Polyakov, M. Shin, J. A. Freitas, M. Skowronski, D. W. Greve, and R. G. Wilson, J. Appl. Phys. **80**, 6349 (1996).
[87] B. J. Ryan, M. O. Henry, E. McGlynn, J. Fryar, Physica. B **340-342**, 452 (2003).
[88] B. Kim, I. Kuskovsky, I. P. Herman, D. Li, and G. F. Neumark, J. Appl. Phys. **86**, 2034 (1999).
[89] L. Macht, J. L. Weyher, A. Grzegorczyk, and P. K. Larsen, Phys. Rev. B **71**, 073309 (2005).
[90] C. H. Seager, D. R. Tallant, J. Yu, and W. Götz, J. Luminescence **106**, 115 (2004).
[91] J. L. Lyons, A. Janotti, and C.G. Van de Walle, Phys. Rev. B **89**, 035204 (2014).
[92] D. O. Demchenko, I. C. Diallo, and M. A. Reshchikov, Phys. Rev. Lett. **110**, 087404 (2013).
[93] M. A. Reshchikov, D. O. Demchenko, A. Usikov, H. Helava, and Yu. Makarov, Phys. Rev. B, **90**, 235203 (2014).
[94] S. G. Christenson, W. Xie, Y. Y. Sun, and S. B. Zhang, J Appl. Phys. **118**, 135708 (2015).
[95] T. Mattila and R. M. Nieminen, Phys. Rev. B **55**, 9571 (1997).
[96] E. Engel and R. M. Dreizler, *Density Functional Theory: An Advanced Course* (Springer-Verlag, New York, 2011).
[97] R. O. Jones and O. Gunnarsson, Rev. Mod. Phys. **61**, 689 (1989).
[98] M. Born and J. R. Oppenheimer, Ann. Physik **84**, 458 (1927).
[99] A. R. Leach, *Molecular Modelling: Principles and Applications* (Pearson Prentice Hall, United Kingdom, 2001) 2nd Edition.

[100] D. R. Hartree, Math. Proc. Cam. Phil. Soc. **24**, 89-110 (1928).
[101] L. H. Thomas, Proc. Cambridge Phil. Roy. Soc. **23**, 542 (1927).
[102] E. Fermi, Rend. Accad. Naz. Lincei **6**, 602 (1927)
[103] P. A. M Dirac , Proc. Cambridge Philos. Soc. **26**, 376 (1930).
[104] J. C Slater, Phys. Rev. **81**, 385 (1951).
[105] K. Burke and friends, *The ABC of DFT*. Retrieved from http://dft.uci.edu/doc/g1.pdf (2007, April 10).
[106] L. H. Thomas and K. Umeda, J. Chem. Phys. 26, **293** (1957).
[107] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964).
[108] E. Fermi, Zeitschrift für Physik **36**, 902-912 (1926).
[109] P.A.M. Dirac, Proc. Roy. Soc. **A112**, 281-305 (1926).
[110] R. G. Parr and W. Yang, *Density Functional Theory of Atoms and Molecules* (Oxford University Press, New York, 1989).
[111] C. C. J. Roothan, Rev. Mod. Phys. **23**, 69 (1951).
[112] R. Fitzpatrick, *Variational Principle*. Retrieved from http://farside.ph.utexas.edu/teaching/qmech/Quantum/node127.html (2010, July 20).
[113] A. I. M. Rae, *Quantum Mechanics* (Taylor & Francis Group, New York, 2008) 5th Edition.
[114] A. Szabo and N. S. Ostlund, *Introduction to Advanced Electronic Structure Theory* (Dover Publications Inc., New York, 1996).
[115] D. B. Cook, *Handbook of Computational Quantum Chemistry* (Oxford University Press, New York, 1998).
[116] E. S. Kryachko and E. V. Ludeña, *Energy Density Functional Theory of Many-Electron Systems* (Kluwer Academic Publishers, Boston, 1990).
[117] J. P. Lowe and K. A. Peterson, *Quantum Chemistry* (Elsevier Academic Press, United Kingdom, 2006) 3rd Edition.
[118] W. Koch and M. C. Holthausen, *A Chemist's guide to Density Functional Theory* (Wiley-VCH Verlag GmbH, Germany, 2001) 2nd Edition.
[119] V. Volterra, *Theory of Functionals and of Integral and Integro-differential Equations* (Dover Publications Inc., New York, 1959).
[120] Retrieved from http://www.mrl.ucsb.edu/~ghf/che230a/ghf_monog_appx_C.pdf
[121] G. Arfken, *Mathematical Methods for Physicists* (Academic Press, Orlando, 1985) 3rd Edition, pps. 303-313.
[122] D. S. Sholl and J. A. Steckel, *Density Functional Theory: A Practical Introduction* (John Wiley & Sons, New Jersey, 2009).
[123] M. Levy, Proc. Nat. Acad. Sci. **76**, 6062 (1979).
[124] M. Levy, Phys. Rev. A **26**, 1200 (1982).
[125] E. Lieb, Int. J. Quant. Chem. **24**, 243 (1983).
[126] E. Lieb, *Density Functional Methods in Physics*, edited by R. M. Dreizler and J. da Provindencia (Plenum, New York, 1985).
[127] C. A. Ullrich, *Time-Dependent Density-Functional Theory: Concepts and Applications* (Oxford University Press Inc., New York, 2012).
[128] N. Balazs, Phys. Rev. **156**, 42 (1967).
[129] E. H. Lieb and B. Simon, Phys. Rev. Lett. **31**, 681 (1973).
[130] E. Teller, Rev. Mod. Phys. **34**, 627 (1962).
[131] C. F. V. Weizsacker, Z. Phys. **96**, 431, (1935).

[132] B. Jacob, R. M. Dreizler, and E. K. U. Gross, J. Phys. B. **14**, 2753 (1981).
[133] R. Baer, *Electron Density Functional Theory: lecture notes (rough draft)*. Retrieved from http://www.fh.huji.ac.il/~roib/LectureNotes/DFT/DFT_Course_Roi_Baer.pdf (October 2009).
[134] J. F. Janak, Phys. Rev. B **18**, 7165 (1978).
[135] Retrieved from http://www.physics.metu.edu.tr/~hande/teaching/741-lectures/lecture-06.pdf.
[136] M. E. Casida, Phys. Rev. B **59**, 4694 (1999).
[137] J. P. Perdew and M. Levy, Phys. Rev. Lett. **51**, 1884 (1983).
[138] W. M. C. Foulkes, L. Mitas, R. J. Needs and G. Rajagopal, Rev. Mod. Phys. **73**, 33 (2001).
[139] O. Gritsenko, R. van Leeuwen, and E. J. Baerends, J. Chem. Phys. **101**, 8955 (1994).
[140] P. Ziesche, S. Kurth, and J. Perdew, Comput. Mater. Sci. **11**, 122-127 (1998).
[141] J. P. Perdew, R. G. Parr, M. Levy, and J. L. Balduz, Phys. Rev. Lett. **49**, 1691 (1982).
[142] L. J. Sham and M. Schlüter, Phys. Rev. Lett. **51**, 1888 (1983).
[143] Y. Wang and J. P. Perdew, Phys. Rev. B **43**, 8911 (1991); J. P. Perdew and Y. Wang, Phys. Rev. B **45**, 13244 (1992).
[144] C. Attaccalite, S. Moroni, P. G.-Giorgi, and G. B. Bachelet, Phys. Rev. Lett. **88**, 256601 (2002).
[145] B. Y. Tong and L. J. Sham, Phys. Rev. **144**, 1-4 (1966).
[146] J. P. Perdew and A. Zunger, Phys. Rev. B **23**, 5048 (1981).
[147] K. Burke, J. P. Perdew, and Y. Wang, *Electronic Density Functional Theory: Recent Progress and New Directions*, edited by J. F. Dobson, G. Vignale, and M. P. Das (Plenum, New York, 1998).
[148] J. P. Perdew and Y. Wang, Phys. Rev. B **33**, 8800 (1986) ; J. P. Perdew, Phys. Rev. B **33**, 8822 (1986).
[149] A. D. Becke, Phys. Rev. A **38**, 3098 (1988).
[150] C. Lee, W. Yang, and R. G. Parr, Phys. Rev. B **37**, 785 (1988).
[151] J. Kohanoff, *Electronic Structure Calculations for Solids and Molecules: Theory and Computational Methods* (Cambridge University Press, New York, 2006).
[152] B. G. Johnson, P. M. W. Gill, and J. A. Pople, J. Chem. Phys. **98**, 5612 (1993).
[153] C. W. Murray, N. C. Handy, and R. D. Amos, J. Chem. Phys. **98**, 7145 (1993).
[154] C. Sosa and C. Lee, J. Chem. Phys. **98**, 8004 (1993).
[155] Y. –H. Kim, I. –H. Lee, S. Nagaraja, J. –P. Leburton, R. Q. Hood, and R. M. Martin, Phys. Rev. B **61**, 5202 (2000).
[156] L. Girifalco, *Statistical mechanics of Solids* (Oxford Univertity Press, Oxford, 2003) p. 125.
[157] J. P. Perdew, J. A. Chevary, S. H. Vosko, K. A. Jackson, M. R. Pederson, D. J. Sigh, and C. Fiolhais, Phys. Rev. B **46**, 6671 (1992).
[158] E. H. Lieb and S. Oxford, J. Int. Chem. **XIX**, 427-439, (1981).
[159] M. Ernzerhof and G. E. Scuseria, J. Chem. Phys. **110**, 5029 (1999).
[160] C. Adamo and V. Barone, J. Chem. Phys. **110**, 6158 (1999).
[161] N. W. Ashcroft and N. D. Mermin, *Solid States Physics* (Saunders College, Philadelphia, 1976).
[162] J.P. Dombroski, S.W. Taylor, and P. M. W. Gill, J. Phys. Chem. **100**, 6272 (1996).
[163] M. Marsman, J. Paier, A. Stroppa, and G. Kresse, J. Phys.: Condens. Matter **20**, 064201 (2008).
[164] P. E. Blöch, Phys. Rev. B **50 (24)**, 17953–17979, (1994).
[165] C. Rostgaard, Cond. Mat. Matr-Sci. **0910.1921v2**, 1-25 (2009).

[166] O. K. Al-Mushadani and R. J. Needs, Phys. Rev. B **68**, 235205 (2003).
[167] Retrieved from http://www.update.uu.se/~jolkkonen/pdf/CRC_TD.pdf (2000).
[168] S. B. Zhang, J. Phys.: Condens. Matter **14**, R881–R903 (2002).
[169] P. Erhart, K. Albe, and A. Klein, Phys. Rev. B **73**, 205203, (2006).
[170] M. H. Cohen and V. Heine, Phys. Rev. **122**, 1821 (1961).
[171] S. Lany and A. Zunger, Modelling Simul. Mater. Sci. Eng. **17**, 084002 (2009).
[172] T. R. Paudel and W. R. L. Lambrecht, Phys. Rev. B **77**, 205202 (2008).
[173] G. Makov and M. C. Payne, Phys. Rev. B **51**, 4015-4022 (1995).
[174] S. Lany and A. Zunger, Phys. Rev. B **78**, 235104 (2008).
[175] F. Oba, A. Togo, I. Tanaka, J. Paier, and G. Kresse, Phys. Rev. B **77**, 245202 (2008).
[176] A. F. Wright and N. A. Modine, Phys. Rev. B **74**, 235209 (2006).
[177] J. Shim, E. K. Lee, Y. J. Lee, and R. M. Nieminen, Phys. Rev. B **71**, 035206 (2005).
[178] C. W. M. Castleton, A. Höglund, and S. Mirbt, Phys. Rev. B **73**, 035215 (2006).
[179] H.-P. Komsa, T. T. Rantala, and A. Pasquarello, Phys. Rev. B **86**, 045112 (2012).
[180] F. Oba, M. Choi, A. Togo, and I. Tanaka, Sci. Technol. Adv. Mater. **12**, 034302 (2011)
[181] G.-Y. Huang and B. D. Wirth, Phys. Rev. B **88**, 085203 (2013).
[182] C. W. M.Castleton, A. Höglund, and S. Mirbt, Phys. Rev. B **73**, 035215 (2006).
[183] M. Choi, F. Oba, and I. Tanaka, Phys. Rev. B **78**, 014115 (2008).
[184] W. Chen, C. Tegenkamp, H. Pfnür, and T. Bredow, Phys. Rev. B **82,** 104106 (2010).
[185] C. Freysoldt, B. Grabowski, T. Hicket, J. Neugebauer, G. Kresse, A. Janotti, and C. G. Van de Walle, Rev. Mod. Phys. **86**, 253-305, (2014).
[186] K. K. Rebane, *Impurity Spectra of Solids: Elementary Theory of Vibrational Structure*, translated from Russian by J. S. Shier (Plenum Press, New York-London, 1970) Chapter 2, pps. 36-54.
[187] A. M. Stoneham, *Theory of Defects in Solids: Electronic Structure of Defects in Insulators and Semiconductors* (Oxford University Press, Oxford, 1975), p 296.
[188] D. L. Dexter, C. C. Klick, and G. A. Russell, Phys. Rev. **100**, 603 (1955).
[189] B. K. Ridley, *Quantum Processes in Semiconductors* (Oxford University Press, NY, 2013), p 207.
[190] A. Alkauskas, Q. Yan, and C.G. Van de Walle, Phys. Rev. B **90**, 075202 (2014).
[191] L. Shi, K. Xu, and L.-W. Wang, Phys. Rev. B **91**, 205315 (2015).
[192] R. H. Bartram and A. M. Stoneham, Solid State Commun. **17**, 1593-1598 (1975).
[193] C. H. Leung and K. S. Song, Solid State Commun. **33**, 907 (1980).
[194] S. Wakita, Y. Suzuki, and M.Hirai, J. Phys. Soc. Jpn. **50**, 2781-2782 (1981).
[195] G. Kresse and J. Furthmüller, Phys. Rev. B **54**, 11169 (1996).
[196] R. Dingle, D. D. Sell, S. E. Stokowski, and M. Ilegems, Phys. Rev. B **4**, 1211 (1971); B. Monemar, Phys. Rev. B **10**, 676 (1974).
[197] H. Morkoç, in *Handbook of Nitride Semiconductors and Devices* (Wiley, New York, 2008), Vols. 1–3.
[198] K. T. Jacob and G. Rajitha, J. Cryst. Growth **311**, 3806–3810 (2009).
[199] I. A. Buyanova, Mt. Wagner, W. M. Chen, B. Monemar, J. L. Lindström, H. Amano, and I. Akasaki, Appl. Phys. Lett. **73**, 2968 (1998).
[200] W. M. Chen, I. A. Buyanova, Mt. Wagner, B. Monemar, J. L. Lindström, H. Amano and I. Akasaki, Phys. Rev. B **58**, R 13351 (1998).
[201] A. F. Wright, J. Appl. Phys. **90**, 6526 (2001).

[202] H. J. von Bardeleben, J. L. Cantin, H. Vrielinck and F. Callens, L. Binet, E. Rauls and U. Gerstmann, Phys. Rev. B **90**, 085203 (2014).
[203] C. Freysoldt, J. Neugebauer, and C. G. Van de Walle, Phys. Rev. Lett. **102**, 016402 (2009).
[204] C. Freysoldt, J. Neugebauer, and C.G. Van de Walle, Phys. Stat. Sol. B **248**, 1067 (2010).
[205] M. A. Reshchikov, A. Kvasov, T. McMullen, M. F. Bishop, A. Usikov, V. Soukhoveev, and V. A. Dmitriev, Phys. Rev. B **84**, 075212 (2011).
[206] M. A. Reshchikov, M. A. Foussekis, J. D. McNamara, A. Behrends, A. Bakin, and A. Waag, J. Appl. Phys. **111**, 073106 (2012).
[207] There is also the +/2+ level very close to the valence band, however a positively charged defect is expected to repel a free hole, and therefore is not observed in experiment.
[208] S. Nakamura, N. Iwassa, M. Senoh, and T. Mukai, Jpn. J. Appl. Phys. **31**, 1258 (1992).
[209] J. Neugebauer and C. G. Van de Walle, Phys. Rev. Lett. **75**, 4452 (1995).
[210] S. Limpijumnong and C.G. Van de Walle, Physica Status Solidi B-Basic Research **228**, 303 (2001).
[211] A. F. Wright, C. H. Seager, S. M. Myers, D. D. Koleske, and A. A. Allerman, J. Appl. Phys. **94**, 2311 (2003).
[212] A. F. Wright and S. M. Myers, J. Appl. Phys. **94**, 4918 (2003).
[213] J. -S. Park and K. J. Chang, Appl. Phys. Express **5**, 065601 (2012).
[214] J. Baur, K. Maier, M. Kunzer, U. Kaufmann, and J. Schneider, Appl. Phys. Lett. **65**, 2211 (1994).
[215] E. Malguth, A. Hoffmann, W. Gehlhoff, O. Gelhausen, M. R. Phillips, and X. Xu, Phys. Rev. B **74**, 165202 (2006).
[216] J. A. Freitas Jr., M. Gowda, J. G. Tischer, J.-H. Kim, L. Liu, and D. Hanser, J. Cryst. Growth **310**, 3968 (2008).
[217] M. A. Reshchikov, R. H. Patillo, and K. C. Travis, Mat. Res. Soc. Symp. Proc. **892**, FF23.12 (2006).
[218] G. Mills, H. Jonsson and G. K. Schenter, Surface Science, **324**, 305 (1995).
[219] L.C. Kimerling, Solid-State Electronics **21**, 1391 (1978).
[220] C. H. Patterson, Phys. Rev. B **74**, 144432 (2006).
[221] F. Gallino, G. Pacchioni, and D. Valentin, J. Chem. Phys. **133**, 144512 (2010).
[222] A. Chakrabarty and C. H. Patterson, J. Chem. Phys. **137**, 054709 (2012)